Pre-Registration Document
Can't Agree to Disagree? Fairness Concerns and Conflict

Ernesto Reuben* Nishtha Sharma†

June 2026

Abstract

This research aims to understand how fairness concerns mediate the impact of income inequality on social conflict. Specifically, we will investigate whether disagreement over fairness perceptions intensifies conflict. Our approach will combine a theoretical model with field evidence and an online experiment to generate insights into the interplay between fairness, inequality, and conflict. In the experiment, participants will be assigned incomes based on merit or luck, with treatments involving information about the allocation rule or not. The experiment will employ a between-subjects design and will collect data on redistribution preferences, contest expenditures, and fairness perceptions along with a demographic survey.

Keywords: fairness, inequality, conflict, redistribution, experiment

*New York University Abu Dhabi; ereuben@nyu.edu
†Charles University; nishthasharma.econ@gmail.com

1 Introduction

We examine how concerns about fairness influence the impact of income inequality on social conflict. While some studies suggest that inequality exacerbates conflict, others find no significant effects. We reconcile these discrepancies by introducing and testing a fairness mechanism that accounts for varying perceptions of income allocation. Including fairness concerns in the model of conflict over the right to redistribute predicts that conflict arises not merely from inequality but from disagreement over whether such inequality is merit-based (and thus fair) or luck-based (and thus unfair). If inequality is universally perceived as either fair or unfair, conflict is minimal as income allocation is either accepted or peacefully renegotiated. However, when there is disagreement about perceptions of the source of inequality, a conflict ensues over the right to redistribute.

We test our theoretical predictions using an online experiment. We first run a separate standalone study in which a sample of participants solves the 10 Raven's puzzles only; the median performance in this benchmark sample establishes a fixed cutoff used to classify high and low performers in the main study. In the main study, Part 1 (the Raven's puzzles and a demographic survey) and Part 2 are combined into a single continuous session. A participant's performance ranking is determined in real time by comparing their score against the fixed benchmark median (participants exactly at the median are equally likely to be classified as high or low). In Part 2, participants receive high or low incomes based on either luck or merit. If income is luck-based, each participant is equally likely to receive good or bad luck, corresponding to high or low income, respectively. If income is merit-based, above-median performers in the Raven's puzzles are allocated high income, while below-median performers receive low income. In addition to varying the income allocation rule (luck or merit) to examine fairness, we randomly vary whether information about the allocation rule (luck or merit) is provided to participants, allowing us to exogenously vary agreement and disagreement about fairness. When the income allocation rule is unknown, self-serving biases may lead high-earners to perceive the rule as merit-based and fair, while low-earners may perceive it as luck-based and unfair.

The experiment employs a between-subjects design with four treatments (luck known, merit known, luck unknown, merit unknown). Each participant is randomly and anonymously matched with six other participants assigned to the same treatment, forming pairs of equal or unequal incomes. In each round, participants report their preferred income redistribution between themselves and their matched partner. After reporting their preferences, they engage in a Tullock contest game to determine which redistribution will be implemented. Participants choose their contest spending and

predict their partner's contest spending in each match. Finally, one of the six rounds is randomly selected for the actual payment. The experiment is asynchronous and effectively uses the strategy method to gather data.

2 The Conceptual Framework

2.1 Initial Allocation and Fairness Preferences

Consider two workers, H (high earner) and L (low earner), who together produce an output $\Pi > 0$. The initial allocation assigns H a larger share than L, i.e., $a_H^{\text{Pre}} > \Pi/2$. Each worker $i \in \{H, L\}$ holds a subjective belief about what constitutes a fair allocation for H, denoted $a_H^{i,\text{Fair}}$. The corresponding fair share for L under i's belief is $a_L^{i,\text{Fair}} = \Pi - a_H^{i,\text{Fair}}$.

