Differential Punishment and the Consequences for Labor Supply*

Katherine B. Coffman and Dalila Figueiredo†

November 26, 2026

Please see the most recent version here.

Abstract

Recent field evidence suggests that Black workers may face differential punishment in the workplace, receiving harsher sanctions or larger backlash in response to similarly negative outcomes. However, it is still largely unknown what drives these differences in punishment and what the consequences of differential punishment are for future workplace interactions. We study these questions using a large, controlled experiment, pairing managers and workers for repeated rounds of a real-effort project whose outcome depends on both worker effort and chance. We randomly vary whether the manager has the opportunity to punish the worker after observing the outcome, and ask how the anticipation and experience of punishment shapes workers' effort decisions. In our setting, conditional on the outcome, managers do not punish Black and white workers differently on average. However, a racial gap appears when effort is observed to be low: at low effort, we estimate that managers are 17.2 percentage points more likely to sanction a Black worker and, on average, sanction them 50% more relative to a white worker. On the worker side, the possibility of punishment does not increase initial worker effort, while the experience of being punished has a significant negative impact on future effort. Punishment also makes workers less willing to continue working for that manager, even at the expense of their own earnings. In our context, punishment lowers joint surplus, with the worker bearing the entire loss, and exacerbates racial disparities.

JEL: J15, J16, J71

Keywords: discrimination, punishment, race, gender, labor supply, real-effort experiment

*Pre-registered: AEARCTR-0017113. We received funding from Harvard Business School and The Max Planck Institute for Behavioral Economics. We would like to thank Omar Abdel Haq and Jannis Bernschein for excellent research assistance, and Marcela Carvalho, Lucas Finamor, Scott Kostysha, Chris Roth, Hannah Ruebeck, Perihan O. Saygin, and Matthias Sutter for their helpful feedback and suggestions.

†Coffman: kcoffman@hbs.edu, Harvard Business School; Figueiredo: figueiredo@econ.mpg.de, Max Planck Institute for Behavioral Economics.

1 Introduction

Punishment is a feature of many educational and professional environments. Teachers discipline students, supervisors sanction employees, and firms penalize misconduct. While punishment can incentivize effort and rule-following, it can also demotivate workers, ending up being costly to the organization. Furthermore, because these decisions are often discretionary, they may be subject to bias and discrimination.

Indeed, a growing body of field evidence suggests that punishment is not always levied equally across groups. For example, Black students are substantially more likely to be expelled or suspended (Kinsler, 2011; Darling-Hammond et al., 2023), darker-skinned football players more likely to receive foul calls (Magistro and Wack, 2023), and non-White employees face more disciplinary procedures in the UK healthcare system (NHS England and NHS Improvement, 2020). Gender disparities have also been documented. Women often face harsher consequences than men following similarly negative outcomes, as seen among financial advisors and surgeons (Egan et al., 2022; Sarsons, 2017), and female leaders receive more scrutiny and are terminated more often after poor firm performance (Carvalho, 2025; Ferrario and Carvalho, 2026). These patterns matter not only because punishment directly affects the workers who receive it in the short-run, but also because discipline can shape how an individual engages with the organization going forward, contributing to decisions to disengage or exit.

This existing literature raises several important questions. First, what causes unequal punishment? Do managers' interpretations of the same outcome vary by worker identity, and if so, is this due to biased beliefs? Field data rarely speak to managers' information, beliefs, or counterfactual punishment decisions, making it challenging to pin down these potential mechanisms. Second, does the threat or experience of punishment affect workers' labor supply, such as how hard they work or whether they stay in the organization? Understanding the full set of consequences of punishment requires studying both the manager's decision to punish and the worker's response to that decision.

We choose to study these questions within the context of a simulated workplace, with a specific focus on differences by race. We design our experiment to explore whether Black and white workers receive different punishments for negative outcomes when worker effort is a key determinant of outcomes. This allows us to ask whether manager stereotypes about effort drive any differences in punishment that we observe. We also track how Black and white workers adjust effort in anticipation of and response to punishment, allowing us to speak to the broader impacts of (differential) punishment across groups.

We conduct a large, synchronous experiment with more than 3,000 online participants. Individuals are randomly assigned to the role of the worker or manager, and paired. The worker must complete a real effort task (slider task), where their effort increases the chance that a joint project succeeds. If the outcome is good, both the manager and the worker receive 10 tokens. If the project fails and the outcome is bad, only the worker gets paid 10 tokens. Thus, while altruism or fear of punishment could motivate effort, the worker may be tempted to shirk. Pairs play two mandatory rounds together, and we surprise workers with the opportunity to opt into a third round with the same manager following Round 2.

The experiment has two key variations. The first is whether the manager can punish the worker following the round (destroying any positive number of the worker's earned tokens). Punishment costs the manager nothing but is socially costly, as any sanctioned tokens are simply destroyed. Importantly, workers know from the outset whether they can be punished, allowing us to study the effects of possible punishment on worker effort.[1] The second treatment variation is whether the manager observes worker effort. In the Hidden-Effort treatment, managers observe only the project outcome (good or bad), whereas in the Known-Effort treatment they learn the worker's effort. This allows us to study whether more information on worker effort mitigates any race or gender gaps in punishment that we might

[1]By contrast, managers start Round 1 knowing they will make a sanction decision and that there is a 50% chance it will be implemented. Following the first round, managers learn whether their sanction is imposed (Punishment arm) or not (No-Punishment arm). If it is, they can punish in the following rounds. Otherwise, they know they will not have another opportunity to punish.

observe.[2]

Managers and workers learn the race and gender of their partners through an avatar that represents them and is visible throughout the study (Ruebeck, 2025).[3] Participants build their avatar at the start of the study, before learning about their partner or the task. The gender of this avatar is fixed by the participant's stated gender in the demographics survey. Participants choose their avatar's skin tone: more than 90% of participants choose an avatar skin-tone that appears consistent with their stated race in the demographics survey. To maximize statistical power to identify differences in levied punishment by worker race, we over-sample Black participants and assign them to the worker role, while white participants are randomly-assigned to either the manager or worker role. As a result, half of our study pairs feature a white manager with a Black worker, while the other half feature a white manager with a white worker. Managers and workers are approximately balanced on gender.

Our controlled experiment exogenously varies the possibility of punishment, while holding fixed the underlying production environment and controlling the information available to managers and workers. We compare otherwise similar manager–worker relationships in which punishment is or is not available, and we track effort, outcomes, and punishment decisions over repeated interactions. By studying the two sides of the interaction together, we can examine not only how punishment is allocated, but also how punishment affects subsequent effort supply and the continuation of the relationship. In addition, we exogenously vary the ability of the manager to perfectly monitor worker effort, allowing us to understand what role beliefs about effort play in driving differential punishment decisions.

Overall punishment rates are somewhat low. Managers punish in fewer than 1/3 of Round 1 decisions; conditional on punishing, the average sanction is just over 5 tokens (of a possible 10). Unsurprisingly, punishment is most common after negative outcomes. Unconditionally,

[2]In the first round, workers know that the manager will observe their effort with a 50% chance. At the end of the first round, workers learn whether the manager observed their effort and know the same will happen in the following rounds.

[3]We use the package developed by Ruebeck (2025) as the foundation of our avatar construction, making only minor changes to suit our setting.

Black workers receive somewhat larger sanctions than white workers, but this largely reflects differences in outcomes. Conditional on the project's success and worker effort, we find no significant race (or gender) gap in punishment, on average.

We find that managers use punishment differently when they can observe effort. Managers who observe effort are more likely to punish and they impose larger sanctions, on average. As expected, punishment decreases as (observed) worker effort increases, consistent with managers punishing based upon perceived responsibility for the outcome.

Importantly, rather than mitigating a disparity in punishment, revealing effort seems to produce a racial gap. Among lower effort workers whose effort is observed, Black workers receive larger sanctions than white workers. At low effort, managers are 17.2pp ($p < 0.05$) more likely to punish Black workers. Additionally, while the estimates are imprecise, our data suggests managers are approximately 8pp ($p \approx 0.14$) more likely to sanction a Black worker the maximum possible amount (10 tokens). The racial gap narrows as effort increases, closing at roughly the sample average effort. The gap operates mainly on the extensive margin, on whether to punish at all, and it is most pronounced among male workers. We find no analogous gender gap at any effort level. The evidence therefore does not point to a broad tendency to punish Black workers more harshly. Rather, differential punishment emerges in a specific part of the performance distribution: when the manager knows that the worker provided relatively little effort. Note that this pattern seems inconsistent with statistical or beliefs-based discrimination, which would predict a larger racial gap when effort is unobserved. Similarly, we find no evidence that managers expect less effort or hold higher standards for Black workers. Instead, our results suggest that the same, precise low-effort signal is punished differently by race, even when stated beliefs are similar.

A key feature of our study is that we can investigate how workers respond to the possibility and experience of being punished. The possibility of punishment does not impact first-round worker effort. Workers exert approximately the same effort, on average, regardless of whether their manager can sanction them. This holds across worker race and gender. However,

workers respond strongly to actually receiving punishment. Leveraging random-assignment, we identify that receiving a Round 1 sanction significantly reduces Round 2 effort. This discouragement effect increases with sanction size, making punishment counterproductive in this environment. These responses vary meaningfully by worker race and gender. White men are substantially more responsive to their manager's sanction decision than women or Black men. They increase effort more in response to their manager sparing them from possible punishment and decrease effort more steeply after being punished.

These repeated rounds mimic an environment where the worker must remain in the same workplace and under the same manager. However, in many settings, workers have the opportunity to exit the relationship. To shed light on this decision, we surprise the worker with the opportunity to opt into a third round with the same manager. Continuing guarantees weakly positive additional earnings, at a modest additional time cost. Workers who decline the third round are also asked whether they would have been (hypothetically) willing to play with a different manager, allowing us to distinguish relationship-specific exit from a general decline in willingness to work. We show that the larger the first-round punishment, the less likely workers are to continue the relationship with the same manager. Black men are the exceptions, and the likelihood to opt-in to Round 3 does not depend on the imposed sanction. The sanction has little effect on the willingness to participate with a different manager.

Taken together, we find that when workers are effectively "stuck" in the relationship and must continue working for the same manager, white men are more likely to reduce their effort in response to harsher punishment. When given the opportunity to exit, workers are more likely to exit the relationship following punishment, except for Black men. Punishment therefore produces different response margins across workers: men, particularly white men, respond more strongly through subsequent effort, even though this lower future effort may itself be punished. Women and Black men, on the other hand, do not adjust their effort in response to being punished, and women are more likely to make a quiet but costly exit

before Round 3 (Dana et al., 2006).

In sum, our evidence suggests that punishment fails to increase productivity and may lead to racial and gender disparities. In our setting, allowing for punishment does not increase effort, and when it is implemented, it can backfire, since it can lead to lower effort and costly exits. Simultaneously, punishment creates a racial disparity in the treatment of low-effort workers, since Black workers are sanctioned more harshly than equally unproductive white workers. At the group level, punishment increases the share of zero-surplus outcomes and reduces average surplus, with the losses falling on the worker, since their payoffs decline at no benefit for the manager. While monitoring alone increases worker effort, combining it with punishment destroys more value than it creates.

Our paper contributes to several strands of literature. First, we add to research on discrimination in workplace evaluations and punishment. While field studies show women and racial minorities sometimes face harsher consequences (Egan et al., 2022; Sarsons, 2017; Carvalho, 2025; Ferrario and Carvalho, 2026), our controlled design demonstrates that differential punishment can emerge even among workers with the same low effort, and even when managers' stated expectations do not differ across groups. We highlight that, at least in our setting, these gaps are not driven by belief-based discrimination (Bordalo et al., 2016, 2019; Coffman et al., 2021, 2025), but instead suggest differential treatment of the same outcome. We also highlight the market-level consequences of differential punishment, showing that punishment matters for worker behavior, and that these effects vary by worker race and gender.

Second, we contribute to a broader line of research on punishment, incentives, and reciprocity. While prior work shows punishment can sustain cooperation (Fehr and Gächter, 1998, 2000; Carpenter, 2007), our results show it can also undermine cooperation: punished workers reduce future effort and are less likely to continue with the manager.

Finally, we contribute to the literature on organizational design and labor-market inequality. In our setting, managerial discretion to punish destroys value, increases the risk

of poor outcomes, and generates unequal treatment of low-effort Black workers. The consequences vary: white men are more likely to retaliate through subsequent effort reduction, while women are more likely to exit. These findings suggest sanction-based institutions may lead to inefficiency, inequality, and costly exits.

2 Experimental Design

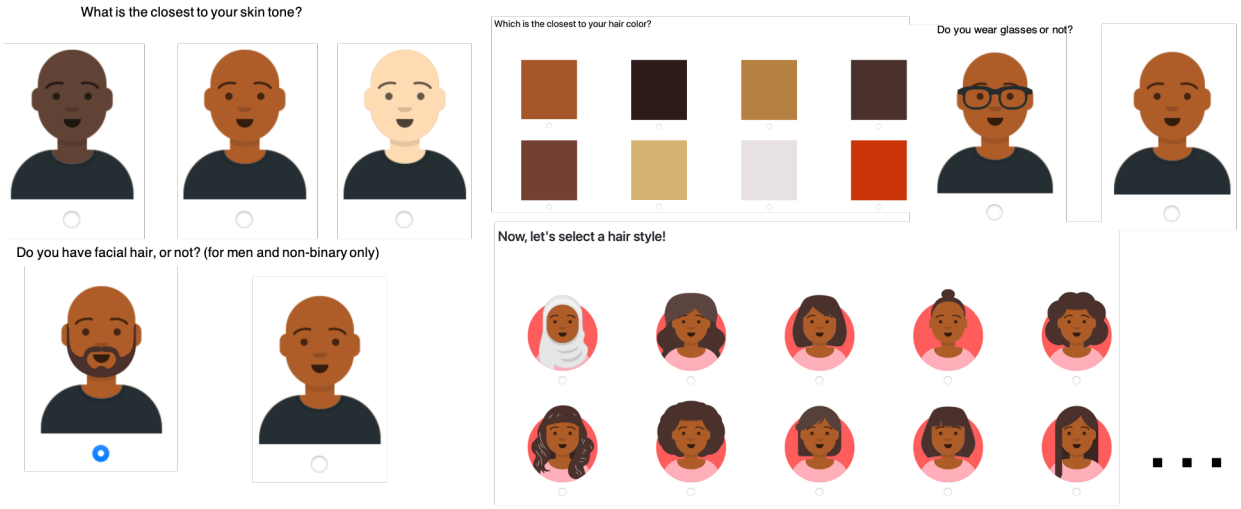
In our experiment, participants first complete an introductory survey and build an avatar that conveys their race and gender. They are then matched into manager-worker pairs and play a two-round effort game in which a good outcome benefits both members of the pair, and the manager may, in some treatments, punish the worker. We conduct the experiment online with real-time synchronous interaction between the two members of each pair. We targeted 3,200 participants, of whom a quarter (800) are Black, and the remainder are white.

Introductory Survey and Avatar Constructions Participants begin with a short demographic survey covering race, gender, education, and employment status, followed by an unincentivized elicitation of economic preferences: time and risk preferences, positive and negative reciprocity, altruism, and trust (Falk et al., 2018).

Because our design requires managers to observe worker race and gender, we need to convey these characteristics unambiguously while limiting the experimenter demand effects that heightened salience of race and gender might otherwise induce. We do so through an avatar, following Ruebeck (2025). At this stage, participants know only that they will play a game with another participant. Each participant builds an avatar that represents them throughout the study: the participant's answer to the gender question fixes the avatar's gender, and participants then choose the avatar's skin tone, hair color and style, facial hair (if male or non-binary), and whether to wear glasses. Figure 2 shows the sequence of choices. Embedding race and gender among several other choices mitigates their salience.

For the purposes of our analysis and in line with our pre-registration, we label participants

Figure 1: Avatar choice



who select the pale skin tone 'white' and those who select the medium or dark tones 'Black.' Reassuringly, fewer than 4% of participants choose a skin tone that appears inconsistent with the race they reported in the survey (see Table A1).

Slider Task and Effort Cost Elicitation Participants are next introduced to the slider task, the real-effort task that the worker will have to complete later. A slider is completed by dragging an indicator from the midpoint of a 0 to 100 scale to a randomly generated target. All participants must read the instructions and complete two sliders before proceeding.

We then measure each participant's cost of effort using a multiple price list. Participants choose, at each of a range of wages from \$0.05 to \$0.50 in \$0.05 increments, whether to complete a fixed number of sliders, S, for that wage or to complete none for nothing. For instance, a given row of the price list would ask a participant to choose whether to (1) complete S sliders and receive 0.30, or (2) do 0 sliders and receive nothing. Across three price lists, we set the fixed number of sliders, S, to 15, 30, or 45. Each participant is randomly assigned to see one of the possible price lists, and one row is selected at random and implemented at the end of the study, thereby incentivizing choices. Participants who chose to complete S sliders in the selected row must complete S sliders to receive the given

wage and finish the study.

The Manager-Worker Game Participants are randomly assigned to roles, pairs, and treatment arms. Each participant is either a manager or a worker. All managers are White, while workers can be either Black or White. As a result, half of the groups are mixed-race (a Black worker and a White manager), and half are same-race (a White worker and a White manager).[4] Because the experiment requires two participants to play simultaneously, pairs are formed in real time from the pool of participants available at each moment, with the matching algorithm designed to maintain a 1:1 ratio of mixed-race and same-race pairs.[5] Pairs are assigned to one of four treatment arms, each containing a balanced number of mixed-race and same-race pairs.

Both members of the pair then learn their role, read the instructions, and learn that the game consists of two rounds, one of which is randomly chosen to determine payoffs. They answer comprehension questions and cannot proceed until they answer correctly. Before the worker begins, the manager completes a short "kick-off" task of counting zeros in a matrix; this task does not affect either player's payoff but is intended to increase the manager's engagement and sense of ownership in the game.

The worker then has 180 seconds to complete up to 50 sliders. Each completed slider raises the probability of a good outcome by one percentage point from a baseline of 5%, so a worker who completes all 50 sliders raises this probability to 55%. The reasons for these odds were twofold: first, we wanted the manager to decide on punishment based on their evaluation of the roles of both effort and luck, and second, we wanted to create a scenario in which bad outcomes were relatively likely, so that we would have more punishment opportunities, thus increasing the power of our analysis. A good outcome pays both players 10 tokens, and a bad outcome pays the worker 10 tokens and the manager 0. This creates an incentive

[4]While it would of course be interesting to study groups with Black managers as well, we made this design decision in order to maximize statistical power for studying differences across *worker* race.

[5]As participants arrive, each white participant is matched as a manager to a waiting Black worker when one is available, and is otherwise placed in a same-race pair. Details of the matching procedure are provided in Appendix A.1.

for the worker to "shirk," as their payoff in the good and bad state is the same. However, the worker may also have reason to exert effort; they may care about the manager's payoff (other-regarding preferences) or, if punishment is possible, they may try to avoid punishment. While the worker works, the manager reports how many sliders they expect the worker to complete and how many they think the worker should complete, allowing us to study the relationship between managers' expectations and their subsequent decisions.

We hypothesized that biased beliefs, or stereotypes, could be a key driver of gaps in punishment. Conditional on observing a negative outcome, managers with biased beliefs may reach different conclusions about the underlying causes of that bad outcome, leading to differences in punishment. One key design goal was to choose a task that could trigger potentially field-relevant racial stereotypes within our stylized context. This led us to choose a real effort task for the worker task, where "poor performance" is likely to be interpreted as resulting from low effort, rather than from low ability. In line with past research in sociology and social psychology, individuals may hold negative stereotypes about the work ethic of Black individuals (Boskin, 1988; Collins, 2004; Lemons, 1977). We hypothesized that this could lead managers to underestimate the effort of Black workers relative to white workers, potentially contributing to racial gaps in punishment. Conversely, it is less clear whether our task triggers relevant gender stereotypes. If anything, manager beliefs that women may be more conscientious (Vianello et al. (2013)) or more other-regarding (Exley et al. (2025)) could benefit female workers in this context.

Treatment Variations We implement a 2×2 design that varies whether the manager can punish the worker and whether the manager observes the worker's effort.

In the punishment arm, the manager can deduct up to 10 tokens from the worker at the end of the round, after the outcome is revealed. Punishment is not costly to the manager, but it is a social loss, as the manager does not keep the tokens deducted from the worker. Workers are informed at the start of the first round whether their manager can punish them.

This variation lets us study how anticipated punishment affects worker effort and how past punishment affects future effort.

In order to maximize statistical power for detecting differential punishment according to worker race, we introduce the possibility of punishment to all *managers* at the start of the game.[6] At the start of the first round, every manager is told that they can deduct up to 10 tokens from the worker, and that this sanction will be imposed with 50% probability. We ask all managers to make a punishment decision in the first round and only after revealing whether they are indeed in the punishment arm. Only decisions from those managers assigned to the punishment arm are implemented. Managers in the no-punishment arm do not have their round 1 decisions implemented, and they make no further punishment decisions.

The other treatment variation regards whether managers observe workers' effort. If the effort is revealed, the manager observes both the outcome and the number of sliders the worker completed at the end of each round. If effort is hidden, managers observe only whether the outcome is good or bad. Managers know from the outset whether they will observe effort. Workers are told at the start of the first round that the manager observes their effort with 50% probability, and that observation in the first round implies observation in the second. In the hidden-effort arms, we additionally ask the manager, after their punishment decision, to guess how many sliders the worker completed. The main goal of this variation is to test whether (biased) beliefs about effort drive punishment decisions, and how those compare with certainty on effort. Although testing classic moral hazard issues is not the study's primary goal (Holmstrom, 1982; Hoppe and Schmitz, 2018), this design also lets us examine how effort revelation affects effort provision in the no-punishment treatment arm.

