Heinrich H. Nax, Stefano Balietti, Ryan O. Murphy and Dirk Helbing

Meritocratic matching can dissolve the efficiency-equality tradeoff: the case of voluntary contributions

Working paper

Meritocratic Matching Can Dissolve the Efficiency-Equality Tradeoff: the Case of Voluntary Contributions Games

Heinrich H. Nax[1], Stefano Balietti[1], Ryan O. Murphy[2] and Dirk Helbing[1]

1 Computational Social Science, ETH Zürich, Switzerland
2 Decision Theory and Behavioral Game Theory, ETH Zürich, Switzerland
∗ **E-mail:** hnax@ethz.ch and sbalietti@ethz.ch

1 Abstract

One of the fundamental tradeoffs underlying society is that between efficiency and equality. The challenge for institutional design is to strike the right balance between these two goals. Game-theoretic models of public-goods provision under 'meritocratic matching' succinctly capture this tradeoff: under zero meritocracy (society is randomly formed), theory predicts maximal inefficiency but perfect equality; higher levels of meritocracy (society matches contributors with contributors) are predicted to improve efficiency but come at the cost of growing inequality. We conduct an experiment to test this tradeoff behaviorally and make the astonishing finding that, notwithstanding theoretical predictions, higher levels of meritocracy increase both efficiency and equality, that is, meritocratic matching dissolves the tradeoff. Fairness considerations can explain the departures from theoretical predictions including the behavioral phenomena that lead to dissolution of the efficiency-equality tradeoff.

Keywords: public-goods, meritocratic matching, efficiency, fairness, inequality.

JEL Codes: C92, D02, D63, H41.

Online Material: http://nodegame.org/games/merit/

2 Introduction

Making policy decisions often requires tradeoffs between different goals. One of the most fundamental tradeoffs is that between efficiency and equality. The basic idea of institutional *meritocracy* (Young, 1958) is to devise a system of rewards that "is intended to encourage effort and channel it into socially productive activity. To the extent that it succeeds, it generates efficient economy. But that pursuit of efficiency necessarily creates inequalities. And hence society faces a tradeoff between equality and efficiency." (Arthur M. Okun, *Equality and efficiency, the big tradeoff*, The Brookings Institution, 1975, p. 1.)

One could argue that inherent to this statement is the view that a certain type of societal activity can be modeled in the language of game theory as a public-goods provision/ voluntary contributions game (Isaac, McCue, and Plott, 1985; Ledyard, 1997; Chaudhuri, 2011). The resulting tradeoff summarizes as follows. In the baseline model, voluntary contributions games create no incentives for contributors and universal free-riding is the only stable equilibrium (Nash, 1950). In such a setting, the "tragedy of the commons" cannot be circumvented (Hardin, 1968). However, even if this outcome is maximally inefficient, one positive thing about it is that it comes with a very high degree of equality (at the cost of low average payoffs). For this reason, the outcome of universal free-riding has been controversially associated with extreme forms of socialism (Mises, 1922; Hayek, 1935). Fortunately, an array of mechanisms exists with the potential to foster contributions to public goods. One such mechanism that has been extensively studied in the literature is punishment (Fehr and Gächter, 2000; Ledyard, 1997; Chaudhuri, 2011). However, mechanisms such as punishment tend to be "leaky buckets" (Okun, 1975), in the sense that some of the

efficiency gains generated by the increase in contributions are spent in order to uphold them (e.g. on punishment costs).

An alternative mechanism, discussed here, is 'meritocratic matching' (Nax, Murphy, and Helbing, 2014) which is inspired by a recent, seminal paper introducing the "group-based meritocratic mechanism" (Gunnthorsdottir et al., 2010). Meritocratic matching generalizes the group-based meritocratic mechanism by introduction of an additional parameter that measures the degree of imprecision inherent to the mechanism's basic functioning and thus bridges the no-mechanism and group-based meritocratic mechanism continuously. Matching is said to be "meritocratic" because cooperators are matched with cooperators, and defectors are matched with defectors (Gunnthorsdottir et al., 2010; Nax, Murphy, and Helbing, 2014), hence "merit" is associated with contribution decisions.

Meritocratic matching differs from what is commonly associated with meritocratic mechanisms, which often feature explicit rewards/punishments, while meritocratic matching works only through action assortativity and not via payoff transfers. Nevertheless, in the real world, many mechanisms and institutions exist that are based on the logic of meritocratic matching. Admissions to schools or types of education, for example, are often based on rewards of past school or exam performances which are a function of the work/effort applicants had invested. An important determinant of what makes places that are more competitive to enter 'better' is the promise of being matched with others who also performed well in the best. Similarly, in professional team sports, clubs aim to hire athletes with good track records, and athletes join teams in order to matched with others. Basically, meritocratic matching mirrors the key features of many systems that feature team-based payments such as on trading desks.

Under meritocratic matching, near-efficient outcomes are supported by payoff-dominant equilibria (Nash, 1950; Harsanyi and Selten, 1988) provided the rate of return (Gunnthorsdottir et al., 2010) and the level of meritocracy exceed certain thresholds (Gunnthorsdottir et al., 2010). The reason for this is that agents have incentives to contribute more in order to be grouped with other high-contributors. As a result, only a small fraction of free-riders continues to exist in these equilibria. Such equilibria are excellent predictors of the population's distribution of play under 'full meritocracy' (Gunnthorsdottir et al., 2010; Gunnthorsdottir and Thorsteinsson, 2011; Gunnthorsdottir, Vragov, and Shen, 2010; J.P. And Rabanal, 2015). Unfortunately, the new equilibria, however desirable in terms of efficiency vis-à-vis tragedy of the commons, typically feature a higher degree of inequality.[1] The contrast between these two outcomes is well illustrated by the tensions that would exist between an ideal Benthiam (utility-maximizing) social planner, on the one hand, and an ideal Rawlsian (inequality-minimizing) social planner on the other: in many games, the Benthiam (Bentham, 1907) would strictly favor perfect action-assortativity, while the Rawlsian (Rawls, 1971) would rather prefer complete non-assortativity. In comparison, a real-world social planner typically exercises a certain degree of 'inequality aversion', aiming for an outcome between these two extremes (Atkinson, 1970).

Essentially, the efficiency-equality tradeoff in designing a meritocratic matching regime boils down to the choice of a systemic degree of assortativity, i.e. the selection of a certain degree of meritocracy. This tradeoff is at the heart of social choice theory (see e.g. (Arrow, 1951; Sen, 1970; Gauthier, 1986; Arrow, Bowles, and Durlauf, 2000)) and welfare economics (see e.g. (Samuelson, 1980; Feldman, 1980; Atkinson, 2012)). Zero meritocracy represents maximal equality, but also minimal efficiency; full meritocracy represents the opposite. For any degree of inequality aversion away from the two extremes (given by (Bentham, 1907) and (Rawls, 1971)), there exist, at least in theory, an intermediate degree of meritocracy that maximizes social welfare (Nax, Murphy, and Helbing, 2014). Unfortunately, this is a difficult tradeoff as the buckets are leaky in both directions: reducing meritocracy increases equality at the expense of efficiency, and increasing meritocracy increases efficiency at the expense of equality.

In this paper we set out to test this tradeoff experimentally by analysis of intermediate regimes of

[1] The new equilibria always have positive variance, while the free-riding equilibrium has variance zero. In what cases this translates into more inequality depends both on the particular structure of the equilibrium given a game and on the measure of inequality that is applied.

meritocracy. We are thus the first to bridge the rich experimental literature on public-goods games under random interactions (zero meritocracy) (Ledyard, 1997; Chaudhuri, 2011) with the more recent literature on full meritocracy (group-based mechanisms) (Gunnthorsdottir et al., 2010; J.P. And Rabanal, 2015). The experiments reveal that the strict tradeoff implied by theory is dissolved in practice. Higher degrees of meritocracy turn out to increase welfare for any symmetric and additive objective function (Atkinson, 1970), including Benthiam utility-maximization (Bentham, 1907) and Rawslian inequality minimization (Rawls, 1971). In other words, meritocracy increases both efficiency *and* equality, leading to unambiguous welfare improvements as we illustrate for a variety of measures. We argue that the dissolution of the tradeoff is driven by the agents' distastes of 'meritocratic' unfairness, and by the corrections to their actions that these considerations imply. The view of fairness that we adopt and test here generalizes the concept of distributive fairness/ inequity aversion (Fehr and Schmidt, 1999; Ockenfels and Bolton, 2000) to settings with positive levels of meritocracy. This fairness definition is a game-theoretic application of a notion related to systemic fairness (Adams, 1965; Greenberg, 1987), which has been long recognized in organizational theory, but not previously applied to game theory (and the problem of public-goods provision in particular). The patterns associated with reactions to between-group comparisons, however, have been noted as robust phenomena without being interpreted as driven by norms of fairness (Bornstein, Erev, and Rosen, 1990; Erev, Bornstein, and Galili, 1993; Bornstein and Erev, 1994; Bornstein, Gneezy, and Nagel, 2002; Bohm and Rockenbach, 2013).

Among our results are the following key findings:

1. *Efficiency increases with meritocracy.* Perfect meritocracy is near-efficient and coincides with the theoretically predicted levels. The zero meritocracy regime lies above the efficiency levels implied by the theoretical equilibrium assuming self-regarding rational choice. For intermediate meritocracy levels, efficiency is above that of zero meritocracy, but below the theoretically expected equilibrium values.

2. *Equality, in contrast to theoretical predictions, also increases with meritocracy.* This finding is robust with regard to several inequality measures, including the payoff of the worst-off subject. In our settings, the often-cited tradeoff between equality and efficiency turns out to be a theoretical construct, rather than a behavioral regularity.

3. *Fairness considerations can explain the dissolution of the tradeoff between efficiency and equality.* According to our definition, agent A considers the outcome of the game "unfair" if another agent B contributed less than A, but B was placed in a better group. As a consequence, agent A is assumed to respond by decreasing his/her contribution.

4. *Higher meritocracy levels increase agents' sensitivity to unfair group matching in lower meritocracy levels.* Our experimental setup expose each participant to two distinct levels of meritocracy. When the second part of the experiment is restarted at a lower meritocratic regime, it turns out that agents' distaste for unfair group matching is magnified.

3 The experiment

3.1 The underlying meritocracy game

A fixed population of n agents plays the following public-goods game repeatedly through periods $T = \{1, 2, ..., t\}$. First, each agent i simultaneously decides to contribute any number of coins c_i between zero and his full budget $B > 0$. The amount not contributed goes straight to his/her private account. The ensemble of players' decisions yields the contribution vector c. Second, Gaussian noise with mean zero and variance $\sigma^2 \geq 0$. Third, k groups of a fixed size $s < n$ (such that $s * k = n$) are formed according to the ranking of the values c' (with random tie-breaking). That is, the highest s contributors form group

G_1, the next highest s contributors form G_2, etc. The resulting group partition is $\rho = \{G_1, G_2, ..., G_k\}$. Finally, based on the grouping and the initial contributions vector c, payoffs ϕ are computed. Each player i in a group G_i with other players $j \neq i$ receives:

$$\underbrace{\phi_i(c)}_{\text{payoff}} = \underbrace{(B - (1-m) * c_i)}_{\text{return from private account}} + \underbrace{\sum_{j \in G_{-i}} m * c_j}_{\text{return from group account}}, \tag{1}$$

where m represents the marginal per capita rate of return, and G_{-i} indicates the members of group G_i excluding i.