Workers derive utility from income and disutility from deviations between actual and fair income. Let $\beta_i \in [0, \infty)$ denote the weight worker i assigns to fairness concerns, normalized such that the weight on own income is 1. Then, utility from income a_i is:

$$U_i(a_i) = a_i - \beta_i \frac{(a_i - a_i^{i,\text{Fair}})^2}{2\Pi} \tag{1}$$

Maximizing $U_i(a_i)$ with respect to a_i yields worker i's most preferred income:

$$a_i^{i*} = a_i^{i,\text{Fair}} + \frac{\Pi}{\beta_i} \tag{2}$$

Thus, their preferred income for the other worker is:

$$a_j^{i*} = \Pi - a_i^{i*} = a_j^{i,\text{Fair}} - \frac{\Pi}{\beta_i} \tag{3}$$

Workers' utility at their preferred allocation is:

$$U_H^{H*} = a_H^{H,\text{Fair}} + \frac{\Pi}{2\beta_H} \tag{4}$$

$$U_L^{L*} = a_L^{L,\text{Fair}} + \frac{\Pi}{2\beta_L} = (\Pi - a_H^{L,\text{Fair}}) + \frac{\Pi}{2\beta_L} \tag{5}$$

Utility under the other player's preferred allocation is:

$$U_H^{L*} = a_H^{L,\text{Fair}} - \frac{\Pi}{\beta_L} - \frac{\beta_H}{2\Pi}\left(a_H^{L,\text{Fair}} - \frac{\Pi}{\beta_L} - a_H^{H,\text{Fair}}\right)^2 \tag{6}$$

$$U_L^{H*} = (\Pi - a_H^{H,\text{Fair}}) - \frac{\Pi}{\beta_H} - \frac{\beta_L}{2\Pi}\left(a_H^{L,\text{Fair}} - \frac{\Pi}{\beta_H} - a_H^{H,\text{Fair}}\right)^2 \tag{7}$$

2.2 Conflict and Post Allocation

If either worker initiates a conflict, they expend effort x_H and x_L, which determine the probability of winning according to the Tullock contest success function:

$$p_H = \frac{x_H}{x_H + x_L}, \qquad p_L = \frac{x_L}{x_H + x_L} \tag{8}$$

The winner enforces their preferred allocation. Expected utility from conflict is:

$$EU_H = p_H U_H^{H*} + (1 - p_H)U_H^{L*} - x_H \tag{9}$$

$$EU_L = p_L U_L^{L*} + (1 - p_L)U_L^{H*} - x_L \tag{10}$$

Let $V_H = U_H^{H*} - U_H^{L*}$ and $V_L = U_L^{L*} - U_L^{H*}$. Substituting the utilities from equations 4–7 gives the explicit values each worker derives from winning the conflict:

$$V_H = (a_H^{H,\text{Fair}} - a_H^{L,\text{Fair}}) + \frac{\Pi}{2\beta_H} + \frac{\Pi}{\beta_L} + \frac{\beta_H}{2\Pi}\left(a_H^{L,\text{Fair}} - a_H^{H,\text{Fair}} - \frac{\Pi}{\beta_L}\right)^2 \tag{11}$$

$$V_L = (a_H^{H,\text{Fair}} - a_H^{L,\text{Fair}}) + \frac{\Pi}{2\beta_L} + \frac{\Pi}{\beta_H} + \frac{\beta_L}{2\Pi}\left(a_H^{L,\text{Fair}} - a_H^{H,\text{Fair}} - \frac{\Pi}{\beta_H}\right)^2 \tag{12}$$

Substituting yields:

$$EU_H = U_H^{L*} + \frac{x_H}{x_H + x_L}V_H - x_H \tag{13}$$

$$EU_L = U_L^{H*} + \frac{x_L}{x_H + x_L}V_L - x_L \tag{14}$$

Maximizing with respect to effort gives best responses:

$$\frac{x_L}{(x_H + x_L)^2}V_H - 1 = 0 \tag{15}$$

$$\frac{x_H}{(x_H + x_L)^2}V_L - 1 = 0 \tag{16}$$

Solving yields equilibrium efforts:

$$x_H^* = \frac{V_H^2 V_L}{(V_H + V_L)^2} \tag{17}$$

$$x_L^* = \frac{V_L^2 V_H}{(V_H + V_L)^2} \tag{18}$$

Equilibrium winning probabilities:

$$p_H^* = \frac{V_H}{V_H + V_L}, \qquad p_L^* = \frac{V_L}{V_H + V_L} \tag{19}$$

The value that each worker derives from winning the conflict, denoted V_H and V_L, increases with the degree of disagreement over what constitutes a fair allocation. Letting $d = a_H^{H,\text{Fair}} - a_H^{L,\text{Fair}}$ denote the gap between the workers' fairness ideals, differentiating equations 11 and 12 gives $\partial V_H/\partial d = 1+\beta_H/\beta_L+\beta_H d/\Pi$ and $\partial V_L/\partial d = 1+\beta_L/\beta_H+\beta_L d/\Pi$, both of which are strictly positive whenever $d > -\Pi(1/\beta_H + 1/\beta_L)$. Since this threshold is negative, both values are strictly increasing in the gap for all $d \geq 0$, i.e. whenever the high-income worker self-servingly claims a weakly larger fair share for herself than the low-income worker grants her.[1] In particular, when the high-income and low-income workers have diverging fairness ideals—i.e., when $|a_H^{H,\text{Fair}} - a_H^{L,\text{Fair}}|$ is large—both players stand to gain more by enforcing their own preferred outcome. Conversely, when both workers agree on what is fair and care equally about fairness (i.e., $a_H^{H,\text{Fair}} = a_H^{L,\text{Fair}}$ and $\beta_H = \beta_L$), they assign equal value to winning the conflict: $V_H = V_L$.

Proposition 1. *Conditional on entering conflict, in equilibrium:*

1. *Conflict expenditures x_H^*, x_L^* increase in the difference between fairness ideals (for $d \geq 0$).*
2. *Conflict expenditures are independent of the initial allocation.*

Substitute x_H^*, x_L^* into expected utility:

$$EU_H^* = U_H^{L*} + \frac{V_H^3}{(V_H + V_L)^2} \tag{20}$$

$$EU_L^* = U_L^{H*} + \frac{V_L^3}{(V_H + V_L)^2} \tag{21}$$

[1] Monotonicity can fail only when d is large and negative—that is, when the low earner believes the high earner deserves substantially *more* than the high earner claims for herself. This "reverse" disagreement lies well outside the empirically relevant range (e.g., with $\Pi = 120$ and $\beta_H = \beta_L = 1$, it would require $d < -240$), and does not arise in our design.

Conflict occurs if and only if at least one player has the incentive to enter conflict, i.e., at least one of inequalities 22 and 23 hold. Note that while equilibrium expenditures (conditional on conflict) are independent of the initial allocation, whether conflict occurs at all does depend on it through these conditions.

$$EU_H^* > a_H^{\text{Pre}} - \beta_H \frac{(a_H^{\text{Pre}} - a_H^{H,\text{Fair}})^2}{2\Pi} \tag{22}$$

$$EU_L^* > (\Pi - a_H^{\text{Pre}}) - \beta_L \frac{(a_H^{\text{Pre}} - a_H^{L,\text{Fair}})^2}{2\Pi} \tag{23}$$

2.3 Causes of Disagreement in Fair Allocation

The model developed above demonstrates that conflict emerges not from inequality per se, but from disagreement over what constitutes a fair allocation. This disagreement may arise through two distinct channels: normative differences in fairness preferences, or divergent beliefs about the source of inequality.

Preferences The first channel reflects heterogeneity in distributive preferences. Some individuals may hold egalitarian ideals and support equal splits regardless of productivity, while others may endorse meritocratic or needs-based principles. Experimental studies consistently show that individuals differ in their fairness ideals and that such heterogeneity affects both redistribution choices and reactions to unequal outcomes (Konow, 2000; Cappelen et al., 2007; Almås et al., 2010). These differences in fairness preferences imply that, even when both individuals fully agree on how the initial allocation was generated, they may still disagree on whether it is fair.