Note that by the start of the second round, both managers and workers know their exact treatment arm. Finally, before the second-round outcome is revealed, we ask the worker whether they would like to play a third round with the same manager, identical to the first two and paid on top of the randomly-selected round. Workers who decline are asked whether

[6]Note that, in contrast, only workers assigned to the punishment arm ever learn about the possibility of punishment.

they would instead have been willing to play with a different manager.[7] These choices allow us to study the effects of punishment and monitoring on workers' labor supply, both on the intensive margin (effort exerted in the second round) and the extensive margin (willingness to participate in the third round).

Implementation We recruit participants through Prolific, restricting the sample to participants located in the United States, aged 18 or older, with at least 100 prior submissions and an approval rating of 95% or higher, who self-identify as white or Black/African-American.[8] We use Prolific's screening to oversample Black participants so that the recruited sample is three-quarters white and one-quarter Black. This way, we target 3200 participants who complete the study, forming 1600 pairs, half mixed-race and half same-race, evenly split across the four treatment arms.

We pre-registered our study with the AEARCT registry.[9] Our first set of pre-specified research questions focus on managers. Do managers punish Black workers more than white workers for negative outcomes? And, if so, are these differences well-explained by managers holding different beliefs about worker effort? Our primary outcome measures for these questions are punishment amounts (# of tokens) levied during Round 1 across treatment and worker race, conditional on project outcome. To test the role of beliefs, we look directly at Round 1 manager expectations and also compare punishment decisions across the Hidden-Effort and Known-Effort arms. We pre-specified using Round 2 beliefs and punishment decisions to further investigate the motivations for punishment.

Our second set of pre-specified research questions ask how the anticipation of punishment shapes the labor supply decisions of workers. Do they anticipate harsher punishment, and if so, does this impact their decisions about how much effort to invest in their work or how willing they are to continue working under that manager? To test the effects of anticipated

[7]Unfortunately, it was not feasible to incentivize this additional choice, as repairing managers and workers at this stage would have been impractical.

[8]Toward the end of data collection, in an effort to meet our targeted sample size, we opened recruitment to the UK. Four percent of our final analysis sample is UK-based.

[9]`https://www.socialscienceregistry.org/trials/17113`, trial 17113.

punishment, we compare effort exerted in Round 1 across treatment and worker race. In addition, we look at the follow-on effects of punishment by investigating effort in Round 2 and Round 3 opt-in decisions, by treatment, race, and Round 1 punishment decisions.

Finally, we explore the combined impacts of these channels on total group earnings by treatment and race. These pre-specified questions and outcomes guide our analysis below.

Note that our pre-registration specifies race as our primary focus, with gender differences as a secondary focus. This is in line with our choice of a real effort task as the worker task, intended to activate race stereotypes more so than gender stereotypes.[10] To streamline writing and presentation, we present results by race and by gender simultaneously below.

Participants must pass several attention and comprehension checks distributed throughout the session. We pre-registered our exclusion criteria and implement them as follows. We excluded from the sample (i) any participant who has failed two attention checks[11],(ii) any participant who dropped out during the study, (iii) any participant whose partner dropped out during the study, and (iv) any pairs with a non-binary participant. Participants who complete the study receive a fixed participation fee of $4 plus bonus payments. First, in the multiple-price-list task, we randomly select one row, and participants are paid according to the accept/reject choice they made for that row. Second, they earn $0.10 per token obtained in one round of the game: in each of the first two rounds, a worker earns up to 10 tokens (reduced by any sanction imposed by the manager), and a manager earns 10 tokens following a good outcome and 0 otherwise, with one of these two rounds selected at random for payment. Finally, if the worker chooses to play a third round, both participants are additionally

[10]We hypothesized that the gender gap could plausibly go either way; male workers could receive greater punishment than female workers (consistent with a manager belief that women may work harder/be more conscientious than men), or male workers could receive less punishment than female workers (perhaps due to differences in perceived power, hierarchy, or status). Our task seems unlikely to generate negative stereotypes about female performance relative to male performance.

[11]Initially, the pre-registration stated that any participant who failed an attention check would be dropped. However, to comply with Prolific's guidance, the code dropped participants only after their second attention check failure. Upon realizing this inconsistency, we updated the pre-analysis plan to align with the implemented code and Prolific's guidelines. At the time of this change, we had already collected 35% of the data. We applied the revised rule retroactively to ensure all participants in the sample meet the new criteria. Restricting the sample to participants who failed no attention check, as originally pre-registered (916 workers; 467 managers in the Known-Effort arm), leaves our main results largely unchanged (Tables A42–A45).

paid $0.10 per token earned in that round.

3 Data and Summary Statistics

Sample construction Table A2 shows the sample from recruitment to the dataset used in the analysis. Of the 6334 participants who began the study, 77% were matched into manager-worker pairs, 10% were excluded from the study for either failing attention checks or having their partner do so, 1.6% of participants were excluded because one of the participants in the group was non-binary, and 50% completed all rounds without dropping out or having their partner drop out. Of the individuals who started the study, 1468 participants (23%) arrived but were never matched to a partner; 18 participants dropped out before being matched, and among matched participants, 570 (9%) dropped out themselves, and 511 (8%) were lost because their partner dropped out. Attrition is similar across roles (26%).

Sample composition Throughout, worker race is defined by avatar skin tone. Avatar skin-tone closely tracks a worker's own reported race: 99% of workers who chose the lightest skin tone self-report being white, while 90% of workers who choose one of the darker two skin tones self-report being Black. In Table A4, we observe that Black and white workers differ along many observables.[12] Black workers are six years younger, less likely to have at most a high-school education, and more likely to be employed. They also report slightly higher risk tolerance, altruism, and negative reciprocity, and lower trust. They also have a lower measured effort cost (0.89 vs. 0.97 cents per slider). Similarly, male and female workers also differ. Males are more likely to be employed and make more comprehension mistakes than females. Male workers exhibit more risk tolerance, patience, and negative reciprocity. However, they report lower positive reciprocity and altruism.

[12]Table A3 reports baseline characteristics by role. The means of managers and workers differ on several observables (not formally tested), but this is explained by the fact that Black and white individuals are different, and all managers are white, and workers can be both. The difference in comprehension mistakes is likely mechanical: managers have to answer more comprehension questions than workers during the study.

Balance across treatment arms. Table A5 reports means of the observed characteristics by treatment arm. No covariate differs significantly across arms at the 5% level, and the share of participants assigned to a mixed-race pair is close to one-half in every arm, confirming that the real-time matching algorithm preserved the intended mixed/same-race ratio without inducing imbalance. Tables A6 and A7 show that balance also holds within role, for managers and workers separately. A handful of point estimates are individually significant, but their number and size are consistent with what would arise by chance under this many comparisons.

Attrition differs across treatment arms (Table A8). Overall, 26% of participants did not complete the study, either because they dropped out or because their partner did. Attrition is lower in the two Known-Effort arms (22.1% and 24.2%) than in the Hidden-Effort arms (27.5% and 28.3%), a difference of 4.8 p.p ($p = 0.01$). The pattern is the same for managers and workers, and for own and partner dropout. Attrition is, however, selected on race and age. Among Black participants, 18% dropped out, compared with 12% of white participants ($p < 0.01$), and participants who dropped out are about two years older than those who completed the study ($p < 0.01$). Dropped and retained participants do not differ on gender, education, or employment (Table A9).

Most worker attrition happens before the first task. Of the 296 workers who dropped out, 276 did so in Round 1, and 259 of these (94%) left before opening the task page, and hence before any outcome or sanction. The remaining 17 opened the task page and then dropped out. Of those, eight completed zero sliders, and nine completed between 1 and 35. Only 20 workers dropped out after Round 1.

Receiving a sanction does not make workers more likely to drop out. Among the 1,641 workers who completed Round 1, 20 (1.2%) dropped out afterward: 14 in the No-Punishment arms and 6 in the Punishment arms. Of the 216 workers who were sanctioned in Round 1, only one dropped out afterward. Comparing workers in the Punishment arms with workers in the No-Punishment arms whose manager chose the same sanction but could not implement

it, the sanction does not predict later dropout ($p = 0.27$). Attrition is therefore not driven by punishment, which is reassuring for the labor supply results.

4 Results

Our experiment allows us to study both sides of the market: how managers decide to punish and how workers respond to both the possibility and realization of punishment. First, we discuss manager decisions; then, we turn to worker decisions. Finally, we discuss the implications of these findings for market efficiency and profit distribution.

4.1 Manager Decisions: The Supply of Punishment

In this section, we study what drives managers' decisions to punish and whether punishment varies by workers' race and gender. First, we document the overall supply of punishment: how often and how heavily managers sanction, and how punishment responds to the round's outcome and observing workers' effort. Then, we analyze whether managers punish Black and white workers and male and female workers differently, conditional on what they observe. Finally, we study the beliefs and reasoning behind managers' decisions, particularly how managers' expectations about effort and the justifications they offer for their choices can account for the documented findings.

4.1.1 Determinants of Punishment

We begin by asking how much punishment managers levy, and what drives their decisions. Recall that in Round 1, every manager makes a punishment decision, knowing that it will be implemented with 50% probability. This design choice lets us collect punishment decisions from all managers, not just those in the Punishment arm, increasing the analysis's statistical power.

Overall, punishment is rare and small in scale. Only 28% of managers impose any sanc-

tion at all, and the average sanction is just 1.54 of the 10 tokens available (Table A10). Conditional on punishing, the average sanction is 5.46 tokens. These low levels of punishment are perhaps surprising, given that punishment is free and bad outcomes are very common (68% of managers face a bad outcome in the first round, as presented in Table 1). Even conditional on a bad outcome, only 37% of managers choose to sanction their worker, with an unconditional average deduction of 2.11 tokens. Conditional on punishing, the average punishment is 5.65 tokens. One possibility is that managers refrain from punishment in part because of fear that workers will reciprocate with lower effort in Round 2, in line with the constraining impacts of counter-punishment documented in (Nikiforakis, 2008). Surprisingly, 9% of managers who receive a *good* outcome nonetheless punish their worker. This is more common when effort is observed (13% in the Known Effort arm) but still sometimes occurs when it is hidden (5%). This is the first indication that information about effort helps explain managers' behavior.

Table 1: The Choice Environment: Outcomes and Punishment

Round-1 outcome	N	% of rounds	% punishing	Mean sanction	Mean | punished
Bad outcome	1035	68.0	37.3	2.11	5.65
Good outcome	487	32.0	8.8	0.34	3.84
All	1522	100.0	28.2	1.54	5.46

Note: This table describes the Round-1 environment managers face, one observation per manager. Outcomes are classified as Bad or Good depending on the realization of the joint project in Round 1. N is the number of managers, % of rounds is the share of managers facing each outcome, % punishing is the share who choose a strictly positive sanction, and Mean sanction is the average number of tokens deducted. Mean | punished is the average sanction among managers who chose a strictly positive sanction.

Indeed, observing effort is a key determinant of sanctions. Managers randomly allocated to Known-Effort arms are 8 pp ($p < 0.01$) more likely to punish and unconditionally sanction almost 0.4 more tokens than their Hidden-Effort counterparts. Note that this is not driven by differences in worker effort across treatments. Conditional on both worker effort and observed outcome, managers who had access to information on worker effort punished more

(Table 2).[13]

Table 2: Punishment Decisions by Treatment Arm

	Punished	Sanction	Punished	Sanction	Punished	Sanction
Punishment arm	-0.015	0.021	-0.015	0.020	-0.011	0.043
	(0.023)	(0.151)	(0.023)	(0.149)	(0.022)	(0.145)
Known effort arm	0.081***	0.390***	0.080***	0.384**	0.069***	0.311**
	(0.023)	(0.150)	(0.023)	(0.149)	(0.022)	(0.145)
Worker's effort			-0.005***	-0.033***	-0.003***	-0.018**
			(0.001)	(0.007)	(0.001)	(0.007)
Good outcome					-0.265***	-1.651***
					(0.021)	(0.122)
Constant	0.248***	1.333***	0.385***	2.193***	0.410***	2.348***
	(0.020)	(0.122)	(0.033)	(0.224)	(0.032)	(0.220)
R^2	0.008	0.004	0.026	0.021	0.098	0.086
Observations	1,522	1,522	1,522	1,522	1,522	1,522

Note: This table shows the effect of each treatment on punishing decisions. Punished is a binary variable that equals one if the manager applied any sanction. Sanction represents the number of tokens the manager fined the worker. Columns 3–4 add the worker's Round-1 effort (sliders completed); columns 5–6 additionally add the Round-1 outcome. Heteroskedasticity-robust standard errors in parentheses. Asterisks denote pointwise statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.1.2 Gender and Racial Punishment Gaps

We now turn to one of the study's central questions: do managers punish workers differently by race and gender? We start by analyzing the unconditional gaps, and then account for the round characteristics.

Unconditionally, managers punish Black workers more. Table A11 shows that managers are 4.3pp ($p < 0.10$) more likely to sanction a Black worker and deduct 0.3 more tokens on average ($p < 0.05$). This gap equals roughly 20% of the mean sanction. We find no meaningful or statistically significant difference by worker gender.[14]

[13]Naturally, there is no difference across punishment arms, since managers only learn whether their sanction is implemented after making their decision. Ex-ante there is a 50% probability that it will be.

[14]This contrasts with field evidence documenting harsher punishment for women, however note that punishment in our context likely hinges on assessments of worker effort, not ability, perhaps a meaningful difference from several field contexts of interest (surgery, firm leadership, etc.)

Importantly, however, differences in outcomes contribute to this gap. Conditional on the outcome of the game, the estimated Black-white gap is smaller and no longer statistically significant (see column (2) of Table 3). The estimates are imprecise, so we cannot confidently conclude that the remaining gap is negligible or that the outcome differences explain most of the raw disparities. We also continue to find no evidence of a gender gap.

Table 3: Punishment Decisions: Treatment and Worker's Characteristics

	S_{g1}				
	(1)	(2)	(3)	(4)	(5)
Black worker	0.302**	0.175	0.167	-0.070	-0.188
	(0.151)	(0.206)	(0.206)	(0.250)	(0.246)
Male worker	-0.011	0.093	0.106	0.107	0.077
	(0.153)	(0.212)	(0.212)	(0.255)	(0.248)
Good outcome (G_{g1})		-1.749***	-1.734***	-1.744***	-1.801***
		(0.180)	(0.181)	(0.180)	(0.181)
Black × Good outcome		0.080	0.085	0.108	0.080
		(0.239)	(0.239)	(0.240)	(0.240)
Male × Good outcome		-0.104	-0.111	-0.114	-0.066
		(0.242)	(0.242)	(0.242)	(0.240)
Known effort arm			0.309**	0.077	0.083
			(0.145)	(0.229)	(0.226)
Punishment arm			0.048	0.049	0.078
			(0.146)	(0.146)	(0.141)
Black × KE arm				0.454	0.462
				(0.290)	(0.285)
Male × KE arm				0.013	0.013
				(0.295)	(0.288)
Constant	1.393***	1.977***	1.791***	1.906***	2.350**
	(0.120)	(0.165)	(0.192)	(0.207)	(0.968)
Manager controls	No	No	No	No	Yes
Observations	1,522	1,522	1,522	1,522	1,522

Note: OLS of the Round-1 sanction S_{g1} (tokens deducted, 0–10) on worker race and gender and their interactions, one observation per manager, all arms pooled. *Black worker* and *Male worker* denote the worker's avatar race and gender (base: White, female). Column 1 reports the unconditional gaps; column 2 adds the Round-1 outcome and its interactions (*Black/Male* × *Good outcome*; base: bad outcome); column 3 adds the treatment-arm indicators; column 4 adds their interactions with the Known-Effort arm (base: Hidden Effort); column 5 adds manager controls (demographics, elicited preferences, the MPL effort cost and list fixed effects, and attention/comprehension counts). Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As expected, we find that the final outcome drives punishment decisions, with managers punishing significantly less when the outcome is good (by about 1.7 tokens, $p < 0.01$). Conditional on outcomes, it is still true that observing effort increases punishment (by an

average of 0.3 tokens, $p < 0.05$). However, we also observe that the monitoring effect is concentrated entirely on Black workers. Once we allow the treatment effect to vary by race, the Known-Effort effect is essentially zero for white workers; only the Black × Known-Effort interaction is marginally significant ($p = 0.105$). When managers observe effort, they punish Black workers 0.4 tokens more than white workers, with no analogous gender gap. Interestingly, information *amplifies* rather than disciplines the racial gap.

In Table 4, we explore this pattern further by focusing specifically on the Known-Effort treatment and looking at heterogeneity by worker effort. Overall, we find that punishment falls with effort: one extra completed slider decreases punishment by 0.05 tokens ($p < 0.01$, column (2)). The racial gap is most pronounced at lower effort levels. At low effort, Black workers are punished 1.4 tokens more than white workers ($p < 0.05$, column (4)). Managers cut Black workers' sanctions faster as effort rises ($-(-0.017 - 0.052) = 0.069, \quad p < 0.05$), but the estimated gap closes only once a Black worker completes about 26 sliders, at the average effort level. Equally low-performing individuals face different sanctions by race when effort is known. We find no parallel pattern by gender.

We also observe that the racial gap is on the extensive margin - whether to punish or not - rather than on the severity of the sanction (Table A13). We estimate that at low effort, managers are 17.2pp ($p < 0.05$) more likely to punish Black workers, but conditional on punishing, we find no statistically significant differences in sanction levels by race (a large but imprecise 1.38 tokens). Additionally, evidence suggests managers are more likely to sanction the maximum possible (10 tokens) if the worker is Black. We estimate that at low effort, Black workers are 8pp ($p \approx 0.14$) more likely to receive 10-token sanctions. Although underpowered, the estimate should be compared to a predicted probability of 12% for white workers with the same effort and outcome.

Splitting the sample further reveals that the racial gap is most pronounced among lower-effort *men* (Table A12). We look separately at men, women, Black, and white workers and, conditional on the outcome, analyze how effort interacts with gender or race within

Table 4: Punishment Decisions: the Role of Effort

	S_{g1}			
	(1)	(2)	(3)	(4)
Black worker	0.403	0.201	1.490**	1.355**
	(0.291)	(0.286)	(0.620)	(0.594)
Male worker	0.099	0.104	0.301	0.279
	(0.302)	(0.296)	(0.621)	(0.598)
Good outcome (G_{g1})	-1.778***	-1.523***	-1.697***	-1.821***
	(0.258)	(0.260)	(0.266)	(0.275)
Black × Good outcome	0.037	-0.024	0.237	0.228
	(0.349)	(0.348)	(0.345)	(0.354)
Male × Good outcome	-0.056	-0.057	0.029	0.032
	(0.355)	(0.354)	(0.352)	(0.351)
Effort (N_{g1})		-0.047***	-0.018	-0.017
		(0.011)	(0.018)	(0.017)
Black × N_{g1}			-0.054**	-0.052**
			(0.021)	(0.020)
Male × N_{g1}			-0.009	-0.009
			(0.021)	(0.020)
Constant	2.018***	3.243***	2.503***	3.677**
	(0.241)	(0.375)	(0.533)	(1.520)
Manager controls	No	No	No	Yes
Observations	770	770	770	770

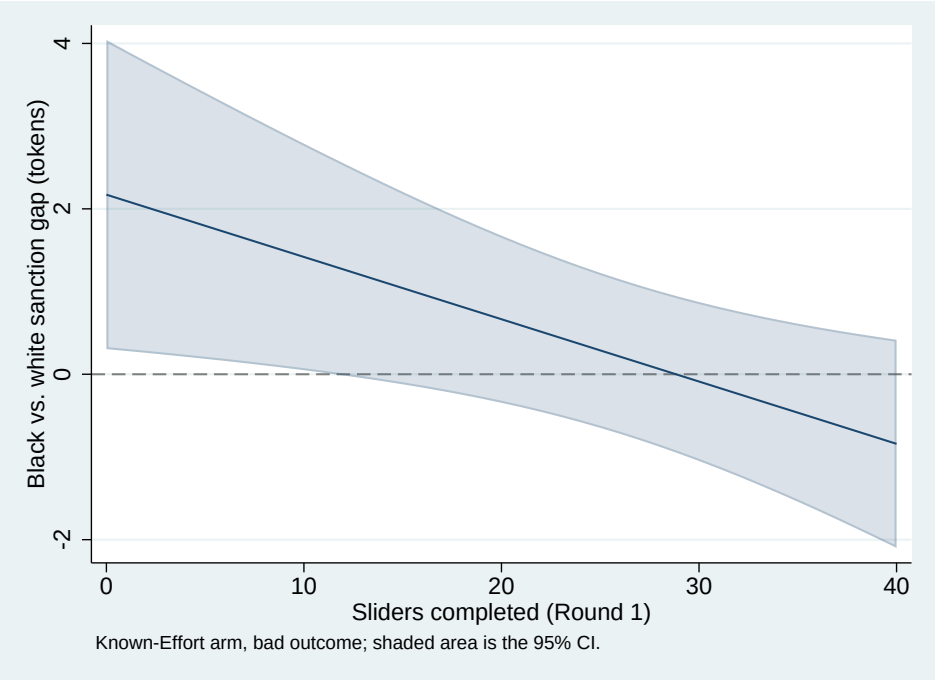
Note: OLS of the Round-1 sanction S_{g1} (tokens, 0–10), restricted to the Known-Effort arm, where the manager observes the worker's actual effort N_{g1} before choosing the sanction; one observation per manager. Column 1 conditions on the outcome and its race and gender interactions; column 2 adds observed effort N_{g1} (sliders completed); column 3 adds *Black/Male* × N_{g1}; column 4 adds manager controls (demographics, elicited preferences, the MPL effort cost and list fixed effects, and attention/comprehension counts). *Black worker* and *Male worker* denote the worker's avatar race and gender (base: White, female). Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

each population. We estimate that, among men, Black workers are fined 2.2 tokens more than White men at low effort ($p < 0.05$). The gap closes as effort rises, by 0.08 tokens per slider completed ($p < 0.05$). For this population, assuming linearity, the punishment gap closes only when a worker completes about 29 sliders, 3 sliders above the average of 26 (see Figure 2).[15] Among women, by contrast, the Black–white gap is smaller and not statistically significant. This pattern does not depend on the linearity assumption: when we instead group workers into effort bins, or allow sanctions to be quadratic in effort, we estimate a racial punishment gap of similar size among lower-effort men that disappears at

[15]Note that the gap becomes statistically indistinguishable from zero (at the 5% level) beyond roughly 12 sliders.

higher effort, although these more flexible estimates are less precise (see Figures A1 and A2).