Note that the game is equivalent to play under the group-based mechanism (here, 'perfect meritocracy') (Gunnthorsdottir et al., 2010) if $\sigma^2 = 0$, and that the case of $\sigma^2 \to \infty$ corresponds to random re-matching (here, 'zero meritocracy') (Andreoni, 1988).

Equilibrium play

To highlight the structure of the Nash equilibria (Nash, 1950) for this class of games, it is useful to evaluate the value of the expected payoff $\mathbf{E}[\phi_i(c)]$ during the decision stage, i.e. before groups are formed. In Eq. (1), the first term, i.e. the private-account return, is completely determined by the agent's contribution choice. The second term, i.e. the group-account return, however, depends on the players' contributions in a probabilistic way. In the case of zero meritocracy (i.e. random re-matching) ($\sigma^2 = \infty$), $\mathbf{E}[\phi_i(c)]$ is strictly decreasing in the player's own contribution because the marginal per capita rate of return is less than one. Under zero meritocracy, the player's own contribution has no effect on group matching, and, therefore, the only equilibrium is universal free-riding. Conversely, for positive levels of meritocracy, the player's contribution choice influences the probability of being ranked in a high group. Hence, making a positive contribution is a tradeoff between the sure loss on the own contribution and the promise of a higher return from the group-account. However, the chances of being ranked in a better group are decreasing with growing variance. As a result, new Nash equilibria with positive contribution levels may emerge: indeed, Nax, Murphy, and Helbing (2014) generalizes the results by Gunnthorsdottir et al. (2010) showing that, if the level of meritocracy stays sufficiently large in addition to some bound on r, there exist a near-efficient pure-strategy Nash equilibria in which a large majority of players contributes the full budget B and a small minority of players contributes nothing.[2]

3.2 Choice of experimental parameters

In order to ensure comparability with the literature on voluntary contributions games under random re-matching (Andreoni, 1988) (as reviewed by Ledyard 1997; Chaudhuri 2011) and particularly under the group-based mechanisms (Gunnthorsdottir et al., 2010), we set the group size $s = 4$ and the marginal per capita rate of return $m = 0.5$ (as in Gunnthorsdottir et al. 2010). Due to laboratory capacity restrictions and as also chosen in many prior experiments, we set $n = 16$. Finally, we need to set different meritocracy levels as represented by variance σ^2 other than $\sigma^2 = 0$ and $\sigma^2 = \infty$.

In order to determine the right and meaningful levels of variance levels, we conducted a series of 16 experimental sessions on Amazon's Mechanical Turk (AMT) with a total of 242 participants using our new NodeGame software (Balietti, 2014). Details about the experiment can be found in Appendix A.2. In each session, all participants played a game with different variance levels which were $\sigma^2 = \{0, 2, 4, 5, 10, 20, 50, 100, 1000, \infty\}$. For all variance levels below $\sigma^2 = 100$, the near-efficient Nash equilibria

[2]Universal free-riding continues to be an equilibrium too. See Theorem 1 in Ref. (Gunnthorsdottir et al., 2010) and Propositions 6 and 7 in Ref. (Nax, Murphy, and Helbing, 2014) for detailed proof and game-theoretic characterization of these equilibria.

exist in the stage game. For higher variance levels, the free-riding Nash equilibrium is the unique Nash equilibrium of the stage game.

Each game was repeated for 25 (or 20) successive rounds. Given $\sigma^2 = 0$, play basically coincided with the levels implied by the near-efficient Nash equilibria almost from first to last round. We evaluated the level of variance starting at which the mechanism started (i) to display contribution levels that differed from the levels implied by the near-efficient Nash equilibria under $\sigma^2 = 0$ initially but reached those over time, and (ii) to exhibit contribution levels that did not stabilize at such levels at all. We found these variance levels to be (i) $\sigma^2 = 3$ and (ii) $\sigma^2 = 20$. Appendix A.3 (Fig. 7) contains further details. Hence, we settled for the following four variances for our laboratory experiment: $\sigma^2 = \{0, 3, 20, \infty\}$. We use the following terminology. We labeled $\sigma^2 = 0$ as PERFECT-MERIT, and $\sigma^2 = \infty$ as NO-MERIT. Intermediate values are labeled as HIGH-MERIT ($\sigma^2 = 3$) and LOW-MERIT($\sigma^2 = 20$).

Note that in the case of these four levels of variance tested in this study, the predicted stage-game Nash equilibria are as follows. For $\sigma^2 = \infty$ (NO-MERIT), the unique stage-game Nash equilibrium is universal free-riding, which is also a Nash equilibrium for all the other variance levels. For $\sigma^2 = \{0, 3, 20\}$, moreover, there exist $\binom{n}{2}$ alternative pure-strategy equilibria where exactly two players free-ride while all others contribute fully. Details on equilibria can be found in Appendix A.1.

3.3 The laboratory experiment

We ran 12 experimental sessions with a total of 192 participants at the ETH Zürich Decision Science Laboratory (DeSciL) using the same NodeGame software as in the pre-tests (Balietti, 2014). Details about the experiment can be found in Appendix A.2. In each session, all participants played two repeated games, one after the other, each one with one of the different variance level $\sigma^2 = \{0, 3, 20, \infty\}$. Each session, therefore, represented a unique order of two of the four possible variance levels (leading to 12 sessions to account for every possible ordered pair). Each repeated game was played for 40 successive rounds ($T = \{1, 2, ..., 40\}$), with population size $n = 16$, group size $s = 4$, and marginal per capita rate of return $m = 0.5$.[3]

4 Results

Overall, we found a significant difference in the mean level of contributions among the four treatments (linear mixed model LMM: $F_{3,8} = 36.8, P < 0.0001$), as Fig. 1 illustrates. Furthermore, Fig. 2 shows how the contribution patterns observed in the laboratory are part of a coherent picture with the results of the AMT pre-tests for different level of variance.

In the following, we first study efficiency, inequality and fairness, focusing on the first part of the experiment. Then, we use the second part of the experiment to determine the agents' sensitivity to changes in meritocracy levels.

4.1 Efficiency

In this section, we evaluate the effect of meritocracy on total payoffs generated, i.e. on efficiency. Theory predicts (Gunnthorsdottir et al., 2010; Nax, Murphy, and Helbing, 2014) that equilibria supported by higher meritocracy levels are more efficient, and we shall show that this predictions holds true in the lab, confirming previous experimental results (Gunnthorsdottir et al., 2010; Gunnthorsdottir and Thorsteinsson, 2011; Gunnthorsdottir, Vragov, and Shen, 2010). Indeed, the levels of efficiency supported by the payoff-dominant equilibria under meritocracy regimes LOW-MERIT, HIGH-MERIT and

[3]We would have liked to reproduce the 80 rounds of play by Gunnthorsdottir et al. (2010), but due to time restrictions as in how long we could keep subjects in the laboratory, we decided to halve this amount in order to be able to run two variance levels per person. Each session lasted roughly one hour.

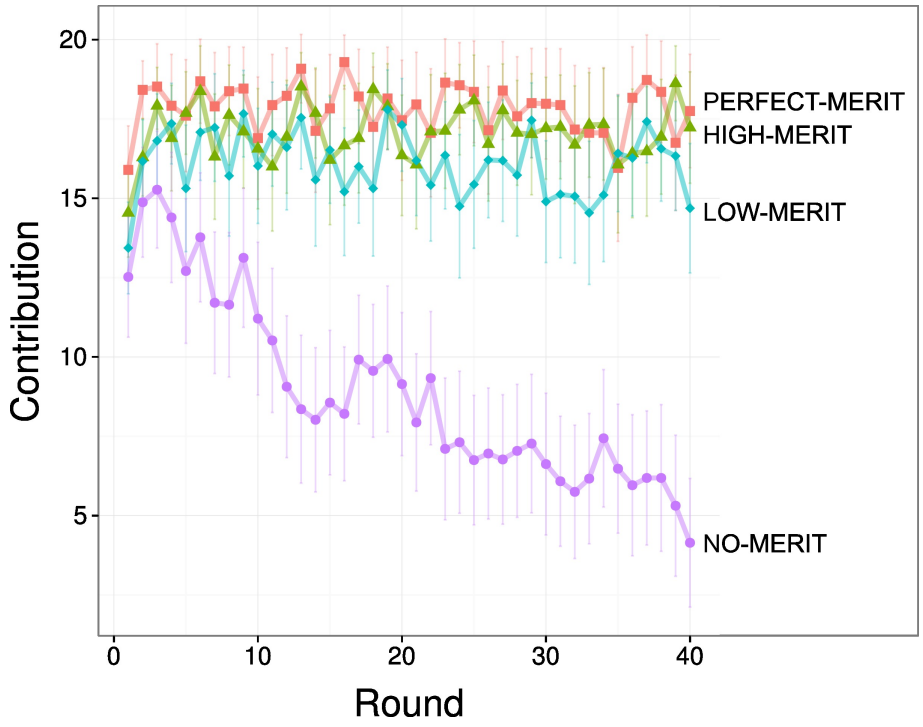

Figure 1. Average contribution levels for perfect-, high-, low-, and no-meritocracy, respectively associated with the values of $\sigma^2 = \{0, 3, 20, \infty\}$. Contribution levels increase as meritocracy increases. In perfect meritocracy, contribution levels are near efficient and approximately coincide with theoretical predictions. Meritocratic treatments are mostly stable over the forty rounds of the game, and do not follow the contribution decay of the random treatment. Error bars represent the 95%-confidence intervals.

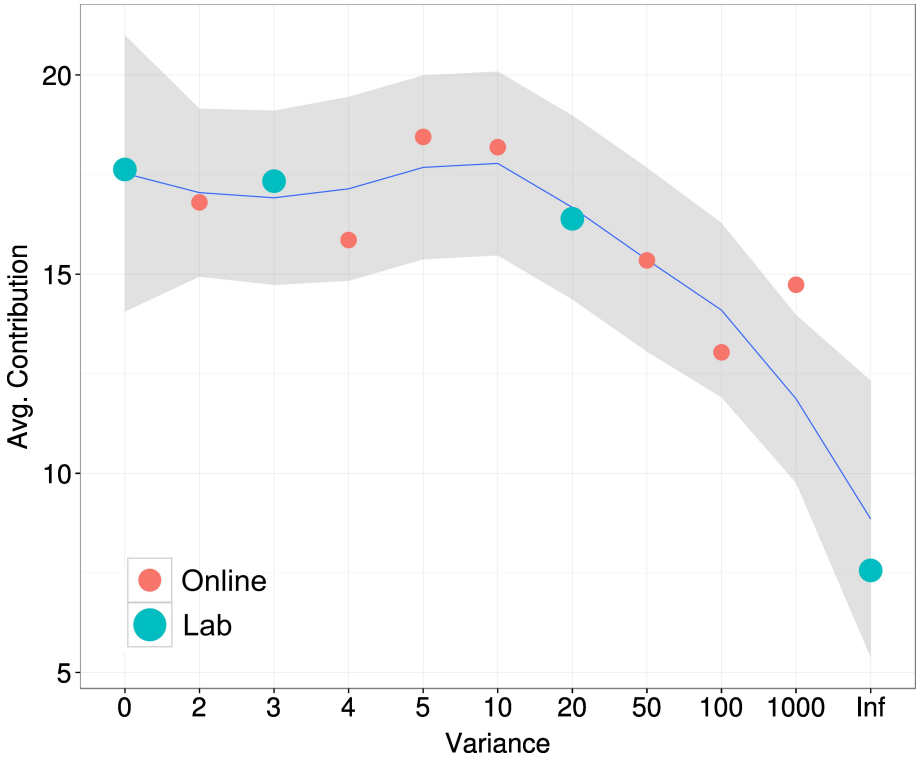

Figure 2. Average contribution levels for different variance levels of online and lab experiments. Contribution levels decrease as variance increases, that is contribution levels increase as meritocracy increases.