Beliefs In contexts where the allocation rule is ambiguous, disagreement about fairness may be driven by differences in beliefs about how the allocation was determined. For example, disagreement may result from different informational signals. If individuals receive conflicting signals about the allocation rule, they may arrive at different conclusions about the source of the observed inequality and hence different fairness ideals. Even when individuals receive the same signals, they may update their beliefs differently due to biases such as base-rate neglect (Cappelen et al., 2024), overreliance on private signals, or motivated Bayesian reasoning (Valero, 2022; Deffains et al., 2016; Cassar and Klein, 2019; Fehr and Vollmann, 2025). These belief updating differences can independently generate disagreement over the source of inequality, even when fairness preferences and information sets are otherwise identical. Our experimental design to study the effect of disagreement in fairness ideas on preference for conflict leverages this channel directly. By withholding information about the allocation rule, we create ambiguity that may activate self-serving attributions and produce disagreement.

3 Experimental Design and Procedures

We test the theoretical predictions developed in the previous section using an incentivized online experiment. The experiment is designed to causally identify the effect of income inequality and the role of fairness disagreement on conflict behavior. Participants are recruited via Prolific. The experiment proceeds in two studies: a standalone benchmarking study and a main study.

We begin with a standalone study that recruits a separate sample of approximately 100 participants. These participants complete a Raven's Progressive Matrices test of ten items, followed by a short demographic and attitudinal survey. They receive a participation fee and a bonus based on the number of puzzles they solve correctly. This study serves a single purpose: to establish a benchmark median performance score for this participant population, which is then used as a fixed cutoff for classifying high and low performers in the main study.

The main study is the primary experimental intervention and consists of two parts run in a single continuous session. In the first part, participants complete the same ten-item Raven's test, which measures individual performance. In the second part, they take part in a redistribution task followed by a contest over whose preferred allocation is implemented. The same survey administered in the standalone study concludes the session.

After consenting, participants are informed that the study has two parts and that they will be paid a participation fee and a bonus that depends on their own and others' decisions. They then complete the Raven's test. Their performance generates a ranking that, under the merit-based treatment, determines their income: high or low status is assigned in real time by comparing each participant's score against the fixed benchmark median from the standalone study, with those falling exactly at the median equally likely to be assigned high or low income. Unlike in the standalone study, participants here receive no piece-rate bonus for the puzzles; performance instead determines their income in the second part under the merit-based rule, and their bonus derives from the outcomes of that part.

Following the contest, participants complete the survey, which covers demographic information, confidence, risk preferences, beliefs about meritocracy and redistribution (using standard World Values Survey items, such as "Hard work leads to success" versus "Success depends on luck and connections"), distributional preferences, and attitudes toward competition and political violence.

The session is subject to an overall time limit—one hour in the standalone study and two hours in the main study—well beyond the time each is expected to take

(approximately 10–15 and 30–45 minutes, respectively); participants are informed of this limit at the outset. To avoid imposing time pressure, the on-screen countdown remains hidden for most of the session and appears only during the final 15 minutes, as a warning that the deadline is approaching.

Part 2 comprises the main experimental intervention. Participants are randomly assigned to one of four between-subject treatment conditions in a 2 × 2 design. The two dimensions of treatment variation are: (1) whether income is allocated based on merit (performance on the Raven's task) or luck (random assignment), and (2) whether participants are explicitly informed about the allocation rule or informed that it is equally likely to be either, as summarized in table 1. Treatment is randomly assigned at the individual level. After receiving their income and treatment-specific information, participants are anonymously and randomly matched with six other individuals (in six independent rounds) who were assigned to the same treatment condition and receive a high or low income.

Allocation Rule \ Information	Known	Unknown
Merit		
Luck		

Table 1: 2 × 2 Treatment Matrix: Allocation Rule × Information Disclosure

At the beginning of each round, participants observe their own and their partner's incomes. Depending on the treatment, they may also be told how those incomes were determined. Participants are then asked to evaluate the fairness of the allocation and to propose how they would redistribute the combined income between themselves and their partner. They are also asked to predict how their partner would choose to redistribute the same income.