Figure 2: Black-white punishment gap among men, by observed effort



Note: Marginal effect of a Black (vs. white) male worker on the Round-1 sanction, from the Known-Effort specification of Table A12, evaluated at a bad outcome across levels of observed effort. The shaded area is the 95% confidence interval.

Managers' sanctions also respond to effort differently across groups. For Black workers, an additional completed slider significantly reduces sanctions by 0.06 tokens for Black women ($p < 0.01$) and by 0.09 tokens for Black men ($p < 0.01$). White workers' sanctions do not move with effort at all. This racial difference in effort sensitivity is itself significant among men (-0.08, $p < 0.05$).

In summary, the racial punishment gap operates through the treatment of lower-effort men. Managers appear most lenient with lower-effort white men and most punitive with lower-effort Black men. The gap does not depend on the manager's own gender: female and male managers both sanction lower-effort Black men more than white men, and neither the gap nor its closing with effort differs significantly between them (Appendix Table A14, Panels A and B).

4.1.3 Beliefs, expectations, and stated reasoning

With these punishment patterns in mind, we now ask how managers' beliefs about effort explain their punishment decisions and how they justify them. We also study whether these factors help account for the racial gap.

While the worker completes the task in each round, we ask managers: (i) how many sliders they believe the worker *should* do, and (ii) how many they think the worker *will* do. On average, managers believe workers should complete 33.5 out of 50 sliders. However, managers also expect them to fall short and believe workers will do about 28 (Tables A15 and A16). In the Hidden-Effort treatment arms, we also ask managers to estimate how many sliders the worker completed after observing a good or bad outcome and after making their punishment decisions. After seeing a bad outcome, managers estimate the worker did 18 sliders on average; after a good outcome, they estimate 30. Overall, workers' race or gender does not significantly predict these expectations.[16]

Tables A17 and A18 show that these beliefs matter for punishment decisions. In the Hidden-effort arms, the manager never learns the true effort, so decisions can depend only on expectations and beliefs. Instead, in the Known-effort arms, managers learn how many sliders the worker completed. Thus, their decisions can account for the difference between expectations and reality. We observe that the difference between how many sliders the manager believes a worker *should* do and either the manager's belief (0.034 tokens, $p < 0.01$) or the worker's actual effort (0.019 tokens, $p < 0.01$) increases the levied sanction.[17]

Importantly, these expectations seem equally informative for sanctions against Black and white workers in the Known-Effort arms. There is no significant interaction between race and expectations. In the Hidden-Effort arms, on the other hand, there is suggestive evidence that the difference between what a manager believes a worker should versus will do matters

[16]The exceptions are that managers believe that males *should* do more sliders than female workers, but they do not think they *will*.

[17]Interestingly, the difference between their guess about how many sliders the worker completed before and after learning the outcome (HE arm) also predicts the sanction amount (0.022, $p < 0.05$), but this loses importance once the manager learns the true effort (0.003, not statistically significant).

more for punishment decisions of Black workers compared to white workers. These point estimates are noisy, but their size is meaningful (0.025 for $Black \times (Should - \hat{N}_{g1})$, $p < 0.14$, and 0.032 for $Black \times (Will - \hat{N}_{g1})$, $p < 0.10$).

One question is whether the differential punishment we observe results from managers believing that punishment is a more effective future motivator for Black workers. Two pieces of evidence suggest this is unlikely to explain the race gap. First, if managers sanctioned Black workers because they expected a larger effort response, their stated expectation of the worker's Round-2 effort should rise more after sanctioning a Black worker. We find that it does not. In fact, managers in the punishment arms who punish a Black worker lower their expectation of future effort by about eight sliders more than managers who punish a white worker, with no such difference in the placebo arm where the sanction is chosen but not applied. Within the Known-Effort sample of male workers, we also do not find evidence of such behavior (Table A19).[18] Second, managers' stated explanations contradict this hypothesis. Managers are no more likely to explain their punishment decision as motivational for Black workers compared to white workers. The racial gap is concentrated among managers who describe their decision as an evaluation of the worker's performance. Among this group, a Black man's low effort is more often attributed to his character, while a white man's is more often excused or treated as something to encourage. Appendix A.2 reports the classification and these results in detail.

4.1.4 Insights from Round 2

The previous analysis relied only on Round 1 decisions, where endogeneity is least threatening. However, Round 2 might help clarify the dynamic path of punishment and shed light on the underlying mechanisms. In Round 2, managers have more information about the paired

[18] An important caveat is that managers were asked to guess the number of sliders completed after deciding on punishment. We did not ask them beforehand because we did not want to prime their decisions. As a result, we cannot distinguish between a manager who formed an expectation of the number of sliders completed before deciding on punishment and a manager who rationalized their punishment decisions ex-post through their stated beliefs.

worker. In the Known-Effort arm, they also know how the worker changed their effort behavior between rounds. We use this information to study whether managers still display a racial gap for low-effort workers. Since only managers in the Punishment treatment arms decide on sanctions in Round 2, our analysis focuses on them. In summary, we will restrict this analysis to managers in the Known-Effort arm with Punishment.

Overall, in this treatment arm, punishment rates fall (from 33% to 27%), and the average unconditional sanction remains constant (from 1.72 tokens to 1.71 tokens). This means that the managers who punish are punishing more severely (5.2 to 6.4 tokens). We are interested in whether managers continue to disproportionately penalize lower-effort Black workers. Table A27 shows that they do not; any racial gaps are no longer significant in Round 2.

In Round 2, effort becomes an even more important predictor of punishment behavior. An increase in effort of one slider leads to a decrease of 0.131 tokens in the sanction ($p < 0.01$), and the raw correlation between the sanction and observed effort roughly doubles, from -0.25 to -0.51. The second-round sanction tracks effort itself, and we no longer find a role for race. Consistent with this, managers also reward improvement for workers of both races (Table A28). Regressing the change in the sanction between rounds on the change in the worker's effort, we observe that each additional slider completed relative to Round 1 lowers the sanction by 0.08 tokens for white workers ($p < 0.01$), and by 0.16 for Black workers ($p < 0.01$). In Round 2, managers can compare each worker with their own first-round effort. As a result, white workers' sanctions, which did not respond to effort in Round 1 (-0.01 per slider), do respond in Round 2. Since Black workers at the bottom of the effort distribution are more likely to move out of it between rounds, and improvement lowers their sanction at least as much as for white workers, crediting effort helps explain why the first-round gap closes.

In the next section, we discuss why we detect the racial gap only in the first round, and what that implies about the nature of the bias.

4.1.5 Discussion

The supply-side results leave us with a pattern to explain. Racial differences in punishment arise when the manager *observes effort* but not when effort is hidden. They also concentrate at *low effort*, where a Black worker near the bottom of the effort distribution is sanctioned by over a token more than an otherwise-identical White worker, and Black men are punished by over two tokens more. The overall gap is mostly explained by willingness to punish rather than by punishment intensity. We do not find the same effect in the second round of the interaction. In this section, we discuss which mechanisms might explain this behavior.

The first candidate, for which there is an extensive literature, is statistical discrimination. If Black workers are, on average, less productive at the task, a manager who cannot perfectly observe effort might treat race as a proxy for productivity. The data provide evidence against this channel. The gap appears exactly when effort *is* observed and vanishes when effort is hidden, which is when race could serve as a proxy. The punishment gap is documented conditioning on observed effort, so equally low-performing Black and white workers are punished differently, seemingly inconsistent with the standard statistical discrimination hypothesis.

The patterns also seem inconsistent with standard taste-based discrimination. If managers simply disliked Black workers, independent of performance, we would likely observe this pattern across treatments and specifications. Instead, we see a much narrower racial disparity.

We also studied the role of expectations, but we found that managers do not hold different expectations for Black and white workers. In the Known-Effort arm specifically, manager expectations in terms of sliders that should or will be completed are statistically indistinguishable across race. Furthermore, it does not appear to be the case that managers expect punishment to be a more effective motivator for Black workers. Conditional on punishing, managers do not expect a larger effort response from Black compared to white workers. If anything, they expect less.

So, what can we say about what drives the racial gap in punishment that we observe? The fact that the gap is mostly concentrated on low-effort Black men may reflect negative stereotypes around work ethic and character (Boskin, 1988; Collins, 2004; Lemons, 1977). The same low slider count may be read as diagnostic of the worker's type when he is a Black man, but discounted as noise or bad luck when he is a White man. Since the manager's expectations do not differ by race, what may differ is the interpretation or meaning attributed to the effort signal. Indeed, the racial gap in punishment is concentrated among managers who describe their decision as an evaluation of the worker's performance. And, the text of manager explanations for Round 1 punishment decisions reveal potential biases along these lines: a Black man's low effort is more often attributed to his character, while a white man's is more often excused or framed as something to encourage, although these differences in language are suggestive rather than precisely estimated.

Finally, the fact that the punishment gap concentrates at low effort suggests that managers may need a pretext to punish Black workers more harshly, and only directly observing low effort provides sufficient pretext (compared to observing more moderate effort or even not observing effort at all). This is consistent with Dana et al. (2007), for example, who show that people choose to be fair because they dislike appearing unfair to others. However, as soon as there is wiggle room, they show less fairness concern. In the second round, fewer workers remain at very low effort, so the 'excuse' arises less often, which may be why we do not detect the gap there.

4.2 Worker Decisions: Anticipation of and Responses to Punishment

In this section, we study how workers adjust their labor supply in response to the possibility and realization of punishment. We first examine how anticipating possible punishment shapes initial effort choices in Round 1. Then, we study how workers' Round 2 effort responds to realized punishment decisions. Finally, we study the extensive-margin decision of

whether to continue working for the same manager in an optional Round 3, asking whether earlier punishment increases the likelihood of opt-out. Throughout, we focus on two dimensions of heterogeneity that are salient in the workplace and in the discrimination literature: race and gender.

4.2.1 Anticipation of punishment and initial effort

We first ask whether the possibility of punishment changes initial effort provision among workers. We compare Round 1 effort decisions (number of sliders completed) across our two randomly-assigned treatments: No Punishment and Punishment. In the Punishment treatment, workers know their manager can sanction them after observing the Round 1 outcome. In the No Punishment treatment, the worker instructions do not mention the possibility of punishment. All other aspects of the treatments are held constant, including the fact that all workers are uncertain whether their manager will observe the exact number of sliders they complete in Round 1. A large experimental literature on punishment in public goods games and other social dilemmas suggests that the possibility of punishment may increase effort (Fehr and Gächter, 2000; Carpenter, 2007).

And yet, we observe no effect of the possibility of punishment on initial effort decisions. On average, workers complete 26 sliders in Round 1 (SD of ≈ 11), just over half the total amount possible. This is nearly identical across treatments. The mean number of sliders completed varies by less than 0.5 sliders across treatments ($p > 0.9$).[19] Furthermore, there is no differential impact of possible punishment by race or by gender. Given the baseline probability of a good outcome (5%) and the returns to effort (1pp per slider), this translates to an expected likelihood of a good outcome of just under one-third in the first round.

Why might effort decisions fail to respond to anticipated punishment in our setting? A few possibilities seem worth considering. First, workers may have been aware of the

[19]The distributions are also indistinguishable across treatments, see Appendix Figure A.3. And the share of workers who do no sliders (approximately 1%) and the share of workers who do all 50 sliders (approximately 1.8%) are also unchanged across treatments.

possibility of punishment but did not expect managers to actually punish them. This is somewhat internally consistent with the overall low rates of punishment we indeed observe in Section 4.1. However, we cannot test this explanation directly. We did not ask workers about their expectations of punishment, as we felt that direct questions on anticipated punishment could raise concerns about experimenter demand or priming effects. A second possibility is that psychological forces offset the financial incentive to increase effort that possible punishment provides. For instance, the possibility of punishment might shift how workers feel about their manager, reducing altruism or goodwill toward the other party; it might also shift how they feel about the work context itself, creating discouragement or low morale (Bowles and Polania-Reyes, 2012). Again, we do not have data that allows us to test these explanations directly.

More broadly, we can ask what factors do seem to matter for Round 1 effort decisions. In Table A31, we predict Round 1 effort from a battery of information we have on workers, including their demographics (race, gender, age, education, employment status), economic preferences (as measured by our introductory survey), and other study behavior. We find many of these factors play a significant role. First, we see clear demographic differences, with Black workers and older workers completing significantly fewer sliders on average, and men completing more. Second, the impact of measured economic preferences is both sizable and intuitive. More risk-tolerant participants complete significantly fewer sliders, while participants who more strongly endorse positive reciprocity complete more sliders. Both of these effects are stronger in the punishment treatment than in the no-punishment treatment. Negative reciprocity predicts less effort, but only in the no-punishment treatment, suggesting that the possibility of punishment may motivate exactly those workers who would expect to punish if in the manager's position. These patterns are consistent with the idea that effort in our setting has both financial and relational components: while some workers may be responding primarily to monetary incentives, others seem to view effort as a reciprocal response to being trusted with the task, in line with previous evidence from the laboratory

(Fehr and Gächter, 1998). Finally, workers who failed an attention check or needed additional attempts to answer comprehension questions exert marginally less effort, consistent with relatively lower engagement in the study. Those who have higher effort cost also exert less effort.

Taken together, the Round 1 evidence suggests that variation in initial effort is driven by workers' characteristics, rather than by the contextual feature of whether punishment is available. This finding may inform models in which the threat of punishment disciplines effort: in our context, merely introducing a punishment technology does not substantially reshape workers' initial labor-supply decisions on net. However, as we describe next, the actual experience of being punished has meaningful consequences for future effort provision.

4.2.2 Responses to Punishment

On average, workers increase effort in Round 2, completing about 4 more sliders ($p < 0.01$). Recall that, at the conclusion of Round 1, workers learn whether there was a good outcome and, if they were assigned to the Punishment treatment, their manager's punishment decision. In addition, workers learn whether or not their manager observes only the outcome for the round (Hidden Effort treatment) or the outcome and their exact number of sliders completed (Known Effort treatment). Table 5 displays the average increase in effort by treatment assignment for the full 2×2. Workers increase their effort in Round 2 across all treatment cells. Workers who learn that their manager has seen (and will see) the number of sliders that they complete increase their Round 2 effort by 1.2 more sliders ($p < 0.05$) than workers in the Hidden Effort treatment.

This section focuses on how Round 1 punishment decisions shape worker effort in Round 2. Table 5 shows that, on average, increases in effort are similar across the No Punishment and Punishment arms. However, as expected, this masks substantial heterogeneity, particularly in the Punishment treatment. To understand this fully, we need to consider not just the possibility of punishment, but the exact amount of punishment that is received. Intu-

itively, a worker who *could* have been punished by their manager but wasn't may react much differently than a worker who received a large sanction, even holding Round 1 effort fixed.

Table 5: Effort Increases in Round 2 by Treatment

Average Increase in Number of Sliders Completed (Round 2 – Round 1)			
	Unknown Effort	Known Effort	
No Punishment	3.25 (0.56)	4.91 (0.37)	$p < 0.05$
Punishment	4.01 (0.51)	4.77 (0.44)	n.s.
	n.s.	n.s.	

Note: Each cell presents the average increase in effort from Round 1 to Round 2. Row ends present p-values from two-sided t-tests across the Unknown and Known Effort treatment, conditional on Punishment Treatment status. Column ends present p-values from two-sided t-tests across the No Punishment and Punishment treatment, conditional on Effort Treatment status.

Table 6 explores these effects, predicting a worker's Round 2 effort, broken out by the manager's sanction decision from Round 1. We also include controls for the worker's Round 1 effort, whether there was a good outcome in Round 1, and an indicator for the Known Effort treatment. In Columns 1 - 4, the key focus is on the indicator for the Punishment treatment, asking whether, conditional on a manager's Round 1 sanction amount, the experience of actually receiving that sanction impacts Round 2 effort. While the estimates are noisy given the data partition, we see the type of heterogeneity we would expect. Column 1 presents data for workers whose managers chose a sanction of 0, i.e., no punishment. Within this subgroup, being assigned to the Punishment treatment has a *positive* impact on future effort. That is, learning that you were not punished when you could have been increases effort, relative to a context where punishment was never a possibility. Columns 2 - 4 reveal that as the sanction amount grows, the effect on future effort becomes directionally negative.

Column 5 pools all workers. We predict Round 2 effort from the indicator for being in the Punishment treatment, the manager's Round 1 sanction choice, and their interaction. The interaction is the key coefficient of interest, allowing us to formally test how effort responds to increasingly larger sanctions in the Punishment treatment. We estimate a significant,

negative coefficient of approximately -0.4 ($p < 0.05$): each additional token deducted in Round 1 lowers Round 2 effort by roughly 0.4 sliders, relative to the counterfactual where the same sanction is chosen but (randomly) not implemented.

Table 6: How Round 2 Effort Responds to Round 1 Punishment Decisions

	Round 2 Effort (Sliders Completed)				
	Manager's Round 1 Sanction Decision				All groups
	$S = 0$	$S \in [1, 3]$	$S \in [4, 6]$	$S \geq 7$	Pooled
Punishment Arm	0.801*	0.586	-0.529	-2.956	0.956**
	(0.473)	(1.528)	(1.934)	(1.987)	(0.465)
R1 Sanction Amt.		0.237	-2.850	-1.944**	-0.180
		(1.116)	(2.702)	(0.797)	(0.127)
Pun. Arm × R1 Sanction Amt.					-0.420**
					(0.197)
R1 effort	0.867***	0.823***	0.589***	0.629***	0.805***
	(0.024)	(0.078)	(0.108)	(0.114)	(0.025)
Good outcome	0.076	-2.701	-4.101	-5.818	0.014
	(0.472)	(1.660)	(3.352)	(5.899)	(0.475)
Known-effort arm	0.966**	2.305	1.106	1.718	1.317***
	(0.473)	(1.665)	(2.001)	(2.078)	(0.461)
Constant	7.422***	6.129*	26.065*	30.374***	8.707***
	(0.787)	(3.254)	(14.386)	(8.183)	(0.784)
Observations	1,093	139	156	134	1,522
Adjusted R^2	0.608	0.486	0.226	0.279	0.527

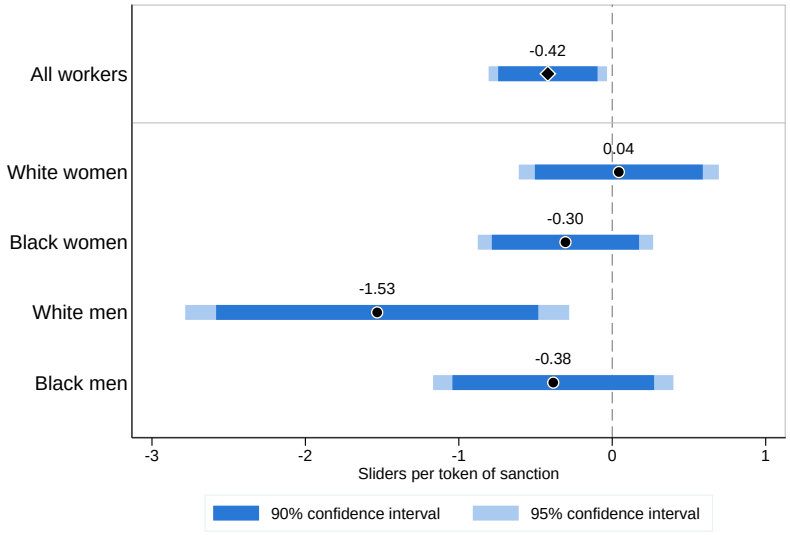
Note: OLS of Round 2 worker effort N_{g2} (sliders) across groups (one observation per group). Columns 1 - 4 split the sample according to Manager's Round 1 sanction decision. Columns 5 and 6 use all groups. Column 6 interacts each regressor of Column 5 with Black and Male; the interactions of Black and Male with Round 1 effort, the Round 1 outcome, and the Known-Effort arm are included but not displayed. All columns use heteroskedasticity-robust standard errors; standard errors displayed in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Our results show that a manager who has the power to punish, but refrains from doing so, elicits higher subsequent effort; a manager who exercises that power, especially at high levels, triggers a reduction in effort. These patterns are consistent with past work on potential backfire effects of fines or sanctions (Gneezy and Rustichini, 2000; Falk and Kosfeld, 2006) and the role of reciprocity in the workplace (Kube et al., 2012, 2013). In our setting, implemented punishment appears to at least somewhat crowd out workers' intrinsic or reciprocal motivation and reduce their willingness to exert effort. Note that this isn't an effect isolated to low-effort workers: we find that the response to an implemented sanction is similar for workers who exerted above or below median Round 1 effort (Table A32).