PERFECT-MERIT represent relatively accurate predictions, while the complete inefficiency prediction of the unique, zero-contribution Nash equilibrium under no-meritocracy (NO-MERIT) understates the achieved efficiency levels in the order of standard magnitudes (Ledyard, 1997; Chaudhuri, 2011).

We measure efficiency as the average payoff over players, $\overline{\phi} = \frac{\sum_{i \in N} \phi_i}{n}$, over the forty rounds. As shown in Fig. 3, when climbing up the meritocracy ladder we find increases in efficiency from $\sigma^2 = \infty$ (NO-MERIT) through $\sigma^2 = \{20, 3\}$ to $\sigma^2 = 0$ (PERFECT-MERIT).

Overall, we observe significant differences in the mean of realized payoffs among the four treatments (linear mixed model LMM: $F_{3,8} = 36.95, P < 0.0001$). Taking NO-MERIT as a baseline, LOW-MERIT led to an increase in the average realized payoff of 7.1611 (Likelihood Ratio Test LRT: $\chi_{(1)} = 12.7, P = 0.0004$), HIGH-MERIT to an increase of 8.1964 (LRT: $\chi_{(1)} = 17.48, P < 0.0001$), and PERFECT-MERIT to an increase of 8.8287 (LRT: $\chi_{(1)} = 16.22, P < 0.0001$). These levels correspond to roughly double those of NO-MERIT. Computing the most conservative (Bonferroni) adjusted p-values on all pairwise differences reveals that the treatment with variance ∞ is significantly different ($P < 0.0001$) from the other three variance levels $\sigma^2 = \{0, 3, 20\}$, which are themselves not significantly different from each other.

For intermediate meritocracy regimes $\sigma^2 = \{20, 3\}$, efficiency is significantly below the level implied by the respective payoff-dominant equilibria (Harsanyi and Selten, 1988), but only by less than five percent. Conversely, under full meritocracy $\sigma^2 = 0$, efficiency is above and within five percent of equilibrium. Note that contribution levels resemble the levels implied by the symmetric mixed-strategy Nash equilibrium identified in Ref. (Nax, Murphy, and Helbing, 2014), but do not perfectly coincide with them, as intermediate contribution levels continue to be selected under $\sigma^2 = \{20, 3\}$, which are dominated even in the mixed equilibrium.

The contribution patterns under $\sigma^2 = 0$ confirm the qualitative patterns of contributions found in (Gunnthorsdottir et al., 2010), instead now we have $n = 16$. For $\sigma^2 = \infty$, we have the same pattern of contributions that, on average, roughly halve every 10-20 rounds as found in many related studies (Ledyard, 1997; Chaudhuri, 2011).

4.2 Equality

Recall from the theory predictions in Ref. (Nax, Murphy, and Helbing, 2014) that equilibria supported by higher meritocracy levels feature more inequality in the distribution of payoffs. In this section, we shall show that laboratory evidence yields diametrically opposite results; namely, higher meritocracy levels lead to outcomes that are more equal in terms of payoff distributions.

One can identify two measures of payoff inequality directly from the moments of the payoff distribution: (i) the payoff of the worst-off (Rawls, 1971), $\underline{\phi} = \min\{\phi_i\}$, and (ii) the variance of payoffs, $\sigma^2 = \frac{\sum_{i \in N} (\phi_i - \overline{\phi})^2}{n}$. A more sophisticated third alternative is (iii) the Gini coefficient. In terms of all measures, our analysis shows that equality increases with meritocracy. Note that the following results are also robust to other measures of inequality (Cowell, 2011) (see *appendix*).

Fig. 4 shows that, like efficiency, equality also increases from $\sigma^2 = \infty$ (NO-MERIT) through $\sigma^2 = \{20, 3\}$ to $\sigma^2 = 0$ (PERFECT-MERIT). These increases are reflected by differences in the Gini coefficient, and by the order of the payoff of the worst-off – Rawlsian inequality. Under NO-MERIT, equality is significantly below the level implied by equilibrium. For all three positive levels of meritocracy, equality is above that achieved by NO-MERIT and above the theoretically implied levels. Details about the statistical tests can be found in Appendix A.3.

4.3 Fairness

We have found that Nash predictions worked well in approximating efficiency levels in the meritocratic regimes LOW-MERIT, HIGH-MERIT and PERFECT-MERIT, but not in NO-MERIT. In this section,

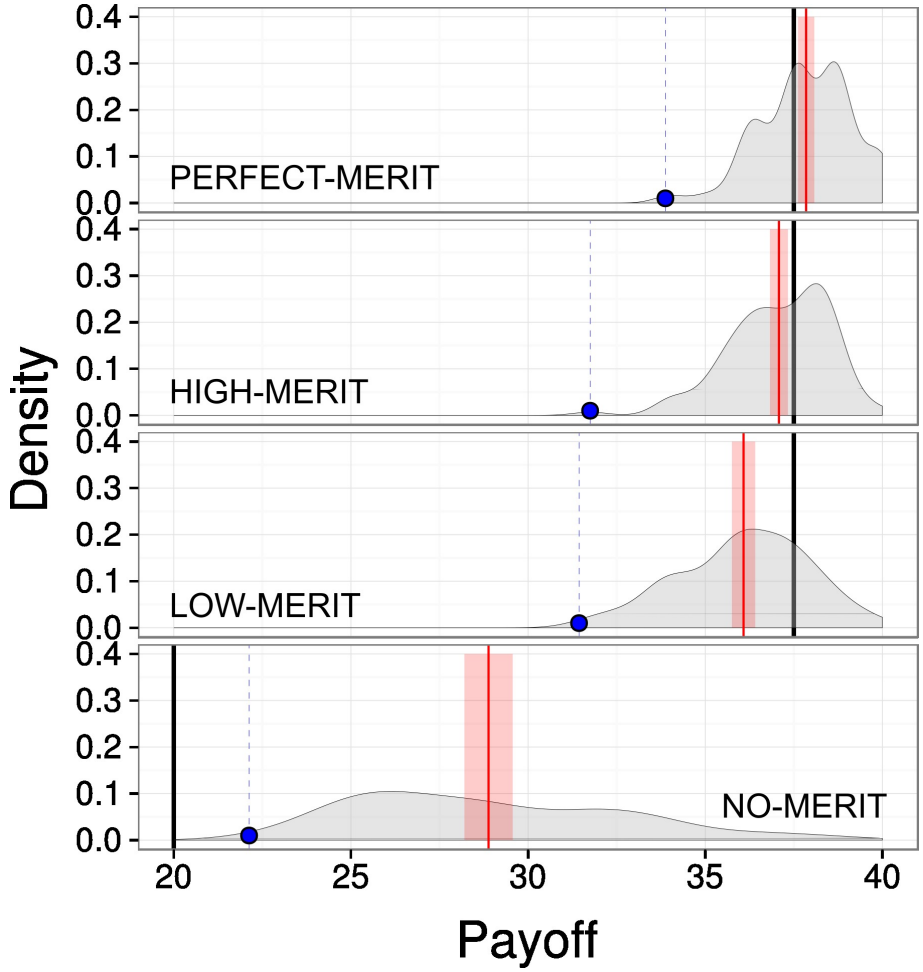

Figure 3. Analysis of efficiency based on smoothed distributions of average payoffs over 40 rounds for perfect-, high-, low-, and no-meritocracy, respectively associated with the values of $\sigma^2 = \{0, 3, 20, \infty\}$. Efficiency, measured as average payoff, increases as meritocracy increases. Black solid lines indicate the mean payoff as implied by the respective payoff-dominant Nash equilibria, red solid lines indicate the mean payoff observed in the experiment, red-shaded areas indicate the 95%-confidence intervals of the mean. Blue dots indicate the payoff of the worst-off player (note that the worst-off player in every equilibrium receives twenty 'coins').

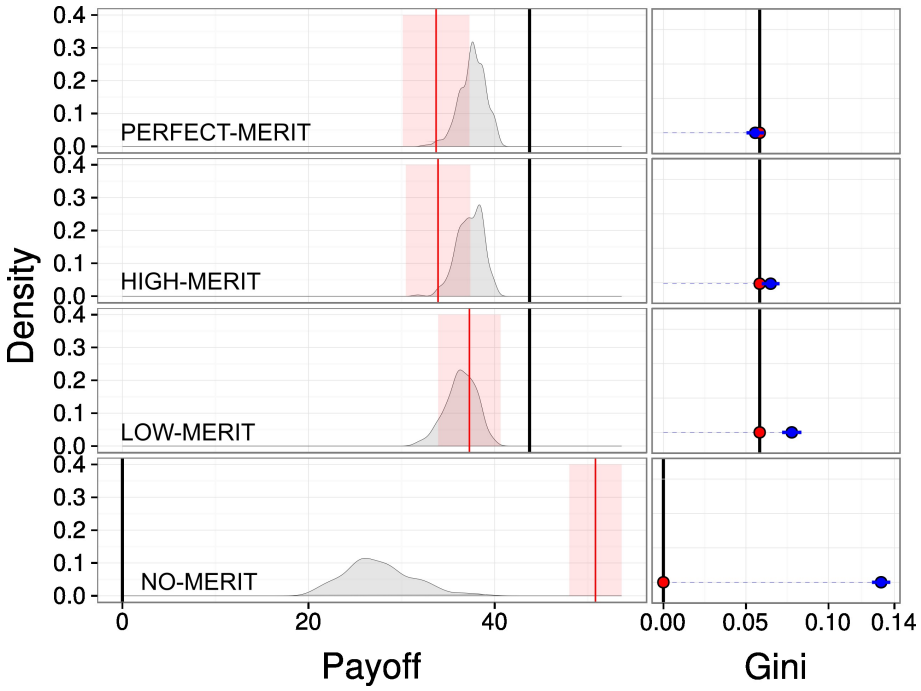

Figure 4. Level of payoff equality for perfect-, high-, low- and no-meritocracy, respectively associated with the values of $\sigma^2 = \{0, 3, 20, \infty\}$. Inequality, measured by the variance of payoff and by the Gini coefficient, decreases, as meritocracy increases. Left panel: Smoothed distributions of average payoffs over 40 rounds. Black solid lines indicate the variance of the payoffs as given by the respective payoff-dominant Nash equilibria, red solid lines indicate the mean variance observed in the experiment, red-shaded areas indicate the 95%-confidence intervals of the mean variance. Right panel: Average Gini coefficient of the distribution of payoffs with 95%-confidence intervals. Black solid lines and and red dots indicate the Gini coefficient implied by the equilibrium (without fairness considerations).

we explore the role of individuals' fairness considerations in explaining these deviations. We shall find evidence for meritocratic fairness concerns that could explain these phenomena and that generalize well-known *fairness* considerations (Fehr and Schmidt, 1999; Ockenfels and Bolton, 2000) in the meritocracy context, allowing for a systemic understanding of the payoff structure.