After expressing their redistribution preferences, participants are informed that their preferred allocation can be implemented only if they win a contest over the right to redistribute. In each of the six rounds, the participant receives 50 tokens and chooses how many tokens to spend to increase their probability of winning. The winner of the contest is determined by a lottery contest success function, where the probability of winning is proportional to the share of total tokens contributed by the participant. Unspent tokens are converted directly into points, which count toward final earnings.

Before making contest decisions, participants complete a comprehension quiz to ensure they understand the rules of the game. They must answer all questions correctly within two attempts to proceed. After completing all six rounds, one round is randomly

selected for payment. Participants receive a breakdown of that round, including the implemented allocation, both players' contest expenditures, the winner, and the resulting payoffs. Final earnings consist of the implemented allocation, any unspent tokens (converted to points), and the base payment.

Outcomes and Hypotheses: The primary outcome is the the level of conflict expenditure, measured by the number of tokens participants choose to spend to influence their probability of winning in each round. We hypothesize that conflict expenditure is highest when incomes are unequal and the source of inequality is unknown, as ambiguity invokes divergent fairness attributions. We test the following hypotheses based on our theoretical framework and experimental design:

Hypothesis 1. Belief updating and redistribution preferences under uncertainty
H1.1 High-income participants are more likely to believe that the allocation was merit-based than low-income participants when the income allocation rule is unknown.
H1.2 High-income participants propose a lower level of redistribution than low-income participants when the income allocation rule is unknown.

Hypothesis 2. Disagreement in redistribution preferences
Given unequal initial allocations, disagreement (between paired participants' preferred final allocations) is greater when income allocation rule is unknown versus known.

Hypothesis 3. Disagreement and conflict expenditure
H3.1 Higher actual disagreement predicts greater perceived disagreement.
H3.2 Higher perceived disagreement leads to higher contest expenditure (i.e., token spending to win the contest to redistribute).

Hypothesis 4. Inequality leads to higher conflict expenditure under the unknown information rule treatments (when it induces disagreement), than known information rule treatments (when it does not induce disagreement).

Taken together, these hypotheses suggest that on average, the conflict expenditure is higher when initial incomes are unequal than equal but the effect of income inequality on contest expenditure is increasing in disagreement over initial income allocation rule (i.e., it is higher under unknown income allocation rule treatment than known income allocation treatments).

4 Sample Size and Empirical Analyses

We calculate our required sample size to detect a minimum of 0.1sd effect size for our main hypothesis:

Hypothesis 4. Inequality leads to higher conflict expenditure under the unknown information rule treatments (when it induces disagreement), than known information rule treatments (when it does not induce disagreement).

To test this hypothesis, we estimate the required sample size to detect an interaction effect between information disclosure (Unknown information allocation rule) and income inequality on conflict expenditure. Our empirical specification is a panel regression of the form:

$$\text{Conflict}_{it} = \beta_0 + \beta_1 \text{Unknown}_i + \beta_2 \text{Inequality}_{it} + \beta_3 (\text{Unknown}_i \times \text{Inequality}_{it}) + \Delta X_i + \gamma_t + \epsilon_{it}$$

where i indexes participants and t indexes rounds. The variable `Unknown` is an indicator for assignment to the treatments where the information allocation rule is unknown, `Inequality` is a round-level indicator for whether initial incomes are unequal, and `Conflict` is the number of tokens spent in the contest. We include individual-level covariates X_i, round fixed effects γ_t, and cluster standard errors at the participant level.

Each participant completes six rounds. We assume that, on average, each participant experiences three equal-income and three unequal-income rounds. Income inequality varies within participants, while treatment assignment is between-subjects. The parameter of interest is β_3, the interaction term, which captures whether the effect of inequality on conflict differs in the unknown allocation rule compared to the known allocation rule conditions.

We aim to detect a minimum interaction effect of 0.1 standard deviations with 80% power and a 5% significance level. The interaction of interest combines a between-subjects factor (assignment to the unknown-rule condition) with a within-subjects factor (round-level income inequality, which varies across a participant's six matches). Because treatment assignment is between-subjects, the precision of the interaction estimate is governed primarily by the number of participants rather than the number of rounds. Under an assumption of intra-cluster correlation $\rho = 0.2$, and drawing on simulation-based power estimates for panel regressions with clustered standard errors, we estimate that we require approximately 425 participants per treatment, with a total of 1700 participants overall in the main study. Simulations matching this design (four equal-sized treatment cells, six rounds per participant with roughly three equal- and three unequal-income rounds) and standard errors clustered at the participant level

yield power of approximately 0.80 at this sample size.