We can ask how these effects vary by race or gender. Table A33 presents the results

fully interacted with our race and gender indicators. Overall, we estimate similar impacts of punishment across race, with no significant differences in how Black workers adjust effort after receiving a sanction compared to white workers (see modest and statistically insignificant coefficient on Black x Pun. Arm x Sanction Amt.) We do, however, estimate a modest gender difference. Men's effort is more responsive to Round 1 punishment decisions than women's, increasing their effort somewhat more after being "spared" punishment (coefficient on Male x Punishment Arm, $p = 0.11$) and decreasing their effort more for each additional token of sanction levied (see coefficient on Male x Pun. Arm x Sanction Amt., $p < 0.10$). Looking at each subgroup, shown in Figure 3, we observe that white men are the ones who adjust their effort more strongly in response to punishment.

Figure 3: Round 2 effort response to a Round 1 sanction, by race and gender



Note: Each marker is the effect of one more token of Round 1 sanction on Round 2 effort (sliders) when the sanction is implemented, relative to the No-Punishment arm where the same sanction is chosen but not implemented. The dark bar is the 90% confidence interval and the light bar the 95% confidence interval, from heteroskedasticity-robust standard errors.

These results indicate that "backlash" to punishment is not uniform. It is particularly strong for men, white men in particular, who respond both more positively to being spared and more negatively to being punished. Women and Black men, in contrast, adjust their effort less in response to the same sanctions. While we can only speculate on mechanisms,

two possibilities seem worth noting. First, punishment may be a larger negative surprise for groups more likely to expect favorable treatment, generating larger effort responses from white men. Alternatively, these differences may stem from differences in future expectations. When deciding whether or not to reduce effort in response to Round 1 punishment, workers may consider how much punishment they expect to receive after Round 2, with white men perhaps expecting less retaliation for lower effort.

4.2.3 Punishment and the decision to quit

Beyond adjusting intensive-margin effort, workers in our study can also respond to punishment by exiting the workplace. After completing Round 2, but before learning the respective outcome, they choose whether to opt in to Round 3. Playing Round 3 is weakly payoff-increasing for both the worker and the manager, as it offers additional expected earnings in exchange for a modest time cost. In addition, managers only learn of the possibility of the third round if their worker opts in. Thus, a worker who opts out both forgoes potential earnings and withholds an opportunity from the manager.

Overall, 64% of workers choose to play Round 3. Average opt-in rates do not differ significantly across the Punishment and No Punishment treatments, mirroring the null average effect of the punishment regime on Round 1 and Round 2 effort. However, once again, realized sanctions matter. Table 7 reports linear-probability models in which the dependent variable is an indicator for choosing to play Round 3 with the same manager (Column 1) or a different manager (Column 2). To minimize endogeneity concerns and parallel the analysis of Table 6, we condition only on Round 1 realizations, including Round 1 effort, whether or not there was a good outcome in Round 1, and an indicator for the Known-Effort treatment.

We find that among workers in the punishment arm whose manager chose a zero sanction, the probability of opting in is about 5pp higher than for comparable workers in the no-punishment arm ($p < 0.10$), echoing the positive effect of avoiding Round 1 punishment on Round 2 effort. Second, for workers in the punishment arm who do receive a Round 1

Table 7: Willingness to Continue Working in Round 3

	Opted in to Round 3	
	With same manager	With any manager
Punishment Arm	0.045*	0.035
	(0.027)	(0.027)
R1 Sanction Amt.	0.002	-0.000
	(0.006)	(0.006)
Pun. Arm × R1 Sanction Amt.	-0.028***	0.011
	(0.008)	(0.008)
Black worker	0.023	0.028
	(0.025)	(0.024)
Male worker	0.026	0.025
	(0.025)	(0.024)
Constant	0.452***	0.511***
	(0.043)	(0.042)
Observations	1,522	1,522
Adjusted R^2	0.033	0.015

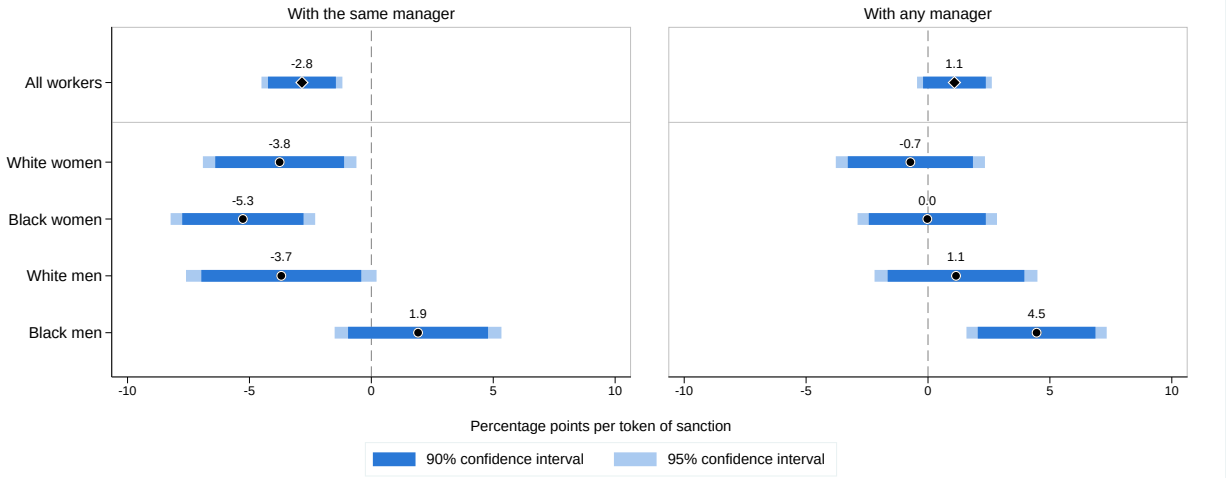
Note: OLS of the worker decision to opt-in to Round 3 with the same manager or with any manager, i.e., a different manager, across all groups. Controls for Round 1 effort, the Round 1 outcome, and the Known-Effort arm. All columns use heteroskedasticity-robust standard errors, displayed in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

sanction, each additional token deducted in Round 1 reduces the probability of opting in to Round 3 by around 3pp relative to the No Punishment counterfactual ($p < 0.01$). In the Punishment treatment, the share of workers who opt in with their existing manager falls from about 70% when they were not punished to roughly 50% when they were punished. Intuitively, higher realized punishment not only discourages effort in the next period but also increases the likelihood that workers exit the relationship altogether, even though doing so is privately and socially costly. These high rates of exit in the Punishment treatment are noteworthy, given that workers have the option to engage in Round 3 but simply "shirk," yielding themselves higher monetary payoffs (even net of a sanction) but at the possible additional psychological cost of being punished or perceived negatively by their manager.

The left panel of Figure 4 shows the results by subgroup (also presented in Table A33). Overall, the higher the implemented sanction, the lower the probability that workers want to continue the relationship with the same manager. Women react more strongly, particularly Black women, who are 5.3p.p less likely to want to participate in a third round. White women follow, at 3.8 p.p. less likely to opt in, ant then white men (3.7p.p). The exception

is Black male workers, who, if anything, are 1.9 p.p. more willing to continue (although this effect is not statistically different from zero).

Figure 4: Opt-in to Round 3 after a Round 1 sanction, by race and gender



Note: Each marker is the effect of one more token of Round 1 sanction on the probability of opting in to Round 3 (percentage points) when the sanction is implemented, relative to the No-Punishment arm where the same sanction is chosen but not implemented. The left panel shows the probability of opting in with the same manager, and the right panel with any manager, i.e., a different manager. The dark bar is the 90% confidence interval and the light bar the 95% confidence interval, from heteroskedasticity-robust standard errors.

Recall that it was men, in particular white men, who most sharply adjusted Round 2 effort on the intensive margin in response to Round 1 punishment. Interestingly, when we shift to extensive margin reactions, women are adjusting most sharply (though white men also exit). Women of both races and white men often take this costly but quiet option in response to being punished, while only white men shirk in a continuing relationship, risking additional punishment. Black men respond on neither margin.

One concern is that punishment could be reducing workers' willingness to participate in the experiment more broadly, for instance by lowering expected earnings or diminishing other-regarding preferences toward generic partners. To address this, we exploit a follow-up choice in which workers indicate whether they would be willing to play Round 3 with a different manager.

Column 2 of Table 7 shows that punishment has less impact on the probability of opting

in to Round 3 *with any manager*: the interaction coefficient for the punishment arm and the sanction amount are small and statistically insignificant (3.5pp, $p = 0.19$). Total willingness to play with a different manager does not decrease: it is 71% among workers who were not punished and 75% among those who were ($p = 0.34$).

The right panel of Figure 4 shows the results by subgroup (also presented in Table A33). Again, the overall results mask interesting heterogeneity. The higher the implemented sanction, the more likely male Black workers are to play with a different manager, by 4.5 p.p. This result, however, relies on a small group of Black workers who did not want to continue playing with the same manager.

Workers' responses are also the same whether their manager is a woman or a man (Appendix Table A14, Panels C and D).

4.3 Group-Level Analysis: Impacts of Punishment on Earnings and Efficiency

Having studied managers' and workers' decisions in isolation, we now ask what punishment does to the pair as a whole. For each group and round, we look at three outcomes: (i) Joint surplus, the total tokens the pair gains; (ii) Welfare, which nets out from the joint surplus the worker's cost of effort; and (iii) Share, the manager's share of the total surplus. Importantly, we measure all payoffs in expected pay, since only one of Rounds 1 and 2 is randomly chosen for payment. Round 3 is paid with certainty if the workers accept to do it, so its tokens count in full. The worker bears the cost of effort in every round, so effort costs always count in full. We compare the Punishment and No-Punishment arms, where, in the first round, managers face identical decisions, but only in one is the sanction implemented. We ask three questions. First, does punishment increase or reduce the pie? Then, who bears the difference? And finally, does observing effort change these answers?

4.3.1 Punishment reduces the pie without value creation

In Table A34, we look at both round 1 and 2 payoffs, with round fixed effects, and observe that punishment makes the pair worse off in both rounds. Relative to the No-Punishment treatment arms, the ability to punish lowers a group's joint surplus by 0.68 tokens per round when effort is hidden and by 0.88 when effort is known (both $p < 0.01$, not different from each other). The loss falls entirely on the worker, whose payoff drops by 0.79 tokens and 0.86 tokens, respectively, and the manager's own payoff is statistically unchanged (by 0.10 and -0.02 tokens, in the hidden and effort arms, respectively, both not statistically significant). Punishment falls only on the worker; the manager receives no direct reward, and it does not buy enough effort to offset the value it destroys. If punishment raised effort, the welfare loss would exceed the surplus loss, because welfare also subtracts the worker's cost of effort. It does not: across Rounds 1 and 2, punishment changes effort by less than half a slider in either arm, and the change is not statistically significant. The two are close in both arms (-0.59 vs. -0.68 tokens with hidden effort and -0.99 vs. -0.88 with known effort). Note that additional effort is socially optimal for a share of workers in this setting. Each slider raises the probability of a good outcome by one percentage point, which is worth 1 cent to the manager if the round is paid. In Rounds 1 and 2, the expected return is therefore 0.5 cents per slider. For 37.4% of workers, the elicited cost of completing one more slider is below this return. For 57% of workers, it is below the 1-cent return of Round 3, which is paid for sure. The median cost is 0.83 cents.[20] However, the elicited cost is likely an upper bound on the cost of effort during the task. In the price-list exercise, each extra slider costs both effort and time. In the task itself, time is not at stake because every worker waits the full 180 seconds regardless of how many sliders they complete, so the only cost of an additional slider is the effort it takes.

If we include Round 3 in the analysis, the surplus loss is similar (Table A35). Good

[20] If we exclude those who rejected all offers, then the median cost per slider is 0.67 cents, 40.8% of workers have a cost below 0.5 cents, and 62.5% have a cost below 1 cent.

outcomes are also about 5pp more frequent in Round 3 than in Round 1 ($p < 0.01$), which is consistent with selection into playing a third round with the same manager.

Cumulating surplus across all three rounds, among the positively selected groups that chose to continue to Round 3, punishment still lowers joint surplus by about 1.9 tokens when effort is hidden and by 3.4 tokens when it is known, a loss borne by the worker. The two arms differ significantly on the welfare measure ($p < 0.05$) and weakly on surplus ($p = 0.13$). Over the two mandatory rounds, the loss is smaller for these groups than in the full sample (0.9 against 1.4 tokens with hidden effort, and 1.3 against 1.8 with known effort), precisely because the groups most damaged by punishment are the ones whose workers opt out of Round 3, so these figures are a lower bound (Table A36).

The fines implemented in Rounds 1 and 2 destroyed roughly 1240 tokens, about \$124, or 12% of the potential group surplus in the Punishment arm. They led to no measurable increase in output. The loss in joint surplus is similar across workers. Splitting the sample by worker race and gender, every subgroup loses between 0.7 and 0.9 tokens of surplus per round, with no statistically significant differences across groups (Table A37). The loss in worker pay, however, is larger for Black workers (0.93 against 0.72 tokens per round, $p < 0.05$), consistent with their being punished more. The efficiency cost of punishment itself is therefore a general feature of the institution, but its burden falls more heavily on Black workers.[21]

In summary, punishment destroys value, and disciplined effort offers no offsetting efficiency gain.

4.3.2 Distributional consequences: relative gains and added risk

Because workers bear all the losses, managers' share of the total pie increases, but without an absolute gain. Punishment also makes group outcomes more unequal and more precarious. It lowers average surplus and increases its variability, increasing the share of zero-surplus

[21]The point estimate of the surplus loss is slightly larger for groups with a Black worker (-0.9 vs. -0.7 for a white worker), but the difference is not statistically significant.

outcomes. The standard deviation of joint surplus rises from 4.7 tokens in the No-Punishment arm to 6.4 under punishment, and the share of groups ending a round with zero-surplus rises from zero to around 8.7% (Table A38). Additionally, under punishment, revealing effort lowers the share of group-rounds at the maximum surplus of 20 by around 4 pp ($p = 0.11$); without punishment, there is no such drop. This motivates us to study the role of information on efficiency.

4.3.3 Neither observation nor time rescues punishment

A natural conjecture is that punishment is inefficient only because managers lack full information. Indeed, when managers observe effort, they target sanctions at low-effort workers (Table A39). However, this does not make it more efficient. Learning that their performance is observed makes workers exert more effort in the second round. However, it does so through information rather than punishment. In fact, punishment pushes in the opposite direction (Table A40). If anything, combining punishment with observing effort seems to be a more damaging institutional setting, although the difference is not statistically significant in Rounds 1 and 2 for the full sample. It is significant for welfare among the groups that continue to Round 3 (Table A41).

Taken together, the group-level evidence is unambiguous: in this setting the power to punish is close to a pure loss. It shrinks the surplus the pair produces, assigns the cost entirely to the worker, adds a tail of zero surplus, and neither observing effort nor repeating the interaction turns it to the group's advantage. Alongside the finding that punishment falls disproportionately on low-effort Black men, this makes the institution not only inefficient but inequitable.

5 Conclusion

In this paper, we study the incidence and impact of punishment in the context of a work relationship. Our controlled experiment varies exogenously whether managers' chosen sanction is implemented, and whether they observe the worker's effort prior to making the decision. Our setting allows us to study the downstream consequences for labor supply on the intensive and extensive margin, with particular attention to differences by race and gender. These questions are challenging to study in observational data, but have important implications for understanding the full costs and benefits of punishment at work.

Focusing on race, we observe that, conditional on the outcome, managers do not treat Black and white workers differently on average. But, we see important race disparities among workers observed to contribute relatively low effort. We find that managers sanction Black workers who did no sliders 1.4 tokens more than white workers that did no sliders. The gap comes mostly through the extensive margin, on whether to sanction or not, and it shrinks as effort rises, closing completely near the average effort level. Managers' stated expectations and standards are the same for Black and white workers, and the gap does not appear when effort is hidden, which is where a race-based inference about effort should matter most. What differs, then, is the response to identical information. We do not find this difference once managers have a second round of information about the worker.

Gender plays a different role. Managers do not punish women and men differently, at any level of effort. Gender matters instead for how workers react. Knowing that the manager can punish does not change first-round effort for any group. Being punished does. White men respond most strongly in effort, working harder when spared and less hard when sanctioned. Women and Black men change their effort much less. Women walk away when given the opportunity: after a sanction, their willingness to continue at all with the same manager drops sharply, from about 72% to 45% among white women, while their willingness to work with a new manager does not change. White men leave as well; Black men do not.

In this environment, punishment benefits no one. Beyond reducing workers' surplus, it

does not increase effort and so fails to increase managers' payoffs. Our results on labor supply further suggest that punishment may distort the talent pool, driving workers away from working relationships that feature punishment.

Our results point to three lessons about punishment in a working relationship. First, the threat of punishment and its use are very different instruments: the threat alone leaves labor supply unchanged, while using punishment changes it, mostly for the worse. Second, workers respond to punishment differently. The most common response is to leave when possible. White men also withdraw effort, and Black men do neither. Third, biases in who is punished do not come from either statistical discrimination or purely taste-based discrimination. In our setting, they appear only when a manager has observed poor performance and must judge it.

References

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *The Quarterly journal of economics*, 2016, *131* (4), 1753–1794.

_ , _ , _ , and _ , "Beliefs about gender," *American Economic Review*, 2019, *109* (3).

Boskin, Joseph, *Sambo: The rise and demise of an American jester*, Oxford University Press, 1988.

Bowles, Samuel and Sandra Polania-Reyes, "Economic incentives and social preferences: substitutes or complements?," *Journal of economic literature*, 2012, *50* (2), 368–425.

Carpenter, Jeffrey P, "The demand for punishment," *Journal of Economic Behavior & Organization*, 2007, *62* (4), 522–542.

Carvalho, Marcela, "Who gets the benefit of the doubt? CEO gender and news about firm performance," *Working paper*, 2025.

Coffman, Katherine B, Christine L Exley, and Muriel Niederle, "The role of beliefs in driving gender discrimination," *Management Science*, 2021, *67* (6).

_ , Scott Kostyshak, and Perihan O Saygin, "Choosing and using information in evaluation decisions," *arXiv preprint arXiv:2507.13798*, 2025.

Collins, Patricia Hill, *Black sexual politics: African Americans, gender, and the new racism*, Routledge, 2004.

Dana, Jason, Daylian M Cain, and Robyn M Dawes, "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organizational Behavior and human decision Processes*, 2006, *100* (2), 193–201.

_ , Roberto A Weber, and Jason Xi Kuang, "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 2007, *33* (1), 67–80.

Darling-Hammond, Sean, Michael Ruiz, Jennifer L Eberhardt, and Jason A Okonofua, "The dynamic nature of student discipline and discipline disparities," *Proceedings of the National Academy of Sciences*, 2023, *120* (17), e2120417120.

Egan, Mark, Gregor Matvos, and Amit Seru, "When Harry fired Sally: The double standard in punishing misconduct," *Journal of Political Economy*, 2022, *130* (5), 1184–1248.

Exley, Christine L, Oliver P Hauser, Molly Moore, and John-Henry Pezzuto, "Believed gender differences in social preferences," *The Quarterly Journal of Economics*, 2025, *140* (1), 403–458.

Falk, Armin and Michael Kosfeld, "The hidden costs of control," *American Economic Review*, 2006, *96* (5), 1611–1630.

_ , **Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde**, "Global evidence on economic preferences," *The Quarterly Journal of Economics*, 2018, *133* (4), 1645–1692.

Fehr, Ernst and Simon Gächter, "Reciprocity and economics: The economic implications of homo reciprocans," *European economic review*, 1998, *42* (3-5), 845–859.

_ **and** _ , "Cooperation and punishment in public goods experiments," *American Economic Review*, 2000, *90* (4), 980–994.

Ferrario, Beatrice and Marcela Carvalho, "Should She Stay or Should She Go? The Relationship Between CEO Gender, Firm Performance, and Turnover," *Working paper*, 2026.

Gneezy, Uri and Aldo Rustichini, "A fine is a price," *The journal of legal studies*, 2000, *29* (1), 1–17.

Holmstrom, Bengt, "Moral hazard in teams," *The Bell journal of economics*, 1982, pp. 324–340.

Hoppe, Eva I and Patrick W Schmitz, "Hidden action and outcome contractibility: An experimental test of moral hazard theory," *Games and Economic Behavior*, 2018, *109*, 544–564.

Kinsler, Josh, "Understanding the black–white school discipline gap," *Economics of Education Review*, 2011, *30* (6), 1370–1383.

Kube, Sebastian, Michel André Maréchal, and Clemens Puppe, "The currency of reciprocity: Gift exchange in the workplace," *American Economic Review*, 2012, *102* (4), 1644–1662.

_ , _ , **and** _ , "Do wage cuts damage work morale? Evidence from a natural field experiment," *Journal of the European Economic Association*, 2013, *11* (4), 853–870.