Meritocratic fairness: definition

In public-goods games with completely random interactions, i.e. in environments with zero meritocracy, a payoff allocation is considered *unfair* if players contribute different amounts and therefore obtain different payoffs (Fehr and Schmidt, 1999). From the perspective of an individual player, unfairness can be *advantageous*, if he/she contributed less than the average, or *disadvantageous* in the opposite situation.

In public-goods games with positive levels of meritocracy, we define an outcome as *fair* if all players are matched into group with contributions that are compatible, that is, there are no players contributing less (more) than others that get matched into a better (worse) group. Conversely, a payoff allocation is considered *unfair* from the viewpoint of a player if there exists at least one other player who contributed less (more) than him/her who is matched into a group with a lower (higher) average contribution level. The more players are matched into such incompatible groups, and the larger the difference in average group payoffs, the higher the level of meritocratic unfairness perceived by that player. More formally, *meritocratic unfairness* of a given payoff allocation is measured by the following two quantities:

$$\begin{array}{ll} MU_{Dis} & = \frac{1}{n-s} * \sum_{j \in N} \max(\Delta_{ij}, 0) * \max(\Delta_{G_j G_i}, 0), \\ MU_{Adv} & = \frac{1}{n-s} * \sum_{j \in N} \max(\Delta_{ji}, 0) * \max(\Delta_{G_i G_j}, 0), \end{array} \tag{2}$$

where for any pair of players, i and j in groups G_i and G_j $(i \neq j)$, Δ_{ij} represents the difference in contributions $c_i - c_j$, and $\Delta_{G_i G_j}$ is the difference in average group contributions $\frac{1}{4}\sum_{k \in G_i} c_k - \frac{1}{4}\sum_{k \in G_j} c_k$.

Contribution decisions: meritocratic fairness and strategic concerns

It has been shown that under random interactions unfair allocations influence players' utilities negatively and that agents respond to unfairness by adjusting their contributions, especially to disadvantageous unfairness (Fehr and Schmidt, 1999; Ockenfels and Bolton, 2000). Disadvantageous unfairness has an accentuated negative effect on a player's utility, while advantageous unfairness has a negative but weaker effect. This gain-loss asymmetry is of course related to some of the most robust findings in experimental economics (Kahneman and Tversky, 1979; Tversky and Kahneman, 1991; Erev, Ert, and Yechiam, 2008). The consequences of the distaste for unfairness are such that, on average, a player responds by decreasing (increasing) his/her contribution after experiencing disadvantageous (advantageous) unfairness (Fehr and Schmidt, 1999; Ockenfels and Bolton, 2000). Importantly, the tendency to decrease is stronger than the tendency to increase due to the asymmetry in distastes. The typical contribution pattern found in repeated public goods experiments (intermediate contribution levels at the beginning, followed by a decay over time) can therefore be explained by heterogeneities in social preferences and the asymmetric reactions to advantageously and disadvantageously fair outcomes related to reciprocity (Ledyard, 1997; Chaudhuri, 2011).

It is reasonable to conjecture that fairness considerations continue to matter in the presence of meritocracy. In line with previous behavioral findings in studies investigating distributional fairness (Fehr and Schmidt, 1999; Ockenfels and Bolton, 2000), we assume that disadvantageous unfairness has a more accentuated negative effect than advantageous unfairness. The consequences of the distaste for meritocratic unfairness in repeated random interactions are assumed to be such that, on average, a player responds by decreasing (increasing) his/her contribution after experiencing disadvantageous (advantageous) meritocratic unfairness. Note that, under this definition, every outcome is meritocratic and fair with probability one under perfect meritocracy (when $\sigma^2 = 0$).

However, an additional subtlety comes from the fact that contributions under meritocratic matching play a double role. On the one hand, they determine a player's payoff within a given group. On the other hand, they also determine the group into which the player is matched. Therefore, players' contribution decisions are a result of fairness considerations and strategic concerns:

Contribution Decision = Meritocratic Fairness + Strategic Concerns.

Our assumptions regarding meritocratic fairness and strategic concerns lead to the following predictions:

- In environments with zero meritocracy, our predictions coincide with those of Ref. (Fehr and Schmidt, 1999; Ockenfels and Bolton, 2000), that is, we expect the typical contribution pattern (intermediate contributions levels at the beginning, then decay over time). The decay is driven by the asymmetry in behavioral responses to disadvantageous versus advantageous unfairness.
- Under perfect meritocracy, starting at the near-efficient Nash equilibrium prediction, we do not expect significant departures from such a best-response state as there is no inherent meritocratic unfairness (by definition).
- For the intermediate meritocracy levels (HIGH-MERIT and LOW-MERIT), starting at the near-efficient Nash equilibrium prediction, we expect decreases as unfairness is expected to occur even in equilibrium. However, other than under zero meritocracy, downward corrections of contributions will not trigger an overall downward decay of contributions because higher amounts become better and fair replies again than contributing zero once substantial decreases of contributions occurred, which were themselves triggered by disadvantageous unfairness. This is due to the fact that there are then new strategic concerns.

Meritocratic fairness: results

Fig. 5 shows the distributions of meritocratic unfairness across different treatments. Similarly to efficiency and inequality, we find increases in fairness from NO-MERIT through all meritocracy levels up to PERFECT-MERIT, and these increases are significant (LMM: $F_{3,8} = 53.74, P < 0.0001$).

Meritocratic unfairness translates directly into departures from the levels of contribution predicted by theory. In particular, we studied how the unfairness level experienced in the previous round impacts the decision to contribute in the following round. To do so, we performed a multilevel regression of between-rounds contribution adjustments with subject and session as random effects, and we tested several models for both distributional (Fehr and Schmidt, 1999) and meritocratic fairness (statistical details are given in the Statistical Analysis section in *Materials and Methods* section and regression tables are available in the Supplementary Information). As expected, applying the notion of distributional fairness *as it is* to a meritocratic environment is not straightforward: the results of the regressions for distributional fairness are often inconsistent across treatments, and, even in many cases contrary to the predictions of the theory. On the other hand, meritocratic unfairness proved a good predictor of the contribution adjustments between rounds across all treatments. Therefore, meritocratic fairness can be seen as natural generalization of distributional fairness in games with positive levels of meritocracy.

4.4 Sensitivity

So far, we have shown that (i) both efficiency *and* equality increase with meritocracy, and that (ii) considerations of 'meritocratic' fairness can explain deviations from the theoretically expected equilibrium. In this section, we show that changes in the level of experienced meritocracy have significant implications

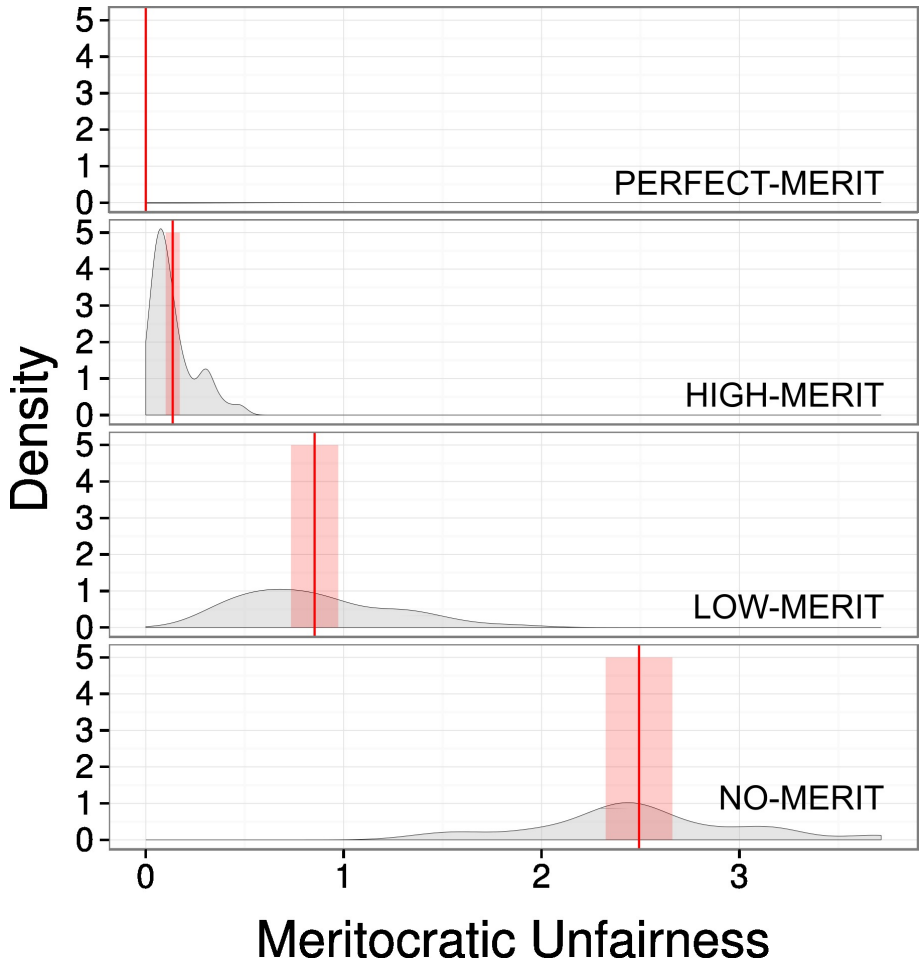

Figure 5. Meritocratic unfairness for perfect-, high-, low-, and no-meritocracy, respectively associated with the values of $\sigma^2 = \{0, 3, 20, \infty\}$. Smoothed distribution of average meritocratic unfairness per round. Unfairness decreases as meritocracy increases. Red solid lines indicate the mean level of meritocratic unfairness observed in the experiment, red-shaded areas indicate the 95%-confidence intervals of the mean.

as well. In particular, we test whether participants coming from a higher (lower) meritocracy level in part 1 are more (less) sensitive to meritocratic unfairness in part 2.

For this analysis, we used the data pertaining of part 2 of the experiment, controlling for which meritocracy level was played in part 1. We divided the dataset in two subsets, depending on whether participants in part 2 experienced a higher or lower meritocracy level than in part 1. In order to obtain a balanced design with respect to the direction of meritocracy changes, we further sampled the data from part 2 to include only the intermediate regimes of meritocracy ($\sigma^2 = \{3, 20\}$). In this way, both conditions could be tested against perfect meritocracy, zero meritocracy, and one intermediate regime. We created a dummy variable for "contribution goes down" (0;1) and performed a multilevel logistic regression with subject and session as random effects. We used the level of disadvantageous meritocratic unfairness experienced in the previous round as a predictor of whether contribution is expected to go up or down in the next round.

Our main finding is that the distaste for meritocratic unfairness is exacerbated after having played a more meritocratic regimes in part 1. That is, if a participant experienced meritocratic unfairness in the previous round, he/she is more likely to reduce the own contribution in the current round if the level of meritocracy in part 2 is lower than in part 1 (Logistic Mixed Regression LMR: $Z = 2.521, P = 0.0117$). The effect in the opposite direction – a lower meritocracy level in part 1 than in part 2 – is not significant (LMR: $Z = 1.522, P = 0.128$).