Because income inequality varies within participants, individual-level heterogeneity partly differences out of the interaction estimate; consequently, a higher intra-cluster correlation tends to *increase* rather than decrease power for this particular interaction. The assumed $\rho = 0.2$ is therefore a conservative benchmark for the parameter of interest, and the design remains adequately powered across plausible values of the intra-cluster correlation, while accounting for repeated measures and clustering at the participant level.

In addition to the main study, we recruit a separate standalone sample of approximately 100 participants who complete Part 1 only. This sample is not used for hypothesis testing; its sole purpose is to establish the benchmark median performance score used to classify high and low performers in the main study. The total recruitment across both studies is therefore approximately 1800 participants.

4.1 Data Quality, Behavioral Measures, and Exclusions

To ensure that participants understand the contest before making decisions, they complete a comprehension quiz and must answer all questions correctly within two attempts to proceed; those who fail are not retained, and our primary analyses are conducted on all retained participants who complete the study. A concern specific to online problem-solving tasks is that the performance ranking used to assign merit-based income may be distorted by participants who do not solve the Raven's puzzles through independent, attentive effort, either because they rely on external tools such as AI assistants or because they disengage and solve inattentively. To address this in a pre-registered way, we passively record per-item response times (without imposing a per-item time limit) and log how participants interact with the study window, specifically instances of navigating away from the page (window blur) and revising answers. These measures do not affect payment or primary inclusion and are used only for robustness checks.

From them we construct three pre-specified flags, each targeting a distinct mechanism, whether tool-assisted or inattentive, by which the ranking could be inflated or rendered uninformative. The first captures a chatbot round-trip pattern: a participant who, on more than one item, both navigates away from the page and records a per-item response time above the 90th percentile, computed separately for each item position across all retained participants so that a dwell time is judged extreme relative to how long others took on the same puzzle rather than on the task as a whole. This mirrors the standard practice of trimming the upper tail to guard against outlying observations. The

second captures heavy off-task behavior, flagging any participant who navigates away five or more times across the ten items. The third captures automated or implausibly fast solving, flagging any participant whose median per-item response time falls below four seconds while their total score is eight or higher, a combination of speed and accuracy inconsistent with independent solving that isolates outliers in the opposite, lower tail. We will re-estimate our main specifications on the subsample excluding participants triggering any of these flags, in the spirit of a robustness check that removes outlying cases, and report results for both the full and flagged samples along with the share flagged by each criterion and the overlap among them.

References

Almås, Ingvild, Alexander W. Cappelen, and Bertil Tungodden (2010), "Fairness and the development of inequality acceptance." *Science*, 328, 1176–1178.

Cappelen, Alexander W., Thomas de Haan, and Bertil Tungodden (2024), "Fairness and limited information: Are people bayesian meritocrats?" *Journal of Public Economics*, 233, 105097.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden (2007), "The pluralism of fairness ideals: An experimental approach." *American Economic Review*, 97, 818–827.

Cassar, Lea and Arnd H. Klein (2019), "A matter of perspective: How failure shapes distributive preferences." *Management Science*, 65, 5050–5064.

Deffains, Bruno, Romain Espinosa, and Christian Thöni (2016), "Political self-serving bias and redistribution." *Journal of Public Economics*, 134, 67–74.

Fehr, Dietmar and Martin Vollmann (2025), "Misperceiving economic success: Experimental evidence on meritocratic beliefs and inequality acceptance." *Journal of Economic Inequality*, 23, 835–855.

Konow, James (2000), "Fair shares: Accountability and cognitive dissonance in allocation decisions." *American Economic Review*, 90, 1072–1091.

Valero, Vanessa (2022), "Redistribution and beliefs about the source of income inequality." *Experimental Economics*, 25, 876–901.