Lemons, J Stanley, "Black stereotypes as reflected in popular culture, 1880-1920," *American Quarterly*, 1977, *29* (1), 102–116.

Magistro, Beatrice and Morgan Wack, "Racial bias in fans and officials: Evidence from the Italian Serie A," *Sociology*, 2023, *57* (6), 1302–1321.

NHS England and NHS Improvement, "NHS Workforce Race Equality Standard: 2019 Data Analysis Report for NHS Trusts," Report Publishing approval number 001283, NHS England and NHS Improvement, London February 2020. Version 1, prepared by the WRES Implementation Team. Available at `https://www.england.nhs.uk/wp-content/uploads/2020/01/wres-2019-data-report.pdf`.

Nikiforakis, Nikos, "Punishment and counter-punishment in public good games: Can we really govern ourselves?," *Journal of Public Economics*, 2008, *92* (1-2), 91–112.

Ruebeck, Hannah, "Causes and consequences of perceived workplace discrimination," *Available at SSRN 4799864*, 2025.

Sarsons, Heather, "Interpreting signals in the labor market: evidence from medical referrals," *Working paper*, 2017.

Vianello, Michelangelo, Konrad Schnabel, Natarajan Sriram, and Brian Nosek, "Gender differences in implicit and explicit personality traits," *Personality and individual differences*, 2013, *55* (8), 994–999.

A Appendix

A.1 Matching Algorithm

Because the experiment requires two participants to play simultaneously, assignment depends on the pool of participants available at each point in time. When a participant selects a

dark skin tone, they are automatically assigned the role of the worker and placed into a mixed-group. Their arrival also increments a "mixed" queue counter, recording that a Black worker is now waiting for a white manager. When a participant selects the pale skin tone, the assignment depends on the current state of that queue — that is, on who has arrived before them. If at least one Black worker is currently queued, the pale participant is assigned the manager role in a mixed group. If no Black worker is queued, the pale participant is assigned to a same-race group and to a role within that group. Once a group is complete (worker and manager), they receive a treatment number drawn from a counter that cycles through the four treatment arms in sequence and considers the balance of group types within each treatment arm. If no complementary partner arrives before the matching timeout (3 minutes), the participant is recorded as unmatched, and participation ends.

Given the low number of Black individuals in Prolific, the later stage of recruitment, and the need to ensure half of the groups were mixed-race, we changed the allocation algorithm to prioritize mixed-race groups. A same-race pair is formed only under more restrictive conditions: at least two light-skinned participants must be waiting, and the first (i.e., longest-waiting, earliest-arrived) of them must have been waiting at least 2.5 minutes, and a quota condition on the ratio of all-white to mixed pairs must be satisfied. When these conditions hold, the two waiting white participants are paired, with the earliest-arrived assigned the role of worker and the next the role of manager, and both given the same treatment from a separate cycling counter. Any participant who remains unmatched beyond the overall matching timeout is recorded as unmatched. In this version, then, arrival order governs matching in real time — mixed pairs are formed as soon as the two required skin tones are jointly present, and all-white pairs are formed only after the earliest waiting white participant has waited a set interval and the quota permits.

A.2 Reasoning

All managers answer about Round 1 decisions, and only managers in the punishment treatment arms answer these questions about other rounds. Trivially, only managers in the punishment arm whose worker chose to play a third round answer the corresponding question. We asked a Large Language Model (LLM) to analyze these reasons blindly and classify them into categories. They are the following: Effort Evaluation (based on the manager's evaluation of the worker's effort/performance), Fairness (based on the manager's sense of fairness / a just outcome), Personal Gain (the manager's own potential gain or loss), Outcome-based (based on the task outcome, regardless of effort), Motivation (desire to motivate the worker to do better in the future), Unwarranted (the manager saw no reason or need to fine the worker), Empathy (understanding or relating to the worker's feelings and situation), Lack Of Control (recognition that the worker had limited control over the outcome), Arbitrary (the amount was chosen at random or without a stated reason), Trust (level of trust in the worker / giving them the benefit of the doubt), and Punishment as Incentive (belief that punishment itself can serve as an incentive to improve).[22] Given the similarity between Motivation and Punishment as an incentive, we combine those two reasons in the analysis. Importantly, each reason is assigned to one category. The two most common justifications are Effort Evaluation and Fairness, followed by Personal gain and Outcome-based (see Table A20).

Table A22 shows that managers' justifications align with their behavior.[23] Those who

[22]The taxonomy was built in two steps. First, a model (Llama 3.3 70B) read a random sample of 200 Round-1 explanations with no predefined list and proposed its own categories with definitions. This produced the nine categories other than Unwarranted and Arbitrary. Second, on reading the labeled statements, we found that explanations giving no reason at all ("it was a wild guess") had been forced into the closest substantive category, most often Lack of Control. We therefore added a tenth category for such statements. Inspection of that category showed it mixed two different things, managers who saw no reason to fine and managers who gave no reason for the amount, so we split it into Unwarranted and Arbitrary. All explanations from all rounds were then classified into the final eleven categories with a single model (gpt-oss-120b, used because the original model's endpoint had been retired) and a single prompt. The model never sees any participant identifier.

[23]The reference category in the table is Personal gain. We dropped trust from the analysis due to the small sample size (27).

motivate sanctioning with effort-based evaluation, the outcome they received, and the use of punishment as a tool to motivate are the ones who implement larger sanctions. Those who justify their punishment decisions with empathy for the worker, lack of control, and unwarranted punishment impose lower sanctions. They also seem to use the information they receive to justify their decisions. Compared to the HE arm, when managers observe effort, they are 24pp ($p < 0.01$) and 2.5pp ($p < 0.10$) more likely to use effort- and motivation-based reasoning, respectively, with corresponding reductions to other reasonings, as shown in Table A23.

If instead we restrict the analysis to where the racial gap is found (managers in the Known-Effort arm matched with a male worker), evaluating the gap at low effort, clear patterns arise. We observe a racial gap among managers who judged the worker's effort or the outcome. For managers who cite fairness, motivation, personal gain, or who saw no reason to fine, we observe no gap. The few who cite empathy show a reversed gap (Table A24). In sum, managers who frame their decision as an assessment of past or future performance sanction a Black worker who completed no sliders by 3.4 tokens more than a White worker, and the gap closes by 0.12 tokens per slider. Managers giving any other reason show no gap (Table A25). This pattern is absent among female workers, for whom there is no racial gap to begin with.

We also run a simple word-count exercise to examine managers' justifications in depth. We categorize justifications into three ideas: blame, encouragement, and softening, capturing the lazy-Black worker stereotype, the effectiveness of punishment, and the benefit of the doubt due to outcome uncertainty, respectively. Then, within each, we counted for the presence of a list of words and short phrases, such as: blame (lazy, nothing, unacceptable, deserve, didn't try, no effort), encouragement (encourage, motivate, hope, next round, do better), and softening (fair, first try, not his fault, felt bad, he tried). The full lists are in the note to Table A26. A justification counts as using an idea if it contains at least one word or phrase from the corresponding list. We then compare the share of justifications that use

each idea for Black and White workers, with a test of proportions. Again focusing on the population where the racial gap exists, we find suggestive evidence that blame language is more common for Black workers, and encouragement and softening words are more common for white workers (See Table A26). We are underpowered to detect consistent statistical significance, but the point estimates are meaningful, especially when the worker exerts less effort.

In sum, among managers who justify their decision as an evaluation of the worker's performance, we observe that lower-effort Black men are sanctioned more than white men with the same effort and outcome. Among managers who give any other reason, we find no racial gap. The exception is the few managers who cite empathy, for whom the gap goes in the opposite direction, although this estimate is noisy. The word-count exercise points in the same direction. We find suggestive evidence of more blame language for Black workers and more encouragement and softening language for white workers. However, we are underpowered, and these differences are not consistently statistically significant.

A.3 Tables and Figures

Table A1: Avatar's skin tone and race/ethnicity

	Chosen avatar skin tone			
Self-reported race	Dark brown	Brown	Pale	*N*
Black	34.9	61.5	3.6	714
White	0.8	2.2	96.9	2316
Other race	0.0	14.3	85.7	14
All	8.8	16.2	75.0	3044

Note: This table shows the rate at which each avatar skin tone is chosen within a given self-reported race.

Table A2: Sample

Stage	N	Share of started
Started the study (completed intro demographics)	6334	1.000
Matched into a manager–worker pair	4848	0.765
Pair not excluded under the attention rule	4229	0.668
Completed all rounds (no own/partner dropout)	3148	0.497
No non-binary participant in the pair	3044	0.481
Analysis sample	3044	
Managers	1522	
Workers	1522	
Not in the analysis sample (of 6334 started):		
Dropped out before being matched	18	0.003
Never matched to a partner	1468	0.232
Excluded under the pre-registered attention rule		
Failed two attention checks	321	0.051
Partner failed two attention checks	298	0.047
Attrition		
Dropped out	570	0.090
Partner dropped out	511	0.081
Pair includes a non-binary participant	104	0.016

Note: Sample construction from all participants who started the study to the analysis sample. The rows under "Not in the analysis sample" are mutually exclusive and sum to the difference between the number who started and the analysis sample; a participant's own status takes precedence over the partner's. Participants who failed two attention checks were removed under the pre-registered exclusion rule and are not counted as attrition. "Dropped out" are participants who left the study or timed out; "partner dropped out" includes partners who left after the interactive part of the study.

Table A3: Workers and Managers' Characteristics

	Managers (all)	Workers (all)
Demographics		
Age	43.54	40.77
	(13.31)	(13.32)
Male	0.393	0.415
Black, self-reported	0.013	0.456
White, self-reported	0.982	0.540
High school or less	0.216	0.183
Currently employed	0.749	0.758
Elicited preferences (1–5 scales)		
Risk tolerance	3.26	3.52
	(0.99)	(0.94)
Patience	3.93	3.98
	(0.76)	(0.74)
Altruism	4.04	4.10
	(0.79)	(0.75)
Trust	3.13	3.04
	(0.99)	(1.02)
Positive reciprocity	4.39	4.38
	(0.59)	(0.63)
Negative reciprocity	1.96	2.10
	(0.94)	(0.99)
Task measure		
Effort cost (MPL, cents/slider)	1.03	0.93
	(0.83)	(0.77)
Data quality		
Attention checks failed	0.37	0.40
	(0.48)	(0.49)
Comprehension mistakes	0.72	0.40
	(1.24)	(0.83)
Observations	1522	1522

Note: Baseline characteristics of the analysis sample (one observation per participant), by role, measured in the intro survey before role and treatment assignment. Continuous variables report the mean with the standard deviation in parentheses below; binary variables report the proportion.

Table A4: Workers' Characteristics by Race and Gender

	By race			By gender		
	White	Black	(B−W)	Female	Male	(M−F)
Demographics						
Age	43.95	37.58	-6.37***	40.89	40.59	-0.30
	(13.61)	(12.22)		(13.51)	(13.04)	
High school or less	0.210	0.156	-0.054***	0.180	0.189	0.009
Currently employed	0.720	0.795	0.075***	0.709	0.826	0.116***
Elicited preferences (1–5 scales)						
Risk tolerance	3.30	3.75	0.45***	3.42	3.68	0.26***
	(0.97)	(0.85)		(0.97)	(0.88)	
Patience	3.96	4.00	0.04	3.92	4.07	0.15***
	(0.70)	(0.78)		(0.73)	(0.74)	
Altruism	4.04	4.17	0.13***	4.15	4.04	-0.11***
	(0.77)	(0.73)		(0.74)	(0.75)	
Trust	3.15	2.92	-0.23***	3.02	3.05	0.03
	(0.98)	(1.05)		(1.03)	(1.01)	
Positive reciprocity	4.39	4.37	-0.02	4.41	4.34	-0.07**
	(0.59)	(0.66)		(0.60)	(0.66)	
Negative reciprocity	2.00	2.20	0.20***	2.03	2.21	0.19***
	(0.96)	(1.02)		(0.99)	(1.00)	
Task measure						
Effort cost (MPL, cents/slider)	0.97	0.89	-0.09**	0.93	0.93	0.01
	(0.74)	(0.79)		(0.79)	(0.75)	
Data quality						
Attention checks failed	0.36	0.44	0.08***	0.38	0.42	0.03
	(0.48)	(0.50)		(0.49)	(0.49)	
Comprehension mistakes	0.33	0.46	0.13***	0.35	0.45	0.10**
	(0.75)	(0.90)		(0.76)	(0.92)	
Observations	761	761		891	631	

Note: Workers' baseline characteristics of the analysis sample (one observation per participant), separate measured in the intro survey before role and treatment assignment. Continuous variables report the mean with the standard deviation in parentheses below; binary variables report the proportion. *Diff. (B−W)* and *Diff. (M−F)* are the black−white and male−female difference, respectively. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Balance across treatment arms

	Treatment arm				Joint
	NP×HE	NP×KE	P×HE	P×KE	p
Age	42.00	42.08	42.09	42.46	0.912
Male	0.403	0.420	0.412	0.379	0.406
Black, self-rep.	0.227	0.237	0.235	0.240	0.911
High school or less	0.205	0.204	0.187	0.203	0.778
Currently employed	0.739	0.763	0.763	0.748	0.630
Risk tolerance	3.40	3.31	3.42	3.44	0.070
Patience	3.91	3.98	3.99	3.94	0.162
Altruism	4.06	4.06	4.10	4.07	0.751
Trust	3.07	3.07	3.06	3.13	0.601
Positive reciprocity	4.39	4.38	4.40	4.36	0.656
Negative reciprocity	2.00	2.02	2.07	2.04	0.595
Effort cost (MPL)	0.98	0.97	0.99	0.98	0.964
Mixed-race pair	0.488	0.511	0.491	0.509	0.878
Observations	750	790	754	750	

Note: Mean of each baseline characteristic by treatment arm (NP: No punishment, P: Punishment; HE: Hidden effort, KE: Known effort), analysis sample, one observation per participant. Binary variables report the proportion. The *Joint p* is from a test that the four-arm means are equal, obtained by regressing the characteristic on the treatment-arm indicators with standard errors clustered by pair. *Mixed-race pair* is the share of participants assigned to a mixed-race group. The final row reports the number of observations per arm.

Table A6: Balance of Baseline Characteristics: Managers

	Treatment arm (Punishment × Known effort)				
	NP×HE	NP×KE	P×HE	P×KE	p
Male	0.36	0.40	0.41	0.39	0.524
	(0.48)	(0.49)	(0.49)	(0.49)	
Age	43.80	44.11	43.63	42.60	0.395
	(13.67)	(13.44)	(13.58)	(12.53)	
High school or less	0.22	0.21	0.21	0.22	0.945
	(0.42)	(0.41)	(0.41)	(0.42)	
Employed	0.74	0.76	0.76	0.74	0.835
	(0.44)	(0.43)	(0.43)	(0.44)	
US resident	0.96	0.96	0.96	0.96	0.996
	(0.20)	(0.20)	(0.20)	(0.19)	
Risk taking	3.21	3.23	3.32	3.28	0.354
	(0.99)	(0.98)	(1.00)	(1.01)	
Patience	3.93	3.91	3.99	3.90	0.315
	(0.76)	(0.74)	(0.72)	(0.83)	
Altruism	4.04	4.06	4.03	4.03	0.957
	(0.80)	(0.79)	(0.77)	(0.82)	
Trust	3.09	3.17	3.08	3.17	0.399
	(0.97)	(0.99)	(1.04)	(0.98)	
Good intentions	2.89	3.02	2.96	3.01	0.303
	(0.98)	(1.04)	(0.99)	(1.01)	
Positive reciprocity	4.39	4.38	4.43	4.34	0.268
	(0.58)	(0.56)	(0.58)	(0.62)	
Negative reciprocity	1.91	1.97	2.00	1.97	0.592
	(0.95)	(0.95)	(0.93)	(0.92)	
Effort cost (per slider)	1.00	1.00	1.10	1.03	0.292
	(0.85)	(0.79)	(0.81)	(0.86)	
Observations	375	395	377	375	

Note: Each cell reports the mean of a baseline (pre-treatment) characteristic, with the standard deviation in parentheses below. Baseline characteristics are measured in the intro survey, before role and treatment assignment. The columns report means across the four treatment arms (NP: No punishment, P: Punishment; HE: Hidden effort, KE: Known effort); the accompanying p-value is from a joint F-test that the four-arm means are equal, obtained by regressing the characteristic on the treatment-arm indicators with robust standard errors. The final row reports the number of observations.

Table A7: Balance of Baseline Characteristics: Workers

	Treatment arm (Punishment × Known effort)				
	NP×HE	NP×KE	P×HE	P×KE	p
Male	0.44	0.44	0.41	0.36	0.085
	(0.50)	(0.50)	(0.49)	(0.48)	
Age	40.20	40.04	40.55	42.32	0.092
	(13.22)	(13.21)	(12.55)	(14.18)	
High school or less	0.19	0.20	0.16	0.18	0.683
	(0.39)	(0.40)	(0.37)	(0.39)	
Employed	0.74	0.77	0.76	0.76	0.777
	(0.44)	(0.42)	(0.43)	(0.43)	
US resident	0.96	0.96	0.96	0.96	0.992
	(0.20)	(0.20)	(0.19)	(0.19)	
Risk taking	3.59	3.40	3.52	3.59	0.018
	(0.95)	(0.96)	(0.93)	(0.91)	
Patience	3.90	4.04	3.99	3.98	0.076
	(0.78)	(0.69)	(0.71)	(0.76)	
Altruism	4.08	4.06	4.17	4.11	0.226
	(0.78)	(0.73)	(0.75)	(0.74)	
Trust	3.06	2.97	3.04	3.08	0.505
	(1.04)	(1.01)	(1.01)	(1.03)	
Good intentions	2.88	2.84	2.92	2.90	0.709
	(1.05)	(1.04)	(1.02)	(1.02)	
Positive reciprocity	4.39	4.38	4.37	4.38	0.967
	(0.66)	(0.61)	(0.61)	(0.63)	
Negative reciprocity	2.09	2.07	2.13	2.11	0.847
	(1.01)	(0.97)	(0.96)	(1.04)	
Effort cost (per slider)	0.96	0.94	0.88	0.94	0.553
	(0.76)	(0.74)	(0.77)	(0.81)	
Observations	375	395	377	375	

Note: Each cell reports the mean of a baseline (pre-treatment) characteristic, with the standard deviation in parentheses below. Baseline characteristics are measured in the intro survey, before role and treatment assignment. The columns report means across the four treatment arms (NP: No punishment, P: Punishment; HE: Hidden effort, KE: Known effort); the accompanying p-value is from a joint F-test that the four arm means are equal, obtained by regressing the characteristic on the treatment-arm indicators with robust standard errors. The final row reports the number of observations.

Table A8: Sample Attrition by Treatment Arm

	Treatment arm					Joint
	NP×HE	NP×KE	P×HE	P×KE	All	p
Dropped out	0.146	0.116	0.147	0.130	0.135	0.090
Partner dropped out	0.129	0.105	0.137	0.112	0.121	0.053
Attrition (either)	0.275	0.221	0.283	0.242	0.256	0.061
Managers	0.277	0.221	0.283	0.242	0.256	0.060
Workers	0.274	0.221	0.284	0.242	0.255	0.063
Retained (completed the study)	0.725	0.779	0.717	0.758	0.744	0.061
Participants	1,082	1,052	1,069	1,026	4,229	

Note: Attrition among matched participants (those assigned to a pair and a treatment arm), by treatment arm (NP: No punishment, P: Punishment; HE: Hidden effort, KE: Known effort). Pairs in which a participant failed two attention checks are excluded under the pre-registered rule and are not part of this table. *Dropped out* is the share who left the study or timed out; *Partner dropped out* is the share lost because their partner did; a participant's own status takes precedence over the partner's. *Attrition (either)* and *Retained* sum to one. The *Joint p* tests equality of the share across arms (regression on the treatment-arm indicators, standard errors clustered by pair).

Table A9: Baseline Characteristics of Retained and Dropped-Out Participants

	Retained	Dropped out	Diff. p
Age	42.04	43.86	0.004
Male	0.397	0.367	0.162
Black, self-rep.	0.235	0.342	0.000
White, self-rep.	0.760	0.646	0.000
High school or less	0.201	0.209	0.690
Currently employed	0.752	0.735	0.401
Participants	3,148	570	

Note: Baseline characteristics measured in the intro survey, for participants who completed the study and participants who dropped out (left the study or timed out). Binary variables report the proportion. *Diff. p* is from a heteroskedasticity-robust regression of the characteristic on a dropped-out indicator. Participants excluded under the pre-registered attention rule, participants never matched to a partner, and those lost because their partner dropped out are not included.

Table A10: Summary Statistics: Punishment Decisions

Treatment arm	N	Mean	SD	% punishing	Mean | punished
No punishment × Hidden effort	375	1.31	2.63	26.1	5.01
No punishment × Known effort	395	1.74	3.11	31.6	5.51
Punishment × Hidden effort	377	1.38	2.98	22.0	6.25
Punishment × Known effort	375	1.72	3.01	32.8	5.24
All managers	1522	1.54	2.94	28.2	5.46

Note: This table shows the summary statistics of punishment decisions by treatment arms. N represents the number of observations, Mean is the average sanction managers chose, SD the standard deviation, and % punishing the percentage of managers who chose a strictly positive sanction. Mean | punished is the average sanction among managers who chose a strictly positive sanction.