The different sensitivity to meritocratic unfairness leads to different levels of efficiency and equality overall. Sessions in part 2 with higher sensitivity to meritocratic unfairness – i.e. descending the meritocracy ladder – have significantly lower average payoff (One-sided Kolmogorov-Smirnoff KS: $D^+ = 0.1531, P < 0.0001$), and significantly higher inequality – measured by the average Gini coefficient per round ($D^+ = 0.1583, P = 0.0494$). These results confirm once again that, in our settings, increases in efficiency are followed by inequality reduction, and that meritocratic fairness considerations can explain the dissolution of the classical efficiency-equality tradeoff.

5 Discussion

Economic theory has identified the efficiency-equality tradeoff as one of the most fundamental tradeoffs underlying society (Arrow, 1951; Sen, 1970; Okun, 1975; Gauthier, 1986; Arrow, Bowles, and Durlauf, 2000). In our study, we decided to analyze an environment that succinctly captures the essence of this tradeoff. The well-known public-goods (voluntary-contribution) game (Isaac, McCue, and Plott, 1985) perfectly suited our task, since it naturally relates to many important real-life issues such as climate change, collective action, common-pool resource problems, etc. (Ostrom, 1990; Ostrom, 1999). For this, it has received tremendous attention in the theoretical and experimental literature in and outside of economics (Chaudhuri, 2011).

The standard case of random re-matching and a recently proposed and seminal group-based mechanism (Gunnthorsdottir et al., 2010) were generalized to a class of mechanisms called "meritocratic matching" (Nax, Murphy, and Helbing, 2014). Here, we test these mechanism, we made the astonishing finding that agents seem to be able to 'make the better system work'. That is, meritocratic mechanisms that promise higher efficiency from a theoretic point of view, also turn out to benefit the worst-off and to improve overall distributional equality, despite theory predicting the opposite (Nash, 1951). The reason for this unexpected finding lies in agents' attempts to improve 'fairness' by adjustments of their actions in order to counter situations in which particular agents are better-off (worse-off) despite being associated with low (high) 'merit'. This fairness concept not only explains our results in the new class of assortative games studied by us, but also remains a significant explanatory variable in games with random interactions, and is consistent with previous results for this class of games. The criterion of 'meritocratic' fairness is formally different from the standard formulation of 'distributional' fairness (Fehr and Schmidt, 1999; Ockenfels and Bolton, 2000), but for random interaction environments their predictions agree qual-

itatively. In meritocratic environments, due to the double-role of contributions inherent in the matching mechanism (both as a group-sorting device and as a payoff determinant within groups), the concept of 'meritocratic' fairness is indeed a natural extension of classical fairness criteria when agents are aware of this double-nature.

The results of our study show that meritocracy can dissolve the fundamental tradeoff between efficiency and equality. Creating a public good does not necessarily generate inefficiencies, nor it requires the intervention of a central coercive power for their suppression. Fairness preferences and suitable institutional settings, such as well-working merit-based matching mechanisms, can align agents' incentives, and shift the system towards more cooperative and near-efficient Nash equilibria. Overall, the results of our experiment lend credibility to agents' sensitivity to the famous quote associated with Virgil that "The noblest motive is the public good."

6 Acknowledgements

The authors acknowledge support by the European Commission through the ERC Advanced Investigator Grant 'Momentum' (Grant No. 324247). The authors thank Bary Pradelski, Anna Gunnthorsdottir, Michael Mäs, Stefan Seifert, Jiabin Wu, Yoshi Saijo, Yuji Aruka, and Guillaume Hollard for helpful discussion and comments on earlier drafts, and finally members of GESS at ETH Zurich as well as seminar participants at the *Behavioral Studies Colloquium* at ETH Zürich, at the 25^{th} *International Conference on Game Theory 2014* at Stony Brook, at the *Choice Group* at LSE, at the *TOM Seminar* at PSE and at the Kochi University of Technology for helpful feedback. All remaining errors are ours.

References

Adams, J.S. (1965). "Advances in Experimental Social Psychology". In: ed. by L. Berkowitz. Vol. 2. New York: Academic Press. Chap. Inequity in Social Exchange, pp. 267–299.

Ahn, T., R.M. Isaac, and T.C. Salmon (2008). "Endogenous Group Formation". In: *Journal of Public Economic Theory* 10.2, pp. 171–194.

Andreoni, J. (1988). "Why Free Ride? Strategies and Learning in Public Goods Experiments". In: *Journal of Public Economics* 37.3, pp. 291–304.

Arrow, K., S. Bowles, and S. Durlauf (2000). *Meritocracy and Economic Inequality*. Princeton University Press.

Arrow, K.J. (1951). *Social Choice and Individual Values*. Yale, USA: Yale University Press.

Atkinson, A. B. (2012). "Public Economics after the Idea of Justice". In: *Journal of Human Development and Capabilities* 13.4, pp. 521–536.

Atkinson, A.B. (1970). "On the Measurement of Inequality". In: *Journal of Economic Theory* 2.3, pp. 244–263.

Balietti, S. (2014). *nodeGame: Real-Time Social Experiments in the Browser*. http://nodegame.org.

Bentham, J. (1907). *An Introduction to the Principles of Morals and Legislation*. Clarendon Press.

Bohm, R. and B. Rockenbach (2013). "The Inter-Group Comparison – Intra-Group Cooperation Hypothesis". In: *PLoS ONE* 8, p. 56152.

Bornstein, G. and I. Erev (1994). "The Enhancing Effect of Intergroup Competition on Group Performance". In: *International Journal of Conflict Management* 5.3, pp. 271–283.

Bornstein, G., I. Erev, and O. Rosen (1990). "Intergroup Competition as a Structural Solution to Social Dilemmas". In: *Social Behaviour* 5, pp. 247–260.

Bornstein, G., U. Gneezy, and R. Nagel (2002). "The Effect of Intergroup Competition on Group Coordination: An Experimental Study". In: *Games and Economic Behavior* 41.1, pp. 1–25.

Brekke, K., K. Nyborg, and M. Rege (2007). "The Fear of Exclusion: Individual Effort when Group Formation is Endogenous". In: *Scandinavian Journal of Economics* 109.3, pp. 531–550.

Brekke, K. et al. (2011). "Playing with the Good Guys. A Public Good Game with Endogenous Group Formation". In: *Journal of Public Economics* 95.9, pp. 1111–1118.

Buckley, E. and R. Croson (2006). "Income and Wealth Heterogeneity in the Voluntary Provision of Linear Public Goods". In: *Journal of Public Economics* 90.4-5, pp. 935–955.

Charness, G.B. and C.-L. Yang (2008). "Endogenous Group Formation and Public Goods Provision: Exclusion, Exit, Mergers, and Redemption". In: *Available at SSRN 932251*.

Chaudhuri, A. (2011). "Sustaining Cooperation in Laboratory Public Goods Experiments: a Selective Survey of the Literature". In: *Experimental Economics* 14, pp. 47–83.

Cinyabuguma, M., T. Page, and L. Putterman (2005). "Cooperation Under the Threat of Expulsion in a Public Goods Experiment". In: *Journal of Public Economics* 89.8, pp. 1421–1435.

Coricelli, G., D. Fehr, and G. Fellner (2004). "Partner Selection in Public Goods Experiments". In: *Economics Series* 48.3, pp. 356–378.

Cowell, F. (2011). *Measuring Inequality*. Oxford University Press.

Ehrhart, K. and C. Keser (1999). *Mobility and Cooperation: On the Run*. Tech. rep. s-24. Cirano.

Erev, I., G. Bornstein, and R. Galili (1993). "Constructive Intergroup Competition as a Solution to the Free Rider Problem: A Field Experiment". In: *Journal of Experimental Social Psychology* 29.6, pp. 463–478.

Erev, I., E. Ert, and E. Yechiam (2008). "Loss Aversion, Diminishing Sensitivity, and the Effect of Experience on Repeated Decisions". In: *Journal of Behavioral Decision Making* 21.5, pp. 575–597.

Fehr, E. and S. Gächter (2000). "Cooperation and Punishment in Public Goods Experiments". In: *American Economic Review* 90, pp. 980–994.

Fehr, E. and K.M. Schmidt (1999). "A Theory of Fairness, Competition, and Cooperation". In: *Quarterly Journal of Economics* 114, pp. 817–868.

Feldman, A. (1980). *Welfare Economics and Social Choice Theory*. Boston, USA: Martinus Nijhoff Publishing.

Fischbacher, U., S. Schudy, and S. Teyssier (2014). "Heterogeneous Reactions to Heterogeneity in Returns From Public Goods". In: *Social Choice and Welfare* 43.1, pp. 195–217.

Gauthier, D.P. (1986). *Morals by Agreement*. New York: Oxford University Press.

Greenberg, J. (1987). "A Taxonomy of Organizational Justice Theories". In: *Academy of Management review* 12.1, pp. 9–22.

Gunnthorsdottir, A. and P. Thorsteinsson (2011). "Tacit Coordination and Equilibrium Selection in a Merit-Based Grouping Mechanism: A Cross-Cultural Validation Study". In: *Available at SSRN 1883465*.

Gunnthorsdottir, A., R. Vragov, and J. Shen (2010). "Tacit Coordination in Contribution-Based Grouping with Two Endowment Levels". In: *Research in Experimental Economics* 13, pp. 13–75.

Gunnthorsdottir, A. et al. (2010). "Near-Efficient Equilibria in Contribution-Based Competitive Grouping". In: *Journal of Public Economics* 94.11, pp. 987–994.

Hardin, H. (1968). "The Tragedy of the Commons". In: *Science* 162, pp. 1243–1248.

Harsanyi, J.C. and R. Selten (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Hayek, F.A. von (1935). "Collectivist Economic Planning". In: ed. by F.A. von Hayek. Chap. The Nature and History of the Problem, pp. 1–40.

Isaac, Mark R., Kenneth F. McCue, and Charles R. Plott (1985). "Public Goods Provision in an Experimental Environment". In: *Journal of Public Economics* 26, pp. 51–74.

J.P., Rabanal and O.A. Rabanal (2015). "Efficient Investment via Assortative Matching: a Laboratory Experiment". In: *Available at SSRN 2565196*.

Kahneman, D. and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk". In: *Econometrica* 47.2, pp. 263–291.

King, R.G. and S. Rebelo (1990). *Public Policy and Economic Growth: Developing Neoclassical Implications*. Tech. rep. National Bureau of Economic Research.

Ledyard, J.O. (1997). "Public Goods: A Survey of Experimental Research". In: *The Handbook of Experimental Economics*. Ed. by J. H. Kagel and A. E. Roth. Princeton, NJ: Princeton University Press, pp. 111–194.

Mises, L. von (1922). *Die Gemeinwirtschaft: Untersuchungen über den Sozialismus*. Jena, Germany: Gustav Fischer Verlag.

Nash, J. (1950). "Equilibrium Points in N-Person Games". In: *Proceedings of the National Academy of Sciences (PNAS)* 36, pp. 48–49.

— (1951). "Non-Cooperative Games". In: *Annals of Mathematics* 54, pp. 286–295.

Nax, H.H., R.O. Murphy, and D. Helbing (2014). *Stability and Welfare of 'Merit-Based' Group-Matching Mechanisms in Voluntary Contribution Games*. Submitted.