Table A11: Unconditional Race and Gender Gaps in Punishment

Worker group	N	Mean sanction	SD	% punishing
By worker race				
White worker	761	1.39	2.80	26.0
Black worker	761	1.69	3.07	30.4
Difference (Black − White)		0.30**		4.3*
By worker gender				
Female worker	891	1.54	2.92	28.6
Male worker	631	1.54	2.98	27.6
Difference (Male − Female)		-0.00		-1.0

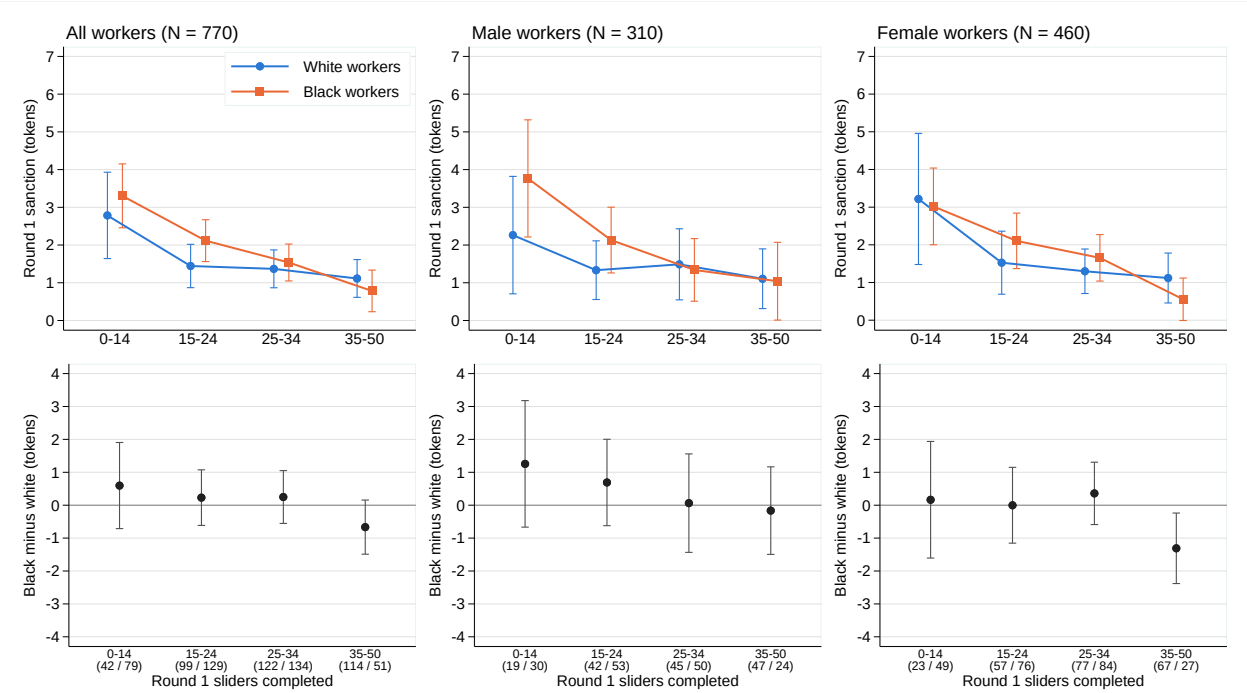
Note: This table reports unconditional Round-1 punishment by worker race and by worker gender, one observation per manager. N is the number of managers, Mean sanction is the average number of tokens deducted, SD is its standard deviation, and % punishing is the share who choose a strictly positive sanction. Difference rows report the Black−White and Male−Female gap in the mean sanction and in the share punishing (percentage points); asterisks denote statistical significance from a heteroskedasticity-robust regression of the sanction on the single group indicator: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Punishment Decisions: Effort, Gender and Race

	S_{g1}			
	Women	Men	Black	White
Black worker	0.862	2.172**		
	(0.782)	(0.949)		
Male worker			0.727	0.159
			(0.896)	(0.833)
Good outcome (G_{g1})	-1.877***	-1.654***	-1.729***	-1.896***
	(0.309)	(0.391)	(0.361)	(0.304)
Black × Good outcome	0.417	-0.040		
	(0.476)	(0.587)		
Male × Good outcome			0.072	0.077
			(0.547)	(0.478)
Effort (N_{g1})	-0.025	-0.012	-0.058***	-0.019
	(0.019)	(0.018)	(0.019)	(0.018)
Black × N_{g1}	-0.038	-0.075**		
	(0.027)	(0.031)		
Male × N_{g1}			-0.030	-0.006
			(0.032)	(0.026)
Constant	6.848***	0.400	4.849**	2.814
	(2.063)	(2.099)	(2.008)	(1.874)
Manager controls	Yes	Yes	Yes	Yes
Observations	460	310	393	377

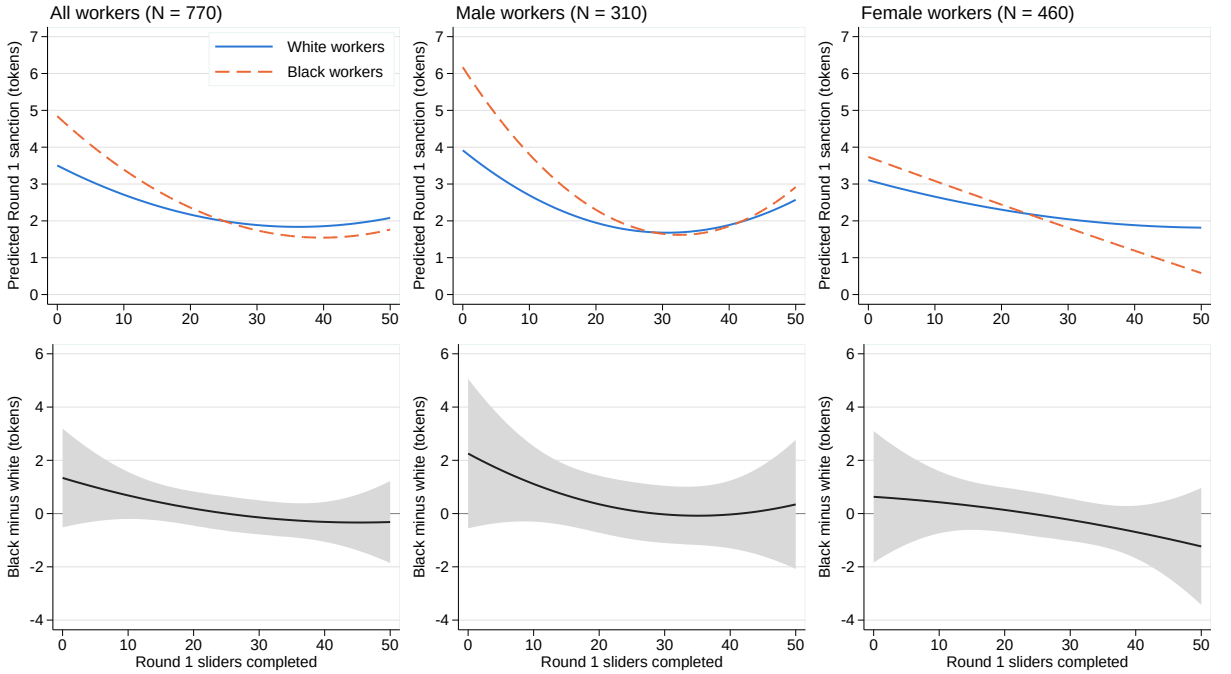
Note: The column-4 Known-Effort specification of Table 4, re-estimated within worker subgroups (women, men, Black, White; Known-Effort arm, one observation per manager). In the gender-split columns, the worker's race and its interaction with observed effort N_{g1} are shown (the gender indicator is constant within column and omitted); in the race-split columns, the worker's gender and its interaction with effort are shown. All columns condition on the outcome and its interactions, observed effort, and the full set of manager controls. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A1: Round 1 sanction by effort bin and worker race (Known-Effort arm)



Note: Known-Effort arm, Round 1, one observation per manager. The columns show all workers, male workers and female workers. Workers are grouped into four bins by the number of sliders completed in Round 1: 0–14, 15–24, 25–34 and 35–50. In the top panel is plotted the mean sanction S_{g1} (tokens) of white and Black workers in each bin, with 95% confidence intervals. In the bottom panel, it is the Black-minus-white difference in the sanction in each bin, from an OLS regression of S_{g1} on bin indicators, their interactions with the Black-worker indicator, the outcome, Black $\times$ Good outcome, and manager controls. The all-workers regression also includes the Male-worker indicator and its interactions with the outcome and with the bin indicators. Bars are 95% confidence intervals based on heteroskedasticity-robust standard errors. The numbers in brackets below each bin are the white / Black workers in that bin.

Figure A2: Round 1 sanction and worker race, quadratic in effort (Known-Effort arm)



Note: Known-Effort arm, Round 1, one observation per manager. The columns show all workers, male workers, and female workers. Estimates are from an OLS regression of the sanction S_{g1} (tokens) on the Black-worker indicator, observed effort N_{g1} and its square, the interactions of Black with N_{g1} and with its square, the outcome, Black $\times$ Good outcome, and manager controls. The all-workers regression also includes the Male-worker indicator and its interactions with N_{g1}, its square, and the outcome. In the top panel we plot the predicted sanction for white and Black workers at a bad outcome, with the other covariates at their sample means. In the bottom panel, the implied Black-minus-white difference in the sanction at a bad outcome. The shaded area is the 95% confidence interval, based on heteroskedasticity-robust standard errors.

Figure A3: Distribution of Sliders Completed in Round 1

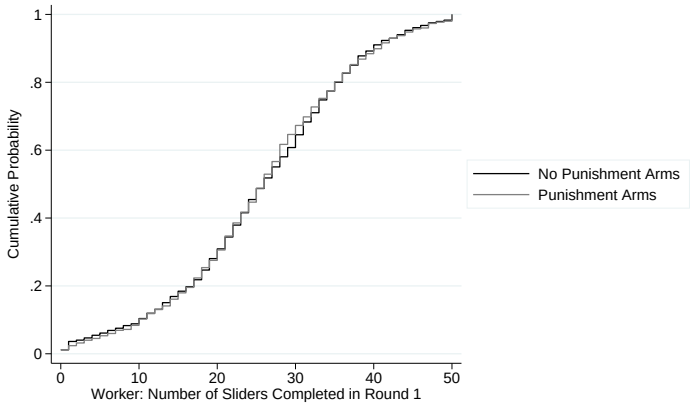


Table A13: Effort and Race: Intensive or Extensive Margins

	$\mathbf{1}[S > 0]$	$S \mid S > 0$	$\mathbf{1}[S = 10]$
Black worker	0.172**	1.378	0.085
	(0.085)	(0.959)	(0.057)
Black × Effort	-0.006**	-0.071*	-0.004*
	(0.003)	(0.036)	(0.002)
Effort (N_{g1})	-0.004	0.020	-0.000
	(0.002)	(0.033)	(0.002)
Male worker	0.006	0.403	0.015
	(0.083)	(0.923)	(0.060)
Good outcome (G_{g1})	-0.304***	-2.431**	-0.097***
	(0.047)	(0.965)	(0.024)
Constant	0.695***	4.352	0.212
	(0.219)	(2.793)	(0.149)
Mean dep. var.	0.322	5.379	0.078
Observations	770	248	770

Note: Decomposition of the Known-Effort-arm Round-1 sanction into three margins, one observation per manager. Column 1 is a linear probability model for any punishment ($\mathbf{1}[S_{g1} > 0]$); column 2 is OLS for the sanction amount conditional on punishing ($S \mid S > 0$); column 3 is a linear probability model for the maximum sanction ($\mathbf{1}[S_{g1} = 10]$). All columns use the fullest Known-Effort specification: worker race and gender, the outcome, observed effort N_{g1} and their interactions, plus manager controls. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: Main Results by Manager Gender

	Female manager	Male manager	p: equal
Panel A. Round-1 sanction S_{g1}, Known-Effort arm, all workers (N = 464, 306)			
Black worker	1.055	1.935*	0.490
	(0.742)	(1.043)	
Black × N_{g1}	-0.055**	-0.057*	0.955
	(0.026)	(0.034)	
Male worker	0.062	-0.076	0.916
	(0.779)	(1.053)	
Male × N_{g1}	-0.006	0.006	0.777
	(0.027)	(0.034)	
Panel B. Round-1 sanction S_{g1}, Known-Effort arm, male workers (N = 173, 137)			
Black worker	1.967	1.664	0.875
	(1.263)	(1.452)	
Black × N_{g1}	-0.097**	-0.036	0.360
	(0.048)	(0.047)	
Panel C. Round-2 worker effort (sliders), all workers (N = 924, 598)			
Punishment arm	0.706	1.282*	0.537
	(0.616)	(0.702)	
Pun. arm × Sanction	-0.477*	-0.311	0.686
	(0.245)	(0.331)	
Male worker	-0.833	-0.330	0.599
	(0.626)	(0.722)	
Black worker	0.496	1.920***	0.118
	(0.615)	(0.670)	
Panel D. Opts in to Round 3 with the same manager (0/1), all workers (N = 924, 598)			
Punishment arm	0.013	0.090**	0.166
	(0.035)	(0.044)	
Pun. arm × Sanction	-0.027**	-0.031**	0.830
	(0.011)	(0.013)	
Male worker	0.015	0.042	0.597
	(0.032)	(0.039)	
Black worker	0.024	0.020	0.937
	(0.032)	(0.040)	

Note: Key coefficients of the main specifications re-estimated separately for female and male managers (the manager's own gender). Panels A and B use the column-4 Known-Effort specification of Table 4 (all workers) and its male-worker version (Table A12), with the manager's gender dropped from the control set; Panel C uses column 5 of Table 6, adding indicators for the worker's race and gender; Panel D uses column 1 of Table 7. The last column is the p-value of a Wald test that the coefficient is equal across manager gender, from a pooled regression in which every regressor is interacted with the manager-gender indicator. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A15: Distribution of Managers' Beliefs and Perceived Effort Gaps (1)

Measure (sliders)	N	Mean	SD	p25	Median	p75
Manager: worker *should* do	1522	33.55	15.33	20	35	50
Manager: worker *will* do	1522	28.03	14.42	16	26	40
Expectations by worker race						
Should do, White worker	761	33.45	15.29	20	35	50
Should do, Black worker	761	33.65	15.38	20	35	50
Difference (Black − White)		0.19				
Will do, White worker	761	27.85	14.37	16	25	40
Will do, Black worker	761	28.20	14.48	18	29	40
Difference (Black − White)		0.35				
Expectations by worker gender						
Should do, Female worker	891	32.96	15.53	20	30	50
Should do, Male worker	631	34.39	15.01	25	35	50
Difference (Male − Female)		1.43*				
Will do, Female worker	891	28.16	14.41	15	30	40
Will do, Male worker	631	27.84	14.44	16	25	40
Difference (Male − Female)		-0.32				

Note: This table describes managers' Round-1 beliefs about worker effort, one observation per manager, all measured in sliders. It reports the manager's beliefs about how many sliders the worker *should* and *will* do, elicited from all managers before the outcome is realized, overall and separately by the worker's race and gender, with difference rows giving the Black−White and Male−Female gap in means; asterisks denote statistical significance from a heteroskedasticity-robust regression of the expectation on the single group indicator: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A16: Distribution of Managers' Beliefs and Perceived Effort Gaps (2)

Measure (sliders)	N	Mean	SD	p25	Median	p75
Manager's guess, Hidden Effort only ($\hat{N}_{g1}$)	752	22.15	14.15	10	20	30
if good outcome	258	29.52	15.34	15	30	45
if bad outcome	494	18.30	11.79	10	16	25
Guess by worker race						
Good outcome: White worker	142	28.25	15.66	15	25	45
Good outcome: Black worker	116	31.08	14.86	20	30	45
Good outcome: Difference (Black − White)		2.82				
Bad outcome: White worker	242	17.98	11.42	10	15	25
Bad outcome: Black worker	252	18.60	12.14	10	20	27
Bad outcome: Difference (Black − White)		0.61				
Guess by worker gender						
Good outcome: Female worker	143	28.96	15.12	15	28	45
Good outcome: Male worker	115	30.23	15.65	15	30	50
Good outcome: Difference (Male − Female)		1.27				
Bad outcome: Female worker	288	18.02	11.47	10	15	25
Bad outcome: Male worker	206	18.68	12.24	10	17	25
Bad outcome: Difference (Male − Female)		0.66				

Note: This table describes managers' Round-1 beliefs about worker effort, one observation per manager, all measured in sliders. It reports the manager's guessed effort ($\hat{N}_{g1}$), overall, and separately by outcome, worker's race and gender. Because the guess is elicited only in the Hidden-Effort arm, these rows are restricted to that arm and so have roughly half as many observations. Difference rows giving the Black−White and Male−Female gap in means; asterisks denote statistical significance from a heteroskedasticity-robust regression of the expectation on the single group indicator: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A17: Expectations

	Hidden-effort arm: S_{g1}				
	(1)	(2)	(3)	(4)	(5)
Good outcome (G_{g1})	-1.701***	-1.340***	-1.347***	-1.482***	-1.473***
	(0.163)	(0.164)	(0.165)	(0.179)	(0.179)
Should $-\ \hat{N}_{g1}$		0.034***	0.021**		
		(0.008)	(0.010)		
Will $-\ \hat{N}_{g1}$				0.022**	0.005
				(0.010)	(0.011)
Black worker	-0.028	0.001	-0.288	-0.004	-0.195
	(0.198)	(0.196)	(0.206)	(0.199)	(0.188)
Black $\times$ (Should $-\ \hat{N}_{g1}$)			0.025		
			(0.016)		
Black $\times$ (Will $-\ \hat{N}_{g1}$)					0.032*
					(0.018)
Male worker	0.065	0.050	0.037	0.108	0.079
	(0.200)	(0.197)	(0.197)	(0.201)	(0.201)
Punishment arm	0.093	0.099	0.106	0.109	0.099
	(0.197)	(0.194)	(0.194)	(0.197)	(0.197)
Constant	1.866***	1.347***	1.495***	1.624***	1.740***
	(0.212)	(0.233)	(0.228)	(0.238)	(0.233)
R^2	0.083	0.111	0.115	0.091	0.096
Observations	752	752	752	752	752

Note: OLS of the Round-1 sanction S_{g1} (tokens, 0-10) in the Hidden-Effort arm, where the manager does not observe effort and the benchmark is the manager's *guessed* effort $\hat{N}_{g1}$; one observation per manager. *Should* $-\ \hat{N}_{g1}$ is the normative shortfall (sliders the worker *should* do minus guessed effort) and *Will* $-\ \hat{N}_{g1}$ the expectational shortfall (sliders the worker *will* do minus guessed effort). Column 1 is a baseline (outcome, worker gender, punishment arm); columns 2-3 add the normative shortfall and its interaction with worker race; columns 4-5 replace it with the expectational shortfall and its race interaction. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A18: Expectations

	Known-effort arm: S_{g1}				
	(1)	(2)	(3)	(4)	(5)
Good outcome (G_{g1})	-1.784***	-1.654***	-1.654***	-1.764***	-1.765***
	(0.176)	(0.174)	(0.174)	(0.178)	(0.178)
Should $-$ N_{g1}		0.019***	0.012		
		(0.006)	(0.008)		
Will $-$ N_{g1}				0.003	0.001
				(0.006)	(0.008)
Black worker	0.415*	0.329	0.215	0.398*	0.390*
	(0.213)	(0.213)	(0.223)	(0.216)	(0.217)
Black $\times$ (Should $-$ N_{g1})			0.014		
			(0.012)		
Black $\times$ (Will $-$ N_{g1})					0.004
					(0.013)
Male worker	0.082	0.054	0.055	0.081	0.079
	(0.219)	(0.217)	(0.217)	(0.219)	(0.219)
Punishment arm	0.006	0.020	0.013	0.007	0.005
	(0.214)	(0.213)	(0.212)	(0.214)	(0.214)
Constant	2.015***	1.873***	1.915***	2.010***	2.012***
	(0.230)	(0.233)	(0.231)	(0.231)	(0.231)
R^2	0.078	0.090	0.092	0.078	0.078
Observations	770	770	770	770	770

Note: OLS of the Round-1 sanction S_{g1} (tokens, 0-10) in the Known-Effort arm, where the manager observes the worker's actual effort N_{g1}; one observation per manager. *Should* $-$ N_{g1} is the normative shortfall (sliders the worker *should* do minus observed effort) and *Will* $-$ N_{g1} the expectational shortfall (sliders the worker *will* do minus observed effort). Column 1 is a baseline (outcome, worker gender, punishment arm); columns 2-3 add the normative shortfall and its interaction with worker race; columns 4-5 replace it with the expectational shortfall and its race interaction. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A19: Belief revision about Round-2 effort and the Round-1 sanction

	$\text{Will}_{g2} - \text{Will}_{g1}$			
	Continuous sanction		Punished dummy	
	No punishment	Punishment	No punishment	Punishment
Sanction (S_{g1})	-0.035	-0.255		
	(0.620)	(0.458)		
Sanction × Black worker	0.322	0.290		
	(0.721)	(0.711)		
Punished ($S_{g1} > 0$)			-4.821	-2.133
			(3.431)	(3.543)
Punished × Black worker			6.436	0.369
			(4.783)	(5.186)
Black worker	-1.630	-5.425	-2.947	-5.131
	(2.501)	(3.277)	(2.643)	(3.396)
Good outcome (G_{g1})	-1.795	2.739	-2.497	2.552
	(2.454)	(2.965)	(2.494)	(2.904)
Round-1 effort (N_{g1})	0.616***	0.596***	0.603***	0.580***
	(0.092)	(0.121)	(0.093)	(0.118)
Constant	-34.196**	-20.931	-31.784**	-20.119
	(15.242)	(13.851)	(15.192)	(13.766)
Observations	174	136	174	136
R^2	0.307	0.362	0.315	0.364

Note: OLS of the change in the manager's stated expectation of the worker's effort between Round 1 and Round 2, $\text{Will}_{g2} - \text{Will}_{g1}$ (sliders), on the Round-1 sanction and its interaction with worker race; restricted to the Known-Effort arm and to managers matched with a male worker; one observation per manager. Columns 1 and 2 use the sanction S_{g1} (tokens, 0–10); columns 3 and 4 use an indicator for a positive sanction ($S_{g1} > 0$). Each pair is estimated separately for the No-Punishment arm. *Black worker* denotes the worker's avatar race (base: White). All columns condition on the Round-1 outcome G_{g1}, observed effort N_{g1}, and the full set of manager controls. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A20: Reasoning

Reason (LLM-induced)	N	% of answers
Effort evaluation	388	25.5
Fairness	238	15.6
Personal gain	215	14.1
Outcome-based	198	13.0
Motivation	115	7.6
Unwarranted	99	6.5
Empathy	94	6.2
Lack of control	93	6.1
Arbitrary	36	2.4
Trust	27	1.8
Punishment as incentive	19	1.2
All	1522	100.0

Note: Distribution of managers' Round-1 open-ended punishment rationales across categories generated bottom-up by a large language model (the model's own, data-driven taxonomy, distinct from the pre-specified categories used elsewhere). Each answer is assigned a single primary category, so the shares are mutually exclusive and sum to 100%. N is the number of answers in each category and % of answers is its share. One observation per manager's Round-1 rationale. The table covers the 1,522 managers in the analysis sample.

Table A21: Reasoning, by whether the manager sanctioned

	Not sanctioned ($S_{g1} = 0$)		Sanctioned ($S_{g1} > 0$)		
Reason (LLM-induced)	N	%	N	%	Diff. (pp)
Effort evaluation	244	22.3	144	33.6	+11.2***
Fairness	167	15.3	71	16.6	+1.3
Personal gain	186	17.0	29	6.8	-10.3***
Outcome-based	138	12.6	60	14.0	+1.4
Motivation	58	5.3	57	13.3	+8.0***
Unwarranted	95	8.7	4	0.9	-7.8***
Empathy	80	7.3	14	3.3	-4.1***
Lack of control	86	7.9	7	1.6	-6.2***
Arbitrary	9	0.8	27	6.3	+5.5***
Trust	24	2.2	3	0.7	-1.5**
Punishment as incentive	6	0.5	13	3.0	+2.5***
All	1093	100.0	429	100.0	

Note: Distribution of managers' Round-1 rationales across the same LLM-generated categories as Table A20, separately for managers who chose a sanction of zero ($S_{g1} = 0$) and managers who chose a positive sanction ($S_{g1} > 0$). Every manager, in both punishment arms, made a Round-1 sanction choice. Within each group, N is the number of answers in a category and % its share; shares sum to 100% within each group. The last column is the difference in shares, sanctioned minus not sanctioned, in percentage points, with a two-sample test of proportions. Rows are ordered by the overall share, as in Table A20. One observation per manager; pair-complete sample. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A22: Reasoning

	S_{g1}			
	(1)	(2)	(3)	(4)
Good outcome (G_{g1})	-1.776***	-2.009***	-1.995***	-2.050***
	(0.121)	(0.139)	(0.139)	(0.139)
Known effort arm	0.299**	-0.023	-0.028	-0.005
	(0.147)	(0.147)	(0.147)	(0.145)
Punishment arm	0.062	0.029	0.030	0.054
	(0.147)	(0.141)	(0.142)	(0.138)
Effort evaluation		1.446***	1.448***	1.376***
		(0.245)	(0.245)	(0.245)
Fairness		0.911***	0.896***	0.805***
		(0.245)	(0.245)	(0.241)
Outcome-based		1.663***	1.668***	1.642***
		(0.279)	(0.280)	(0.275)
Empathy		-0.422*	-0.426*	-0.458*
		(0.244)	(0.246)	(0.258)
Lack of control		-0.758***	-0.743***	-0.777***
		(0.224)	(0.224)	(0.225)
Motivation/incentive		1.428***	1.437***	1.376***
		(0.328)	(0.327)	(0.322)
Unwarranted		-0.471**	-0.463**	-0.418*
		(0.216)	(0.217)	(0.222)
Arbitrary		3.425***	3.426***	2.936***
		(0.607)	(0.610)	(0.624)
Black worker			0.220	0.108
			(0.140)	(0.140)
Male worker			0.030	0.027
			(0.143)	(0.140)
Constant	1.947***	1.356***	1.231***	1.785*
	(0.141)	(0.193)	(0.215)	(0.963)
Categories joint p		0.000	0.000	0.000
Manager controls	No	No	No	Yes
R^2	0.082	0.175	0.177	0.219
Observations	1,495	1,495	1,495	1,495

Note: OLS of the Round-1 sanction S_{g1} (tokens, 0-10) on indicators for the manager's (LLM-classified, bottom-up) reasoning category, one observation per manager, restricted to the nine sizable categories. Column 1 is a baseline (outcome and treatment-arm indicators); column 2 adds the reasoning-category fixed effects (base category: *Personal gain*); column 3 adds worker race and gender; column 4 adds manager controls. *Categories joint p* is an F-test that the categories differ. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A23: Reasoning

	Effort	Fair.	Outc.	Gain	Empat.	Control	Motiv/inc	Unwarr.	Arbit.
Known effort arm	0.244***	-0.068***	-0.047***	-0.053***	-0.037***	-0.040***	0.025*	-0.014	0.009
	(0.022)	(0.019)	(0.017)	(0.018)	(0.013)	(0.012)	(0.014)	(0.013)	(0.008)
Black worker	-0.009	0.043**	-0.016	0.007	0.006	-0.017	-0.018	-0.005	-0.004
	(0.022)	(0.019)	(0.017)	(0.018)	(0.013)	(0.013)	(0.015)	(0.013)	(0.008)
Male worker	0.022	0.002	-0.004	-0.017	0.004	-0.005	-0.007	-0.006	0.003
	(0.022)	(0.019)	(0.017)	(0.018)	(0.013)	(0.012)	(0.015)	(0.013)	(0.008)
Mean share (%)	25.5	15.6	13.0	14.1	6.2	6.1	8.8	6.5	2.4
Observations	1,522	1,522	1,522	1,522	1,522	1,522	1,522	1,522	1,522

Note: Linear probability models for the manager's use of each (LLM-classified, bottom-up) reasoning category; one regression per column, the dependent variable being a 0/1 indicator for that category, one observation per manager. Reported are the coefficients on the Known-Effort arm and on worker race and gender, each net of the others, the outcome, the punishment arm, and manager controls. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A24: Racial gap in Round-1 sanctions by managers' stated reason

Stated reason	N	Gap at zero effort	Gap at 10 sliders	Black × Effort
Effort evaluation	126	3.887***	2.404**	-0.148***
		(1.281)	(0.953)	(0.042)
Outcome	27	2.966	2.159	-0.081
		(1.970)	(1.478)	(0.070)
Personal gain	32	0.587	0.584	-0.000
		(1.733)	(1.170)	(0.067)
Empathy	17	-5.718**	-3.792**	0.193*
		(2.762)	(1.856)	(0.104)
Motivation	31	-0.137	0.304	0.044
		(3.383)	(2.315)	(0.116)
Unwarranted	20	1.598	0.974	-0.062
		(1.377)	(0.993)	(0.044)
Fairness	34	-1.442	-0.052	0.139*
		(2.027)	(1.350)	(0.084)
p-value: equal across all reasons		0.046		0.007
Other (categories with fewer than 15 managers, pooled)	23		estimated, not shown	
Observations	310			

Note: OLS of the Round-1 sanction S_{g1} (tokens, 0-10) on indicators for the manager's (LLM-classified, bottom-up) reasoning category, each interacted with the Black-worker indicator and with observed effort N_{g1}; restricted to the Known-Effort arm and to managers matched with a male worker. Controls include the Round-1 outcome G_{g1} and its interaction with race, the punishment arm, and the full set of manager controls. The p-value row reports Wald tests that the gap at zero effort (column 3) or the effort slope (column 5) is equal across all categories. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A25: Racial gap in Round-1 sanctions: performance reasons versus other reasons

	N	Gap at zero effort	Gap at 10 sliders	Black × Effort
Performance reasons	184	3.424***	2.211***	-0.121***
		(1.126)	(0.834)	(0.037)
Other reasons	126	0.027	0.146	0.012
		(1.126)	(0.761)	(0.044)
Difference		3.397**	2.066*	-0.133**
		(1.621)	(1.102)	(0.060)
Observations	310			

Note: OLS of the Round-1 sanction S_{g1} (tokens, 0–10) on an indicator for a performance-based reason, interacted with the Black-worker indicator and with observed effort N_{g1}; restricted to the Known-Effort arm and to managers matched with a male worker; one observation per manager. *Performance reasons* are the (LLM-classified, bottom-up) categories in which the manager judges the worker's performance, past or expected: effort evaluation, outcome, and motivation. *Other reasons* are fairness, empathy, personal gain, lack of control, trust, unwarranted (no reason to fine), and arbitrary (no reason for the amount). Each row reports, for managers in that group, the Black–White difference in the sanction at zero effort, at 10 sliders, and its change per additional slider; *Difference* is the performance-minus-other gap in each column. All quantities come from one regression that also conditions on the Round-1 outcome G_{g1} and its interaction with race, the punishment arm, and the full set of manager controls. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A26: Word-list check on the justifications for low-effort workers, by worker race

	White worker	Black worker	Difference	p-value
All *workers with 26 or fewer sliders (N = 68 White, 98 Black)*				
Blame / character words	0.147	0.224	+0.077	0.214
Encouragement / forward-looking words	0.191	0.102	-0.089	0.102
Softening / excuse words	0.324	0.286	-0.038	0.602
Fined *workers with 26 or fewer sliders (N = 23 White, 42 Black)*				
Blame / character words	0.130	0.262	+0.131	0.218
Encouragement / forward-looking words	0.304	0.190	-0.114	0.297
Softening / excuse words	0.391	0.190	-0.201*	0.078
All *workers with 12 or fewer sliders (N = 16 White, 23 Black)*				
Blame / character words	0.188	0.348	+0.160	0.274
Encouragement / forward-looking words	0.188	0.043	-0.144	0.145
Softening / excuse words	0.500	0.261	-0.239	0.126
Fined *workers with 12 or fewer sliders (N = 6 White, 15 Black)*				
Blame / character words	0.333	0.467	+0.133	0.577
Encouragement / forward-looking words	0.333	0.067	-0.267	0.115
Softening / excuse words	0.667	0.067	-0.600***	0.004

Note: Share of managers' written Round-1 justifications that contain at least one term from a fixed list, by worker race; restricted to the Known-Effort arm and to managers matched with a male worker, one observation per manager. The three lists correspond to the explanations of the gap considered in the text: *Blame/character* (lazy, laziness, nothing, unacceptable, deserve, didn't try, did not try, didn't even try, did not even try, no effort, lack of effort, poor effort, minimal effort, bare minimum, didn't care, did not care, not a team player, not serious, slack, accountable, consequence, punish), *Encouragement/forward-looking* (encourag-, motivat-, hope, hoping, next round, second round, try harder, do better, improve, incentiv-, serious moving, lesson, learn), and *Softening/excuse* (fair, first try, first round, first time, not his fault, not their fault, didn't want to, did not want to, feel bad, felt bad, empath-, kind, understand, he tried, they tried, effort was, did his best, did their best). Justifications are lower-cased, and each term is matched at the start of a word, so it also matches longer words that begin with it (for example, *punish* also matches *punished* and *punishment*). A hyphen marks a truncated stem, and *didn't* is matched with or without the apostrophe. Panels restrict the sample by the worker's Round-1 effort (26 or fewer sliders, the sample mean; 12 or fewer) and by whether the manager imposed a positive sanction (*Fined*). *Difference* is the Black-minus-White difference in shares; the p-value is from a two-sample test of proportions. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A27: Low-effort penalty in the Known-Effort arm, Round 1 vs. Round 2

	S_{gt}, KE arm	
	Round 1	Round 2
Black worker	1.799**	0.277
	(0.864)	(0.996)
Male worker	0.020	-0.772
	(0.901)	(1.068)
Good outcome	-1.941***	-1.073**
	(0.443)	(0.422)
Black × Good outcome	0.268	0.374
	(0.533)	(0.573)
Effort (N_{gt})	-0.011	-0.131***
	(0.026)	(0.022)
Black × Effort	-0.069**	-0.013
	(0.031)	(0.029)
Male × Effort	-0.003	0.025
	(0.033)	(0.030)
Constant	3.805*	6.169***
	(2.135)	(1.985)
Observations	375	375

Note: OLS of the Round-t sanction S_{gt} (tokens, 0-10) on worker race and gender, the Round-t outcome, observed Round-t effort N_{gt} (sliders completed), and the interactions of race and gender with the outcome and with effort, estimated separately for Round 1 and Round 2; one observation per manager. *Male* × *Good outcome* is included but not displayed. Sample restricted to the Known-Effort ∩ Punishment arm. All columns include manager controls (demographics, elicited preferences, the MPL effort cost and list fixed effects, and attention/comprehension counts). Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A28: Do managers credit effort improvements equally by race?

	Δ Sanction (R2−R1)	
	(1)	(2)
Δ Effort (R2−R1)	-0.082***	-0.137***
	(0.022)	(0.033)
Black worker	0.436	0.452
	(0.439)	(0.406)
Black × Δ Effort	-0.080*	-0.072
	(0.048)	(0.044)
Male worker		-0.860*
		(0.439)
Constant	0.375	-0.843
	(0.253)	(2.383)
Manager controls	No	Yes
Observations	375	375

Note: OLS of the change in the sanction (Round 2 − Round 1) on the change in observed effort and its interaction with worker race. Sample restricted to Known-Effort ∩ Punishment arm. Column 2 adds the worker's gender, its interaction with the change in effort (included but not displayed), and manager controls (demographics, elicited preferences, the MPL effort cost and list fixed effects, and attention/comprehension counts). Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A29: Manager's Reasoning and Traits and Differential Punishment

	Round-1 sanction (tokens)		
	Performance reasons	Traits	Both
Black worker	0.027	2.092**	0.120
	(1.126)	(0.975)	(1.174)
Black × Effort (N_{g1})	0.012	-0.073**	0.010
	(0.044)	(0.033)	(0.045)
Black × Performance reasoner	3.397**		3.143*
	(1.621)		(1.651)
Black × Performance reasoner × Effort	-0.133**		-0.127**
	(0.060)		(0.060)
Black × Negative reciprocity (z)		0.610*	0.698**
		(0.342)	(0.323)
Black × Altruism (z)		-0.061	-0.140
		(0.364)	(0.357)
Performance reasoner	2.312**		2.596**
	(1.173)		(1.215)
Performance reasoner × Effort	-0.046		-0.053
	(0.039)		(0.040)
Negative reciprocity (z)		-0.122	-0.202
		(0.211)	(0.215)
Altruism (z)		0.263	0.347
		(0.284)	(0.279)
Effort (N_{g1})	0.019	-0.015	0.021
	(0.029)	(0.018)	(0.030)
Good outcome (G_{g1})	-1.689***	-1.646***	-1.671***
	(0.400)	(0.392)	(0.397)
Observations	310	310	310
R^2	0.313	0.238	0.326

Note: OLS of the Round-1 sanction S_{g1} (tokens, 0-10); restricted to the Known-Effort arm and to managers matched with a male worker; one observation per manager. Column 1 interacts the Black-worker indicator and observed effort N_{g1} with an indicator for a performance-based stated reason (effort evaluation, outcome, or motivation, from the LLM classification of managers' explanations). Column 2 interacts the Black-worker indicator with two manager traits from the survey, negative reciprocity and altruism, each standardized within the sample. Column 3 includes both sets of interactions. All columns condition on the Round-1 outcome G_{g1} and its interaction with race, the punishment arm, and the full set of manager controls, with the raw altruism and negative-reciprocity scores replaced by their standardized versions in columns 2 and 3. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A30: Manager Traits and Differential Punishment

	Sanction (S_{g1})	Punished ($\mathbf{1}[S > 0]$)
Black worker	0.155	0.017
	(0.143)	(0.022)
Black × Male (z)	0.048	-0.003
	(0.146)	(0.023)
Black × Age (z)	-0.325**	-0.021
	(0.149)	(0.023)
Black × H.S. or less (z)	-0.022	-0.015
	(0.145)	(0.022)
Black × Employed (z)	-0.044	0.004
	(0.144)	(0.022)
Black × Risk-taking (z)	0.024	-0.008
	(0.156)	(0.023)
Black × Time disc. (z)	-0.188	-0.032
	(0.161)	(0.024)
Black × Altruism (z)	0.451***	0.051**
	(0.158)	(0.024)
Black × Trust (z)	-0.191	-0.002
	(0.227)	(0.034)
Black × Best intentions (z)	0.083	-0.016
	(0.211)	(0.034)
Black × Pos. reciprocity (z)	-0.101	0.006
	(0.157)	(0.024)
Black × Neg. reciprocity (z)	0.280*	0.058**
	(0.163)	(0.024)
Black × Effort cost per slider (z)	0.012	0.016
	(0.145)	(0.022)
Controls	Yes	Yes
Observations	1,522	1,522

Manager characteristics z-scored. Controls: outcome, KE arm, outcome×KE.

Note: OLS (column 1, sanction S_{g1} in tokens) and linear probability (column 2, any punishment $\mathbf{1}[S_{g1} > 0]$) of the Round-1 decision on worker race interacted *jointly* with twelve manager characteristics, one observation per manager, all arms pooled. Each manager characteristic is standardized (mean 0, SD 1); the effort cost per slider is standardized within the price list the manager saw (15, 30, or 45 sliders), because the same wage offers imply a mechanically lower per-slider cost on longer lists. All models control for the outcome, the Known-Effort arm, their interaction, and the twelve characteristic main effects (not shown). Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A31: Predictors of Effort in Round 1

Sliders Completed in Round 1			
	All	No punishment	Punishment
Black worker	-5.63***	-5.53***	-5.67***
	(0.58)	(0.83)	(0.83)
Male worker	0.96*	0.69	1.38*
	(0.56)	(0.78)	(0.82)
Age	-0.19***	-0.19***	-0.19***
	(0.022)	(0.031)	(0.032)
High school	3.70	9.05	0.23
	(4.26)	(6.76)	(4.90)
Trade/vocational	1.29	6.71	-2.74
	(4.30)	(6.79)	(5.04)
Associate (in prog.)	4.51	7.23	4.07
	(4.47)	(6.97)	(5.36)
Associate degree	1.60	5.77	-1.18
	(4.28)	(6.79)	(4.92)
Bachelor's (in prog.)	4.39	10.5	0.018
	(4.37)	(6.86)	(5.12)
Bachelor's degree	3.05	7.81	-0.26
	(4.23)	(6.72)	(4.84)
Advanced (in prog.)	1.87	5.41	-0.26
	(4.38)	(6.94)	(5.04)
Advanced degree	0.38	4.63	-2.46
	(4.25)	(6.75)	(4.88)
Job seeking	3.49***	2.23	4.26***
	(1.13)	(1.63)	(1.58)
Employed	0.47	0.27	0.48
	(0.81)	(1.12)	(1.14)
Male manager avatar	0.77	1.29	0.30
	(0.56)	(0.80)	(0.78)
Effort cost	-0.65*	-1.15**	-0.32
	(0.36)	(0.53)	(0.49)
Risk-taking	-1.42***	-0.99**	-1.91***
	(0.32)	(0.47)	(0.44)
Patience	0.50	0.54	0.42
	(0.38)	(0.53)	(0.54)
Altruism	-0.18	0.40	-0.89
	(0.42)	(0.62)	(0.55)
Trust	-0.51	-0.30	-0.64
	(0.40)	(0.51)	(0.62)
Good intentions	-0.47	-0.39	-0.69
	(0.39)	(0.50)	(0.63)
Positive reciprocity	1.77***	1.42**	2.40***
	(0.48)	(0.67)	(0.68)
Negative reciprocity	-0.55*	-1.13***	-0.055
	(0.28)	(0.40)	(0.39)
Failed attention check	-1.08*	-1.44*	-0.79
	(0.57)	(0.82)	(0.79)
Failed comprehension	-0.63**	-1.76***	-0.073
	(0.31)	(0.50)	(0.38)
Constant	33.8***	27.6***	38.7***
	(5.16)	(7.99)	(6.28)
Observations	1,522	770	752
Adjusted R^2	0.160	0.173	0.152

Note: OLS of Round 1 worker effort N_{g1} (sliders) across groups (one observation per group). Column 1 uses all groups; columns 2 and 3 split by whether punishment was possible (No-punishment vs. Punishment arm). All columns use heteroskedasticity-robust standard errors; standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A32: Round 2 effort responses to punishment, by Round-1 effort