Ockenfels, Axel and Gary E. Bolton (2000). "ERC: A Theory of Equity, Reciprocity, and Competition". In: *American Economic Review* 90.1, pp. 166–193.

Okun, A.M. (1975). *The Big Tradeoff*. Washington D.C.: Brookings Institution Press.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, U.K.: Cambridge University Press.

— (1999). "Coping with Tragedies of the Commons". In: *Annual Review Political Science* 2, pp. 493–535.

Page, T., L. Putterman, and B. Unel (2005). "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency". In: *The Economic Journal* 115.506, pp. 1032–1053.

Rawls, J. (1971). *A Theory of Justice*. Belknap Press.

Rebelo, S. (1991). "Long-Run Policy Analysis and Long-Run Growth". In: *Journal of Political Economy* 99.3, pp. 500–521.

Samuelson, P.A. (1980). *Foundations of Economic Analysis*. Cambridge, USA: Harvard University Press.

Sen, Amartya (1970). "The Impossibility of a Paretian Liberal". In: *Journal of Political Economy* 78.1, pp. 152–57.

Tamai, T. (2010). "Public Goods Provision, Redistributive Taxation, and Wealth Accumulation". In: *Journal of Public Economics* 94.11-12, pp. 1067–1072.

Tversky, A. and D. Kahneman (1991). "Loss Aversion in Riskless Choice: A Reference Dependent Model". In: *Quarterly Journal of Economics* 106.4, pp. 1039–1061.

Young, M. (1958). *The Rise of the Meritocracy, 1870-2033: An Essay on Education and Equality*. Transaction Publishers.

A Materials and Methods

A.1 Equilibrium structure

Our stage games with $n = 16$, $s = 4$, $B = 20$ and $m = 0.5$ have the following equilibria dependent on which variance level of $\sigma^2 = \{0, 3, 20, \infty\}$ is played. When $\sigma^2 = \infty$ (NO-MERIT), the only equilibrium is $c_i = 0$ for all i. $c_i = 0$ for all i is also an equilibrium for all other variance levels. In that equilibrium, all players receive a payoff of $\phi_i = 20$. However, when $\sigma^2 = \{0, 3, 20\}$, there also exist exactly $\binom{n}{k}$ unique pure-strategy equilibria such that $c_i = 0$ for exactly two agents and $c_j = 20$ for the remaining fourteen. In that equilibrium, for the case when $\sigma^2 = 0$ (PERFECT-MERIT), payoffs are such that twelve of the fourteen players who contribute $c_i = 20$ are matched in groups with each other and receive $\phi_i = 40$. The remaining four players are matched in the worst group. Of those, the two players who contribute $c_i = 0$ receive a payoff of $\phi_i = 40$, while the two players who contribute $c_i = 20$ receive a payoff of $\phi_i = 20$. For the cases when $\sigma^2 = 3$ (HIGH-MERIT)/$\sigma^2 = 20$ (LOW-MERIT), payoffs in the last group are as in the case when $\sigma^2 = 0$ (PERFECT-MERIT) in over 99.9%/ 99% of all cases. In the remaining cases, payoffs are such that 6 out of fourteen players who contribute $c_i = 20$ are matched in groups with each other and receive $\phi_i = 40$. The remaining 6 players who contribute $c_i = 20$ are matched in a group with one player who contributes $c_i = 0$ and receives a payoff of 30. The two players who contribute $c_i = 0$ receive a payoff of $\phi_i = 50$ each. The near-efficient Nash equilibrium collapses when the variance reaches a level of about $\sigma^2 = 100$ (see propositions 6 and 7 in Ref. (Nax, Murphy, and Helbing, 2014)).

A.2 Experimental design

A total of 192 voluntary participants took part in one session consisting of two separate games each. Each session lasted roughly one hour. There were 16 participants in each session and 12 sessions in total. All sessions were conducted at the ETH Decision Science Laboratory (DeSciL) in Zürich, Switzerland, using the experimental software NodeGame (Balietti, 2014). DeSciL recruited the subjects using the Online Recruitment System for Economic Experiments (ORSEE). The experiment followed all standard behavioral economics procedures and meets the ethical committee guidelines. Decisions, earnings and payments were anonymous. Payments were administered by the DeSciL administrators. In addition to a 10 CHF show-up fee, each subject was paid according to a known exchange rate of 0.01 CHF per coin. Overall, monetary rewards ranged from 30 to 50 CHF, with a mean of 39 CHF.

Each session consisted of two games, each of which was a forty-round repetition of the same underlying stage game, namely a public-goods game. The same fixed budget was given to each subject every period. Each game had separate instructions that were distributed at the beginning of each game. After reading the instructions, all participants were quizzed to make sure they understood the task. The two games differ with respect to the variance level that is added to players' contributions. There were four variance levels ($\sigma^2 = \{0, 3, 20, \infty\}$), and each game had equivalent instructions. Instructions contained full information about the structure of the game and about the payoff consequences to themselves and to the other agents. We played every possible pair of variance levels in both orders to have an orthogonal balanced design, which yields a total of 12 sessions. As the game went on, players learnt about the other players' previous actions and about the groups that formed. Each of our 192 participants made forty contribution decisions in each of the two games in his session. This yields 80 choices per person per session, hence a total of 15,360 observations. More details, including a copy of a full instructions set, are provided in the following subsections.

Instructions of the lab experiment

Each experimental session consisted of two separate games (part 1, part 2), each played with a different variance level. We exhausted all possible pair of variance levels in both orders, for a total of 12 different

combinations. Consequently, we prepared 12 different instruction texts that took into account whether a variance level was played in the first or in the second part, and in the latter case also considered which variance level was played in part 1.

Together with the main instructions sheet, we provided an additional sheet containing tabulated numerical examples of fictitious game-rounds played at the current variance level. This aimed to let participants get an intuitive feeling of the consequences of noise on contributions and final payoffs.

All instructions texts can viewed at the address http://nodegame.org/games/merit/. Here we report the instruction text for variance level equal 20 played in the part 1.

Instructions for Variance Level = 20, Part 1

Welcome to the experiment and thanks for your participation. You have been randomly assigned to an experimental condition with 16 people in total. In other words you and 15 others will be interacting via the computer network for this entire experimental session.

The experiment is divided into two parts and each part will last approximately 30-40 minutes long. Both parts of the experiment contribute to your final earnings. The instructions for the first part of the experiment follow directly below. The instructions for the second part of the experiment will be handed out to you only after all participants have completed the first part of the experiment. It is worth your effort to read and understand these instructions well. You will be paid based on your performance in this study; the better you perform, the higher your expected earnings will be for your participation today.

Your decision.

In this part you will play 40 independent rounds. At the beginning of each round, you will receive 20 "coins". For each round, you will have to decide how many of your 20 coins to transfer into your "personal" account, and how many coins to transfer into a "group" account. Your earnings for the round depend on how you and the other participants decide to divide the coins you have received between the two accounts.

Group matching with noise.

For each round you will be assigned to a group of 4 people, that is, you and three other participants. In general, groups are formed by ranking each individual transfer to the group account, from the highest to the lowest. Group 1 is generally composed of those participants who transferred the most to the group account; Group 4 is generally composed of those who transferred the least to the group account. The other groups (2 and 3) are between these two extremes.

However, the sorting process is noisy by design; contributing more will increase a participant's chances of being in a higher ranked group, but a high ranking is not guaranteed. Technical note- The noisy ranking and sorting is implemented with the following process:

1. ***Step 1:*** Preliminary ordering. A preliminary list is created in which transfers to the group account are ranked from highest to lowest. In case two or more individuals transfer the same amount, their relative position in the ranking will be decided randomly.

2. ***Step 2:*** Noisy ordering. From every participant's actual transfer to the group account, we obtain a unique noisy contribution by adding an i.i.d. (independent and identically distributed) normal variable with mean 0 and variance 20. The noisy contributions are then ranked from 1 to 16 from highest to lowest, and a final list is created.

3. ***Step 3:*** Group matching. Based on the final list created at Step 2 (the list with noise), the first 4 participants on that list form Group 1, the next 4 people in the list form Group 2, the third 4 people in the list form Group 3, and the last 4 people form Group 4.

Return from personal account.

Each coin that you put into your personal account results in a simple one-to-one payoff towards your total earnings.

Return from group account.

Each coin that you put into the group account will pay you back some positive amount of money, but it depends also on how much the other group members have transferred to the group account, as described below.

The total amount of coins in your group account is equal to the sum of the transfers to the group account by each of the group members. That amount is then multiplied by 2 and distributed equally among the 4 group members. In other words, you will get a return equal to half of the group account total.

Final Earnings

Your total earnings for the first part of the experiment are equal to the sum of all your rounds' earnings. One coin is equal to 0.01 CHF. This may not appear to be very much money, but remember there are 40 rounds in this part of the experiment so these earnings build up.

Example

Here is an example of one round to demonstrate this decision context, the noisy sorting into different groups, and the different resulting payoffs. In the table below, pay attention to the following facts:

- Groups are roughly formed by ranking how much participants transferred to the group account, but this is not a perfect ranking. For example, participant #8 transferred less to the group account than participant #10, but the noisy sorting process placed him in a higher ranked group.
- Participant #7 transferred 14 of his coins to the group account. This means that he transferred 6 to his personal account. Due to noisy sorting he was ranked first, and assigned to Group 1. The other participants in Group 1 transferred a total of 64 coins to the group account. This amount is doubled and redistributed evenly back to the 4 members of the groupthis is 32 for each participant. So then participant #7 earned 38 coins for this round.
- Participant #12 transferred 7 coins to the group account and transferred the remaining 13 coins to his personal account. He was sorted (with noise) into Group 3 and this group transferred 46 coins in total. This resulted in 23 coins being returned to each of the group members, and thus his total payoff is 36 coins (23 returned

  from the group account and the 13 he kept in his personal account).

Player ID	Group	Transfer to group account	Transfer to personal account	Total to group account	Amount returned to player	Total earnings for the round
7	1	14	6	64	32	38
6	1	13	7	64	32	39
14	1	16	4	64	32	36
4	1	8	12	64	32	44
1	2	14	6	51	25.5	31.5
3	2	20	0	51	25.5	25.5
8	2	11	9	51	25.5	34.5
11	2	19	1	51	25.5	26.5
10	3	17	3	46	23	26
12	3	7	13	46	23	36
16	3	6	14	46	23	37
5	3	16	4	46	23	27
9	4	10	10	18	9	19
2	4	1	19	18	9	28
13	4	5	15	18	9	24
15	4	2	18	18	9	27

Additional examples are provided in a separate sheet for your own reference.

Quiz

Subjects were given a quiz after instructions to test their understanding of the game. Only after "passing" the quiz were subjects allowed to begin play. Details about the quiz can be found at http://nodegame.org/games/merit/.