	Round 2 Effort (Sliders Completed)		
	R1 effort ≤ 26	R1 effort > 26	Pooled
Punishment Arm	0.62	1.25*	0.64
	(0.67)	(0.65)	(0.67)
R1 Sanction Amt.	0.10	-0.63**	0.07
	(0.13)	(0.25)	(0.13)
Pun. Arm × Sanction Amt.	-0.35*	-0.42	-0.36*
	(0.19)	(0.39)	(0.19)
R1 effort above median			2.03**
			(0.98)
Above median × Pun. Arm			0.62
			(0.93)
Above median × Sanction Amt.			-0.68**
			(0.28)
Above median × Pun. Arm × Sanction Amt.			-0.09
			(0.44)
R1 effort	0.82***	0.68***	0.76***
	(0.05)	(0.06)	(0.04)
Good outcome	-0.18	-0.09	-0.13
	(0.67)	(0.67)	(0.47)
Known-effort arm	0.58	1.85***	1.22***
	(0.60)	(0.69)	(0.45)
Constant	8.28***	13.37***	8.92***
	(0.99)	(2.33)	(0.86)
Observations	797	725	1,522
Adjusted R^2	0.335	0.229	0.533

Note: OLS of Round 2 worker effort N_{g2} (sliders), one observation per group, using the specification of Column 5 of Table 6: the Punishment-arm indicator, the manager's Round 1 sanction, and their interaction, conditioning on Round 1 effort, the Round 1 outcome, and the Known-Effort arm. Columns 1 and 2 split the sample at the median of Round 1 effort (26 sliders). Column 3 pools both. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A33: Worker Responses to Punishment by Race and Gender: Four-Cell Specification

	Round 2 effort	Opted in to Round 3:	
	(sliders)	same manager	any manager
Punishment Arm	0.477	0.085*	0.107**
	(0.831)	(0.049)	(0.049)
R1 Sanction Amt.	-0.482*	0.009	0.004
	(0.252)	(0.011)	(0.011)
Pun. Arm × Sanction Amt.	0.043	-0.038**	-0.007
	(0.332)	(0.016)	(0.016)
Black worker	1.864	0.056	0.103
	(1.919)	(0.103)	(0.101)
Male worker	-0.878	0.045	0.051
	(2.494)	(0.113)	(0.111)
Black × Male	-1.421	-0.013	-0.042
	(3.177)	(0.153)	(0.150)
Black × Punishment Arm	-0.358	-0.030	-0.077
	(1.149)	(0.072)	(0.070)
Black × R1 Sanction Amt.	0.384	0.001	0.002
	(0.297)	(0.016)	(0.016)
Black × Pun. Arm × Sanction Amt.	-0.348	-0.015	0.007
	(0.442)	(0.022)	(0.021)
Male × Punishment Arm	1.615	-0.069	-0.110
	(1.457)	(0.077)	(0.075)
Male × R1 Sanction Amt.	0.417	-0.001	-0.004
	(0.407)	(0.018)	(0.018)
Male × Pun. Arm × Sanction Amt.	-1.575**	0.001	0.019
	(0.719)	(0.026)	(0.023)
Black × Male × Punishment Arm	-0.141	0.012	0.060
	(1.921)	(0.110)	(0.108)
Black × Male × R1 Sanction Amt.	-0.286	-0.029	-0.014
	(0.527)	(0.025)	(0.024)
Black × Male × Pun. Arm × Sanction Amt.	1.496*	0.071**	0.026
	(0.873)	(0.034)	(0.031)
Constant	8.115***	0.430***	0.468***
	(1.507)	(0.074)	(0.073)
Observations	1,522	1,522	1,522
Adjusted R^2	0.532	0.038	0.019
Implied Pun. Arm × Sanction Amt., white women	0.043	-0.038**	-0.007
	(0.332)	(0.016)	(0.016)
Implied Pun. Arm × Sanction Amt., Black women	-0.305	-0.053***	-0.000
	(0.291)	(0.015)	(0.015)
Implied Pun. Arm × Sanction Amt., white men	-1.532**	-0.037*	0.011
	(0.638)	(0.020)	(0.017)
Implied Pun. Arm × Sanction Amt., Black men	-0.384	0.019	0.045***
	(0.399)	(0.017)	(0.015)

Note: Specifications of Column 6 of Table 6 and Columns 2 and 4 of Table 7, adding Black × Male and its interactions with each regressor (one observation per group). The bottom rows report the implied Pun. Arm × Sanction Amt. coefficient for each race-by-gender cell, i.e., the sum of the relevant coefficients, which equals the estimate from a separate regression on that cell; delta-method standard errors in parentheses. Controls for Round 1 effort, the Round 1 outcome, and the Known-Effort arm, and their interactions with Black, Male, and Black × Male, are included but not displayed. All columns use heteroskedasticity-robust standard errors; standard errors displayed in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A34: Group Efficiency (R1 - R2)

	Joint surplus	Wkr pay	Mgr pay	Mgr share	Welfare
Punishment arm	-0.682***	-0.786***	0.104	0.033**	-0.588***
	(0.154)	(0.069)	(0.124)	(0.013)	(0.215)
Known effort arm	-0.014	0.000	-0.014	-0.001	0.040
	(0.123)		(0.123)	(0.012)	(0.188)
Punishment × Known effort	-0.196	-0.072	-0.124	-0.014	-0.397
	(0.217)	(0.098)	(0.175)	(0.018)	(0.298)
Round 2 FE	0.074	-0.047	0.122	0.017*	-0.264***
	(0.094)	(0.030)	(0.083)	(0.009)	(0.096)
Constant	6.610***	5.024***	1.586***	0.156***	4.240***
	(0.101)	(0.015)	(0.098)	(0.010)	(0.143)
R^2	0.020	0.121	0.001	0.004	0.016
Group-rounds	3,044	3,044	3,044	2,913	3,044

Note: OLS on the pooled Rounds 1–2 group-round panel (one observation per group per round, all arms). Dependent variables: *Joint surplus* (manager plus worker payoff, tokens), *Worker pay*, *Manager pay*, *Manager share* (manager's share of realized group earnings, defined when surplus is positive), and *Welfare* (surplus net of the worker's effort cost). Payoffs are in expected pay: tokens in Rounds 1 and 2 are weighted by one half, because only one of the two rounds is randomly chosen for payment; effort costs are not weighted, because they are incurred in both rounds. Regressors are the two treatment arms and their interaction (base: No-punishment × Hidden-effort) and a Round-2 fixed effect. Standard errors are clustered by group in parentheses. One token = \$0.10. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A35: Group Efficiency (R1 - R3)

	Joint surplus	Wkr pay	Mgr pay	Mgr share	Welfare
Punishment arm	-0.754***	-0.982***	0.228	0.039***	-0.617***
	(0.188)	(0.086)	(0.153)	(0.012)	(0.238)
Known effort arm	0.175	0.001	0.174	0.008	0.311
	(0.146)	(0.004)	(0.146)	(0.011)	(0.199)
Punishment × Known effort	-0.428	-0.066	-0.362*	-0.024	-0.714**
	(0.266)	(0.124)	(0.211)	(0.017)	(0.329)
Round 2 FE	0.074	-0.047	0.122	0.017*	-0.264***
	(0.094)	(0.030)	(0.083)	(0.009)	(0.096)
Round 3 FE	6.721***	4.583***	2.138***	0.032***	6.373***
	(0.190)	(0.072)	(0.162)	(0.010)	(0.192)
Constant	6.607***	5.118***	1.488***	0.151***	4.196***
	(0.115)	(0.024)	(0.110)	(0.009)	(0.153)
R^2	0.376	0.648	0.075	0.006	0.313
Group-rounds	4,018	4,018	4,018	3,837	4,018

Note: As Table A34, estimated on the Rounds 1-3 group-round panel with Round-2 and Round-3 fixed effects. Round-3 observations are limited to the groups that continued to Round 3 (a selected subsample), so the Round-3 rows condition on continuation. Tokens in Round 3, which is paid for sure, are weighted by one, so the Round-3 fixed effect also reflects the higher probability of payment. Dependent variables and regressors are otherwise identical; standard errors are clustered by group in parentheses. One token = \$0.10. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A36: Group Efficiency: Cumulative (R1-R3)

	Joint surplus	Wkr pay	Mgr pay	Welfare
Punishment arm	-1.882***	-2.788***	0.906	-1.223
	(0.716)	(0.319)	(0.591)	(0.886)
Known effort arm	0.619	0.000	0.619	1.603**
	(0.560)		(0.560)	(0.732)
Punishment × Known effort	-1.538	-0.142	-1.395*	-2.908**
	(1.009)	(0.461)	(0.810)	(1.220)
Constant	26.891***	20.000***	6.891***	18.908***
	(0.401)	(0.000)	(0.401)	(0.540)
R^2	0.031	0.140	0.003	0.026
Groups	974	974	974	974

Note: OLS across groups (one observation per group), restricted to groups that completed all three rounds. Dependent variables are per-group *cumulative* totals over Rounds 1–3: *Joint surplus*, *Worker pay*, *Manager pay*, and *Welfare* (surplus net of worker effort cost), in tokens. Totals are in expected pay: tokens in Rounds 1 and 2 are weighted by one half (only one of the two rounds is paid) and tokens in Round 3 by one; effort costs are counted in full in all three rounds. Regressors are the two treatment arms and their interaction (base: No-punishment × Hidden-effort). Heteroskedasticity-robust standard errors in parentheses. One token = \$0.10. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A37: Group Efficiency, by worker's characteristics (R1–R2 & R1–R3)

	R1–R2				R1–R3			
	Women	Men	Black	White	Women	Men	Black	White
Joint surplus	-0.733***	-0.828***	-0.890***	-0.671***	-0.845***	-1.130***	-1.149***	-0.792***
	(0.139)	(0.174)	(0.154)	(0.151)	(0.170)	(0.214)	(0.192)	(0.184)
Worker pay	-0.790***	-0.871***	-0.925***	-0.719***	-0.987***	-1.056***	-1.130***	-0.901***
	(0.061)	(0.082)	(0.073)	(0.064)	(0.078)	(0.102)	(0.091)	(0.084)
Manager pay	0.057	0.043	0.035	0.048	0.142	-0.074	-0.019	0.109
	(0.114)	(0.137)	(0.121)	(0.126)	(0.137)	(0.166)	(0.149)	(0.150)
Welfare	-0.723***	-0.866***	-0.792***	-0.787***	-0.848***	-1.159***	-1.074***	-0.883***
	(0.194)	(0.235)	(0.209)	(0.213)	(0.211)	(0.267)	(0.232)	(0.234)
Manager share	0.025**	0.030**	0.028**	0.024*	0.031***	0.023*	0.026**	0.028**
	(0.012)	(0.015)	(0.013)	(0.013)	(0.011)	(0.013)	(0.012)	(0.012)

Note: Each cell reports the coefficient on the Punishment arm from a separate regression of the group outcome (rows: joint surplus, worker pay, manager pay, welfare, and manager share) on the punishment and known-effort arm indicators and round fixed effects, estimated within the worker subgroup heading the column (Women, Men, Black, White). The two panels use the Rounds 1–2 panel and the Rounds 1–3 panel (Round-3 completers only). Payoffs are in expected pay, as in Tables A34 and A35. Standard errors clustered by group in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A38: Group Efficiency: distribution

Treatment arm	N	Mean	SD	% zero	% max
No punishment × Hidden effort	750	13.3	4.7	0.0	32.9
No punishment × Known effort	790	13.3	4.7	0.0	32.7
Punishment × Hidden effort	754	11.9	6.4	8.8	32.6
Punishment × Known effort	750	11.5	6.3	8.7	28.7
All	3044	12.5	5.6	4.3	31.7

Note: Descriptive distribution of joint group surplus (tokens) on the Rounds 1–2 group-round panel, by treatment arm. Surplus is the number of tokens the pair earns in a round if that round is chosen for payment. Since one of the two rounds is paid with equal chance, this is the distribution of the surplus in the paid round. N is the number of group-rounds; % zero and % max are the shares with zero surplus (a bad outcome with no offsetting) and with the maximum surplus of 20 (a good outcome with no fine implemented), respectively. One token = \$0.10.

Table A39: Punishment and effort

	S_{g1}	
	(1)	(2)
Effort (N_{g1})	-0.006	0.011
	(0.009)	(0.009)
Effort × Known effort	-0.056***	-0.058***
	(0.014)	(0.013)
Known effort arm	1.817***	1.811***
	(0.402)	(0.394)
Punishment arm	0.022	0.045
	(0.149)	(0.144)
Good outcome		-1.665***
		(0.122)
Constant	1.488***	1.611***
	(0.267)	(0.259)
R^2	0.033	0.098
Observations	1,522	1,522

Note: OLS of the Round-1 sanction S_{g1} (tokens, 0-10) across groups (one observation per group, all arms). Column 1 is the baseline; column 2 adds a good-outcome control. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A40: Effort Provision by treatment arm

	Round-2 effort N_{g2}		
	All groups	No punishment	Punishment
Known effort arm	1.278***	1.616**	0.936
	(0.459)	(0.653)	(0.643)
Round-1 effort	0.808***	0.830***	0.785***
	(0.024)	(0.028)	(0.039)
Round-1 sanction	-0.598***		-0.605***
	(0.157)		(0.160)
Punishment arm	1.228***		
	(0.450)		
Constant	8.368***	7.636***	10.364***
	(0.723)	(0.829)	(1.126)
R^2	0.528	0.528	0.529
Observations	1,522	770	752

Note: OLS of Round-2 worker effort N_{g2} (sliders) across groups (one observation per group). Column 1 uses all groups; columns 2 and 3 split by whether punishment was possible (No-punishment vs. Punishment arm). Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A41: Group Efficiency and Observing Effort

	Effort	Joint surplus	Welfare
A. Punishment×Known-effort by round — full sample (Rounds 1–2)			
Round 1	0.11	-0.18	-0.43
	(1.17)	(0.28)	(0.34)
Round 2	-0.79	-0.21	-0.37
	(1.32)	(0.29)	(0.37)
B. Punishment×Known-effort by round — Round-3 completers			
Round 1	-0.87	-0.09	-0.43
	(1.44)	(0.35)	(0.41)
Round 2	-0.37	-0.32	-0.80*
	(1.55)	(0.36)	(0.45)
Round 3	-0.45	-1.13	-1.69**
	(1.60)	(0.74)	(0.77)
C. Difference across rounds (completers, triple interaction; tokens if the round is paid)			
Round 3 − Round 1	0.42	-0.96	-1.17
	(1.20)	(0.93)	(0.92)
Round 3 − Round 2	-0.08	-0.50	-0.57
	(0.94)	(0.93)	(0.93)

Note: Each cell reports the *Punishment* × *Known-effort* interaction coefficient from a separate regression of the column outcome—worker effort (sliders), joint surplus (tokens), and welfare (surplus net of effort cost, tokens). Panel A estimates it round by round on the full Rounds 1–2 sample; Panel B does so for the groups that reach Round 3 (Round 3 is played only by these continuers); Panel C reports, from a round × punishment × known-effort triple-interaction model, how the Round-3 interaction differs from Rounds 1 and 2. In Panels A and B, payoffs are in expected pay, as in Tables A34 and A35. In Panel C, every round is measured in tokens if the round is paid, so that the rounds are in the same units. Panels A–B use heteroskedasticity-robust standard errors; Panel C clusters by group. Standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.4 Robustness: No Failed Attention Checks

Table A42: Robustness: Punishment Decisions and Effort, Participants with No Failed Attention Check

	(1)	(2)	(3)	(4)
Black worker	0.118	-0.175	1.306	1.042
	(0.370)	(0.353)	(0.809)	(0.769)
Male worker	-0.175	-0.165	0.559	0.633
	(0.379)	(0.358)	(0.806)	(0.747)
Good outcome	-1.940***	-1.697***	-1.845***	-1.803***
	(0.336)	(0.332)	(0.328)	(0.330)
Black × Good outcome	0.033	0.039	0.258	0.035
	(0.423)	(0.417)	(0.399)	(0.414)
Male × Good outcome	0.418	0.417	0.540	0.517
	(0.435)	(0.429)	(0.403)	(0.411)
Effort (sliders)		-0.065***	-0.026	-0.027
		(0.013)	(0.023)	(0.021)
Black × Effort			-0.061**	-0.055**
			(0.026)	(0.024)
Male × Effort			-0.029	-0.032
			(0.025)	(0.023)
Constant	2.109***	3.861***	2.837***	6.375***
	(0.317)	(0.474)	(0.725)	(1.885)
Manager controls	No	No	No	Yes
Observations	467	467	467	467

Note: Table 4 re-estimated on the sample of participants who failed no attention check, the exclusion rule in the original pre-registration (see Section 2). OLS of the Round-1 sanction S_{g1} (tokens, 0–10), Known-Effort arm, one observation per manager. Column 1 conditions on the outcome and its race and gender interactions; column 2 adds observed effort N_{g1}; column 3 adds *Black/Male* × N_{g1}; column 4 adds manager controls (as in the main text, minus the attention-check count, which is zero for all managers here). Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A43: Robustness: Effort, Gender and Race, Participants with No Failed Attention Check

	(1) Women	(2) Men	(3) Black	(4) White
Black worker	-0.140	2.835**		
	(1.033)	(1.207)		
Male worker			2.129*	-0.748
			(1.132)	(1.080)
Good outcome	-1.768***	-1.181***	-1.762***	-1.925***
	(0.372)	(0.430)	(0.416)	(0.370)
Black × Good outcome	0.111	-0.032		
	(0.550)	(0.669)		
Male × Good outcome			0.385	0.622
			(0.650)	(0.557)
Effort (sliders)	-0.047*	-0.044**	-0.061***	-0.042*
	(0.025)	(0.019)	(0.021)	(0.025)
Black × Effort	-0.019	-0.108***		
	(0.034)	(0.036)		
Male × Effort			-0.080**	0.003
			(0.037)	(0.031)
Constant	7.081**	6.686***	6.908***	4.983**
	(2.836)	(2.099)	(2.328)	(2.400)
Observations	279	188	225	242

Note: Table A12 re-estimated on the sample of participants who failed no attention check. The column-4 Known-Effort specification of Table A42, estimated within worker subgroups (women, men, Black, White; one observation per manager). In the gender-split columns, the worker's race and its interaction with observed effort N_{g1} are shown; in the race-split columns, the worker's gender and its interaction with effort are shown. All columns condition on the outcome and its interactions, observed effort, and manager controls (minus the attention-check count). Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A44: Robustness: Responses to Punishment by Race and Gender, Participants with No Failed Attention Check

	(1) White	(2) Black	(3) Men	(4) Women
Punishment arm	0.598	0.673	1.810*	-0.214
	(0.829)	(0.860)	(0.974)	(0.768)
R1 sanction	-0.204	-0.068	-0.054	-0.221
	(0.269)	(0.213)	(0.312)	(0.178)
Punishment × sanction	-0.987*	-0.212	-0.948*	-0.173
	(0.520)	(0.324)	(0.495)	(0.312)
R1 effort	0.809***	0.860***	0.820***	0.821***
	(0.044)	(0.044)	(0.048)	(0.043)
Good outcome	-0.862	1.219	-0.358	0.479
	(0.859)	(0.810)	(0.974)	(0.785)
Known-effort arm	0.965	0.425	0.563	1.026
	(0.865)	(0.795)	(0.941)	(0.776)
Male worker	-0.400	0.805		
	(0.889)	(0.812)		
Black worker			1.084	-0.381
			(0.998)	(0.779)
Constant	9.650***	7.014***	8.096***	8.885***
	(1.632)	(1.415)	(1.899)	(1.435)
Observations	488	428	368	548

Note: Table 6 re-estimated on the sample of workers who failed no attention check. OLS of Round 2 worker effort N_{g2} (sliders), one observation per group. Columns 1 and 2 present the results for white and Black workers, respectively; columns 3 and 4 for men and women. All columns control for Round 1 effort, the Round 1 outcome, and the Known-Effort arm. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A45: Robustness: Willingness to Continue with the Same Manager, Participants with No Failed Attention Check

	(1) All	(2) White	(3) Black	(4) Men	(5) Women
Punishment arm	0.058	0.048	0.070	0.028	0.089*
	(0.035)	(0.047)	(0.054)	(0.055)	(0.046)
R1 sanction	0.001	-0.001	0.003	-0.006	0.007
	(0.008)	(0.012)	(0.011)	(0.012)	(0.011)
Punishment × sanction	-0.019	-0.023	-0.017	-0.001	-0.032**
	(0.012)	(0.019)	(0.015)	(0.018)	(0.016)
R1 effort	0.006***	0.006***	0.006**	0.004*	0.008***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Good outcome	0.038	0.067	-0.002	0.108**	-0.019
	(0.035)	(0.048)	(0.054)	(0.053)	(0.047)
Known-effort arm	0.004	0.005	0.003	0.027	-0.018
	(0.032)	(0.043)	(0.048)	(0.050)	(0.042)
Male worker	0.039	0.006	0.075		
	(0.032)	(0.044)	(0.048)		
Black worker	0.021			0.055	0.002
	(0.033)			(0.052)	(0.042)
Constant	0.419***	0.424***	0.429***	0.468***	0.385***
	(0.055)	(0.070)	(0.075)	(0.083)	(0.071)
Observations	916	488	428	368	548

Note: Table 7 re-estimated on the sample of workers who failed no attention check. Linear probability models of the worker's decision to opt in to Round 3 with the same manager, one observation per group. Column 1 presents all workers; columns 2 and 3 present the results for white and Black workers, respectively; columns 4 and 5 for men and women. All columns control for Round 1 effort, the Round 1 outcome, and the Known-Effort arm. Heteroskedasticity-robust standard errors in parentheses. Asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.