Graphical interface of the experiment

The experiment was implemented using the experimental software nodeGame (Balietti, 2014). Besides, offering a textual response of the actions of the players, we also offer a visual summary with contributions bars ordered by group, as shown in Fig. 6. More details about the interface, and the implementation are available at the url: http://nodegame.org/games/merit/

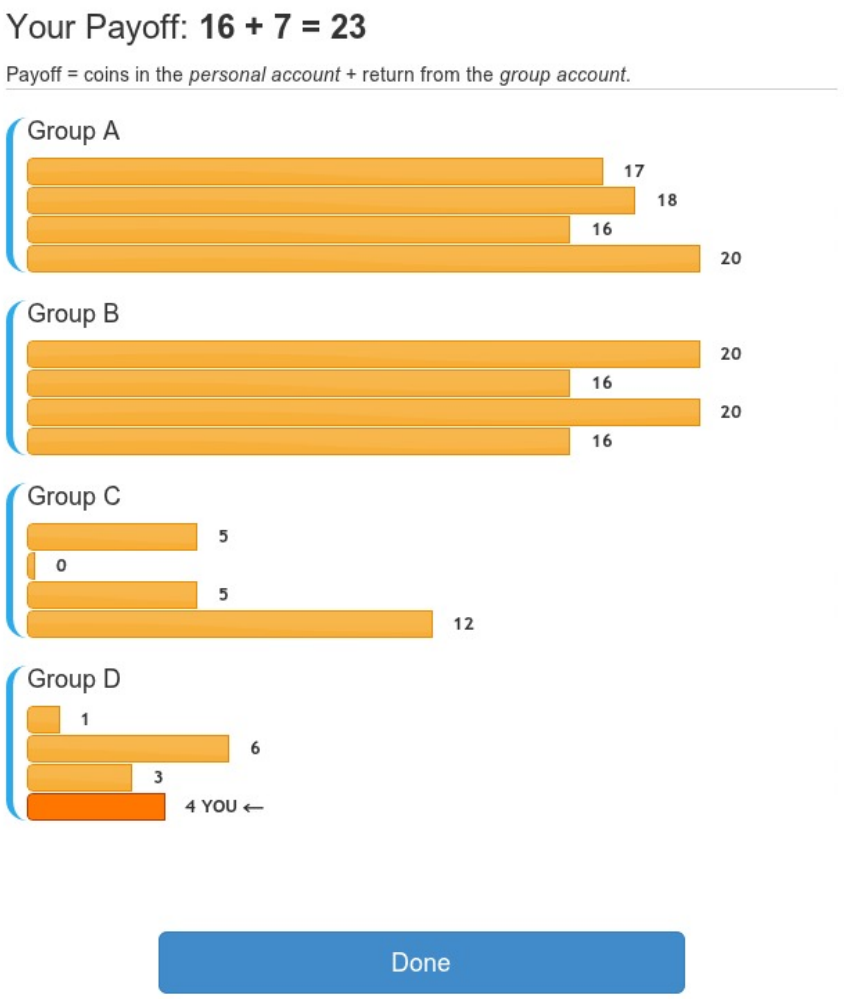

Figure 6. Game interface for displaying the results. Participants' contribution decisions are displayed as horizontal bars of variable length sorted according to their ranking after noise has been applied.

A.3 Statistical analyses

Equality analysis

Overall, we found a significant difference in the variance of realized payoffs in each round among the four treatments (LMM: $F_{3,8} = 7.27, P < 0.0113$). When computing Bonferroni adjusted p-values, the treatment with variance ∞ was found significantly different ($P = 0.0003; P = 0.0004; P = 0.0086$) from the other three variance levels ($\sigma^2 = \{0, 3, 20\}$), which are themselves not significantly different from each other. Taking NO-MERIT as a baseline, LOW-MERIT led to a decrease in the variance of realized payoffs in each round of -13.546 (LRT $\chi_{(1)} = 8.13, P = 0.0043$), HIGH-MERIT to a decrease of -16.914 (LRT $\chi_{(1)} = 9.89, P = 0.0016$), and PERFECT-MERIT to a decrease of -17.122 (LRT $\chi_{(1)} = 6.78, P = 0.0091$).

Similarly, the Gini index differs significantly among the four treatments (LMM: $F_{3,20} = 42.0, P < 0.0001$). Taking NO-MERIT as a baseline, LOW-MERIT led to a decrease in the variance of realized payoff in each round of -0.058901 (LRT $\chi_{(1)} = 18.18, P < 0.0001$), HIGH-MERIT to a decrease of -0.071843 (LRT $\chi_{(1)} = 22.28, P < 0.0001$), and PERFECT-MERIT to a decrease of -0.075453 (LRT $\chi_{(1)} = 22.06, P < 0.0001$). Computing Bonferroni adjusted p-values for all pair-wise differences reveals that the treatment with variance ∞ is significantly different ($P < 0.0001$) from the other three variance levels ($\sigma^2 = \{0, 3, 20\}$), which are themselves not significantly different from each other (see Fig. 4).

Fairness analysis

We find a significant difference in the experienced levels of meritocratic unfairness in each round among the four treatments (LMM: $F_{3,8} = 53.74, P < 0.0001$). When computing Bonferroni adjusted p-values we find that – excluding PERFECT-MERIT for which meritocratic unfairness is always zero by definition – all treatments are statistically significantly different from each other (HIGH-MERIT vs LOW-MERIT $P = 0.0071$, all the other pair-wise comparisons $P < 0.0001$). Taking NO-MERIT as a baseline, LOW-MERIT led to a decrease in the experienced meritocratic unfairness in each round of -1.66 (LRT $\chi_{(1)} = 11.76, P = 0.0006$), HIGH-MERIT to a decrease of -2.36 (LRT $\chi_{(1)} = 18.92, P < 0.0001$).

We also analyzed the effect of meritocratic (dis)advantageous unfairness on contribution adjustments between rounds, by performing a multilevel regression with subject and session as random effects. Our findings reveal that disadvantageous unfairness leads to decreases in treatments LOW-MERIT $-0.18^{***}(0.05)$, and NO-MERIT $-0.25^{***}(0.03)$). For HIGH-MERIT the decrease is consistent in sign and size, but not statistically significant $-0.39(0.21)$. However, if HIGH-MERIT and LOW-MERIT are pooled together the effect turns out to be significant $-0.25^{***}(0.03)$. Meritocratic disadvantageous fairness can, therefore, originate significant differences between the theoretical equilibrium predictions and experimentally observed behavior. Advantageous unfairness leads to increases under some but not under all regimes. Full regression tables are available in the remainder of this Appendix.

Fairness regressions

Here we report the results of the mixed-effects regressions of meritocratic and distributional fairness on contributions adjustments between rounds in part 1 and part 2 of the experiment. As we argued in the main text, distributional fairness cannot easily be generalized to the case of assortative matching. Here we show that a näive extension of the formula in (Fehr and Schmidt, 1999) fails to reproduce the results predicted by theory. In fact, both within-group and across-groups distributional fairness under assortativity often lead to the contradictory result that disadvantageous fairness implies an increase in the contribution levels. However, by taking into account assortativity in the formula of distributional fairness, we developed an extension that is able to reproduce the results predicted by the theory for all treatments.

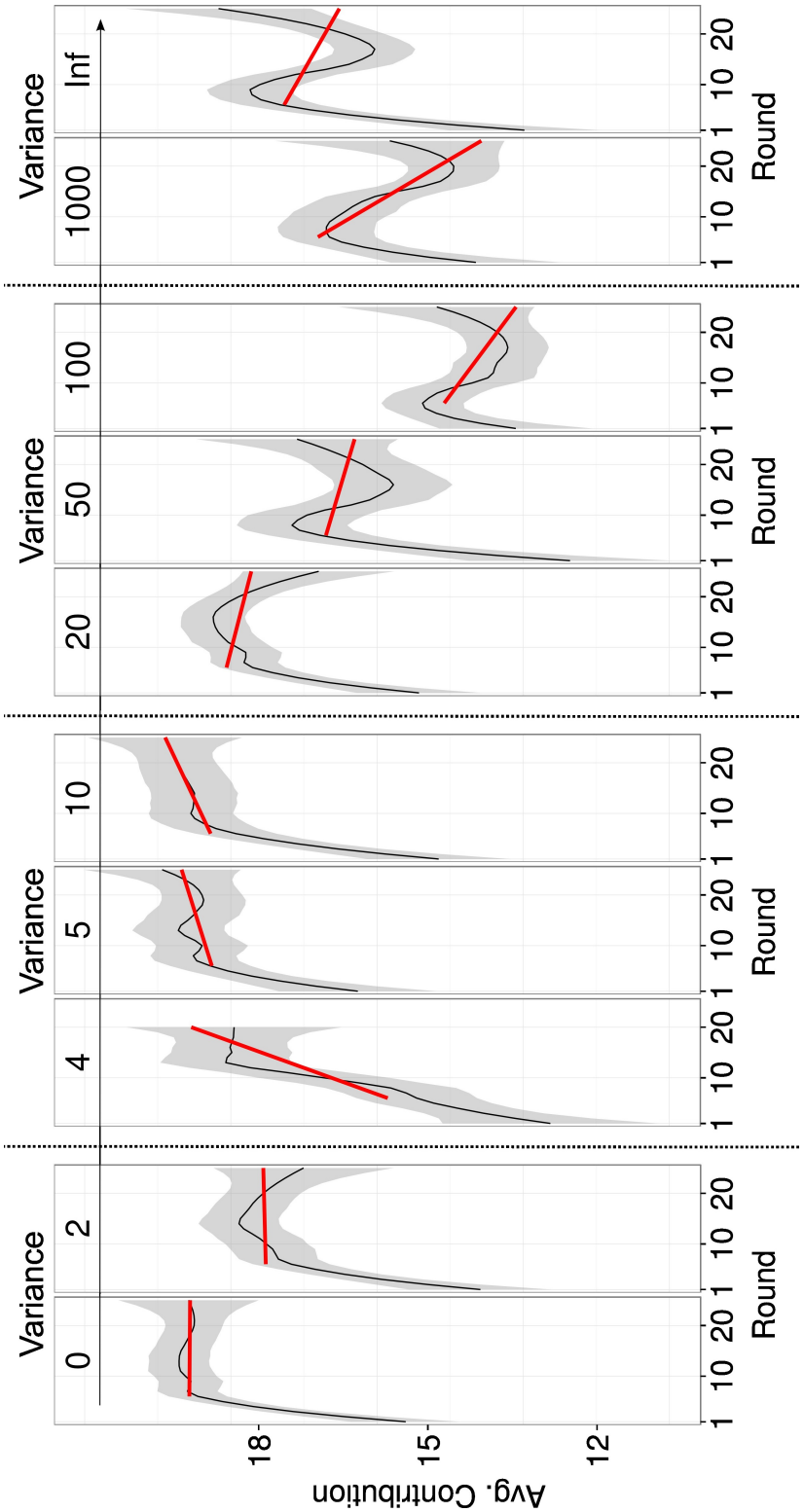

Figure 7. Average contribution levels over time for different levels of variance in experiments played online. Approximately four contribution regimes were found: (i) from 0 to 2 players' contributions stabilize immediately; (ii) from 4 to 10 players' contributions are increasing tending towards the high-efficiency Nash equilibrium; (iii) from 20 to 50 players' contributions are declining towards the zero-efficiency equilibrium; (iv) for extremely high-levels from 1000 to Infinity, the decline of players' contributions is even steeper. The red line shows a fitted linear regressions on the data excluding the first five rounds where players are still learning the dynamics of the game.

Meritocratic fairness

In tables 1 and 2, meritocratic unfairness is used as a predictor. `lag.merit.fair.dis` and `lag.merit.fair.adv` are respectively the amount of *disadvantageous* and *advantageous* meritocratic unfairness experienced by a player in the previous round, measured according to the equations in Section 2 of the main text.

Table 1. Meritocratic fairness predicts contribution differential. (Part 1) The sign of the regression coefficient is always consistent with theory predictions. HIGH-MERIT is significant if pooled together with LOW-MERIT.

	HIGH-MERIT	LOW-MERIT	HIGH-MERIT&LOW-MERIT	NO-MERIT
(Intercept)	0.25	0.15	0.03	0.03
	(0.16)	(0.16)	(0.19)	(0.19)
lag.merit.fair.dis	−0.39	−0.18***	−0.25***	−0.25***
	(0.21)	(0.05)	(0.03)	(0.03)
lag.merit.fair.adv	−0.91**	0.06	0.15***	0.15***
	(0.30)	(0.06)	(0.03)	(0.03)
AIC	12314.36	12284.05	12359.50	12359.50
BIC	12347.56	12317.24	12392.70	12392.70
Log Likelihood	-6151.18	-6136.02	-6173.75	-6173.75
Num. obs.	1872	1870	1872	1872

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2. Meritocratic fairness predicts contribution differential. (Part 2) The sign of the regression coefficient is always consistent with theory predictions. HIGH-MERIT is significant if pooled together with LOW-MERIT.

	HIGH-MERIT	LOW-MERIT	HIGH-MERIT&LOW-MERIT	NO-MERIT
(Intercept)	0.13	0.16	0.11	0.38*
	(0.16)	(0.17)	(0.11)	(0.18)
lag.merit.fair.dis	−0.45	−0.29***	−0.29***	−0.26***
	(0.28)	(0.07)	(0.06)	(0.02)
lag.merit.fair.adv	−0.57	0.00	−0.02	0.04
	(0.32)	(0.07)	(0.07)	(0.02)
AIC	12288.63	12419.05	24699.24	12123.03
BIC	12321.83	12452.25	24736.60	12156.23
Log Likelihood	-6138.31	-6203.53	-12343.62	-6055.51
Num. obs.	1872	1871	3743	1872

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Distributional fairness

The results of the regressions for distributional fairness are shown in tables 3, 4, 5 and 6. Based on the original formula in Ref. (Fehr and Schmidt, 1999), we tried two different extensions of the notion of distributional fairness for meritocratic environments. First, we computed distributional fairness for each player only taking into account the other players within the group into which he/she was matched (Within-group distributional fairness). The regressors in this case are called: `lag.distr.fair.group.dis` and `lag.distr.fair.group.adv`. Then, we also computed distributional fairness across all players, regardless

of the group they belonged to (Across-group distributional fairness). The regressors for across-group distributional fairness are called: `lag.distr.fair.dis` and `lag.distr.fair.adv`.

Table 3. Within-group distributional fairness predicts contribution differential. (Part 1)
The sign of the regression coefficient is often inconsistent with theory predictions.

	PERFECT-MERIT	HIGH-MERIT	LOW-MERIT	HIGH-MERIT & LOW-MERIT	NO-MERIT
(Intercept)	-0.79^{***}	-1.39^{***}	-1.32^{***}	-1.39^{***}	1.40^{**}
	(0.23)	(0.22)	(0.21)	(0.15)	(0.45)
lag.distr.fair.group.dis	-0.03	0.13^{**}	0.01	0.06^{*}	-0.70^{***}
	(0.04)	(0.05)	(0.05)	(0.03)	(0.04)
lag.distr.fair.group.adv	0.76^{***}	0.99^{***}	0.77^{***}	0.88^{***}	0.28^{***}
	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)
AIC	11682.40	11933.18	12025.27	23952.86	11968.23
BIC	11715.59	11966.38	12058.46	23990.22	12001.43
Log Likelihood	-5835.20	-5960.59	-6006.64	-11970.43	-5978.12
Num. obs.	1872	1872	1870	3742	1872

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4. Within-group distributional fairness predicts contribution differential. (Part 2)
The sign of the regression coefficient is often inconsistent with theory predictions.

	PERFECT-MERIT	HIGH-MERIT	LOW-MERIT	HIGH-MERIT & LOW-MERIT	NO-MERIT
(Intercept)	-0.93^{***}	-1.54^{***}	-1.25^{***}	-1.43^{***}	1.60^{***}
	(0.25)	(0.40)	(0.23)	(0.22)	(0.38)
lag.distr.fair.group.dis	-0.10^{*}	0.05	-0.06	0.00	-0.61^{***}
	(0.04)	(0.04)	(0.05)	(0.03)	(0.03)
lag.distr.fair.group.adv	0.88^{***}	1.19^{***}	0.86^{***}	1.02^{***}	0.15^{***}
	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)
AIC	11856.01	11799.36	12109.33	23935.12	11827.92
BIC	11889.21	11832.55	12142.53	23972.48	11861.12
Log Likelihood	-5922.01	-5893.68	-6048.67	-11961.56	-5907.96
Num. obs.	1871	1872	1871	3743	1872

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5. Across-group distributional fairness predicts contribution differential. (Part 1)
The sign of the regression coefficient is often inconsistent with theory predictions.

	PERFECT-MERIT	HIGH-MERIT	LOW-MERIT	HIGH-MERIT & LOW-MERIT	NO-MERIT
(Intercept)	-1.42^{***}	-2.40^{***}	-2.20^{***}	-2.23^{***}	1.04^{*}
	(0.26)	(0.34)	(0.34)	(0.24)	(0.40)
lag.distr.fair.dis	0.22^{***}	0.39^{***}	0.33^{***}	0.35^{***}	-0.44^{***}
	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)
lag.distr.fair.adv	0.44^{***}	0.59^{***}	0.43^{***}	0.48^{***}	0.13^{*}
	(0.08)	(0.10)	(0.08)	(0.06)	(0.05)
AIC	11934.03	12223.59	12225.86	24434.15	12277.90
BIC	11967.23	12256.79	12259.05	24471.51	12311.10
Log Likelihood	-5961.02	-6105.80	-6106.93	-12211.07	-6132.95
Num. obs.	1872	1872	1870	3742	1872

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6. Across-group distributional fairness predicts contribution differential. (Part 2)
The sign of the regression coefficient is often inconsistent with theory predictions.

	PERFECT-MERIT	HIGH-MERIT	LOW-MERIT	HIGH-MERIT & LOW-MERIT	NO-MERIT
(Intercept)	-2.15^{***}	-1.98^{***}	-2.19^{***}	-2.01^{***}	1.96^{***}
	(0.30)	(0.30)	(0.35)	(0.23)	(0.48)
lag.distr.fair.dis	0.21^{***}	0.29^{***}	0.30^{***}	0.29^{***}	-0.49^{***}
	(0.03)	(0.03)	(0.04)	(0.02)	(0.04)
lag.distr.fair.adv	0.65^{***}	0.54^{***}	0.46^{***}	0.48^{***}	-0.04
	(0.09)	(0.09)	(0.09)	(0.06)	(0.04)
AIC	12162.64	12222.36	12374.95	24584.87	12068.03
BIC	12195.83	12255.56	12408.15	24622.23	12101.23
Log Likelihood	-6075.32	-6105.18	-6181.48	-12286.43	-6028.02
Num. obs.	1871	1872	1871	3743	1872

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

A.4 Additional inequality indexes

As stated in the main text, inequality decreases as meritocracy increases. In this section, we show that our finding is robust to the type of inequality measurement chosen. Fig. 8 displays the payoff inequality as measured by a number of different indexes commonly found in the literature of inequality studies (Atkinson, 1970).

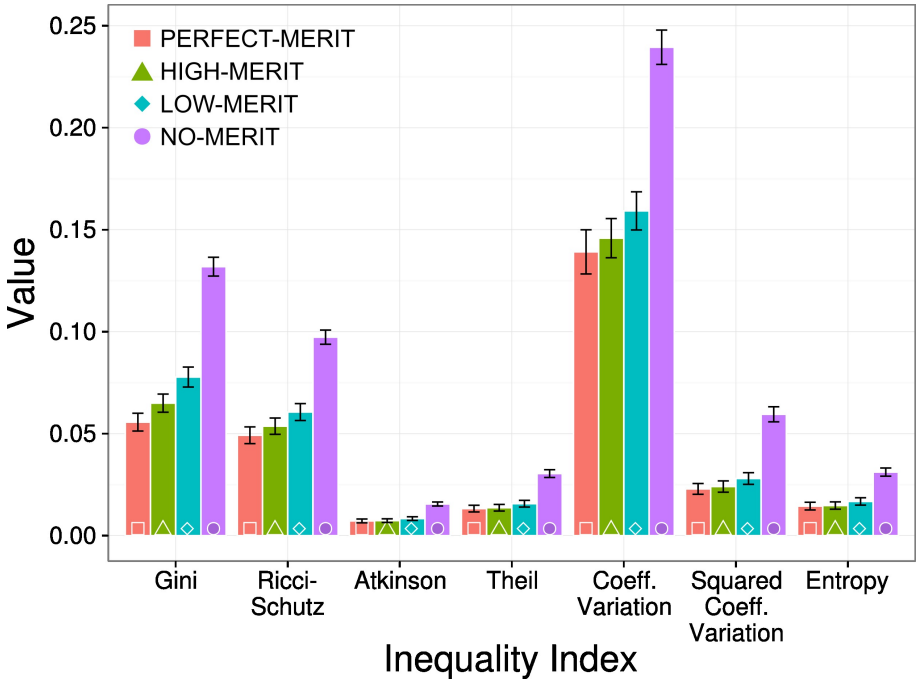

Figure 8. Battery of indexes measuring payoff inequality over the forty rounds for perfect-, high-, low-, and no-meritocracy, respectively associated with the values of $\sigma^2 = \{0, 3, 20, \infty\}$**.** Inequality decreases with meritocracy for a large number of distinct inequality indexes. Error bars represent the 95%-confidence intervals

A.5 Implications

Our model implies that situations consistent with our model assumptions would benefit from higher degrees of meritocracy, both in terms of efficiency and in terms of equality. This positive result relies on several features of the underlying model. It is an avenue for future research to consider these generalizations. First, our model describes an *ex ante* homogeneous population. Differences in payoff are driven by differences in actions and by neutral stochastic elements alone. Heterogeneity in priority given by the matching mechanism and/or heterogeneities in the individual rates of return could influence the results. This is true for any public-goods game including the standard models with random interactions (e.g. (Buckley and Croson, 2006; Fischbacher, Schudy, and Teyssier, 2014)). However, it should be noted that meritocracy may actually mitigate the associated inequality problems. Second, related to heterogeneity, our model allows for no wealth creation, that is, individuals receive a new budget every period and the size of this budget is fixed and constant over time. Players cannot accumulate wealth. The role of wealth creation in public-goods games has received some attention and has been shown to lead to the emergence of different classes of contributions and income (e.g. (Tamai, 2010), see also (King and Rebelo, 1990;

Rebelo, 1991)). Under assortative matching, wealth creation can be problematic as it allows rich players to block out poor players. Third, group sizes are fixed. Alternative models have been proposed (e.g. (Cinyabuguma, Page, and Putterman, 2005; Charness and Yang, 2008; Ehrhart and Keser, 1999; Ahn, Isaac, and Salmon, 2008; Coricelli, Fehr, and Fellner, 2004; Page, Putterman, and Unel, 2005; Brekke, Nyborg, and Rege, 2007; Brekke et al., 2011)).