Identity and Underrepresentation: Interactions between Race and Gender[*]

September 28, 2020

Jean-Paul Carvalho
University of California, Irvine

Bary Pradelski
CNRS

Abstract

Economic outcomes vary significantly across socio-demographic groups. In a model of multi-dimensional identity, we show how economic underrepresentation can evolve through identity-specific norms and stereotypes. Whereas standard approaches treat identity dimensions as independent, our analysis reveals deep connections between inequality and underrepresentation based on race, gender and other characteristics. 'Sterilized interventions' along a single identity dimension are impossible. Interventions that aim to reduce underrepresentation along one identity dimension can increase underrepresentation along another. Underrepresentation can be eliminated along every dimension through a system of (a) self-financing subsidies or (b) role models. Subsidies need to be intersectional, whereas role-model approaches are simpler but less efficient. We identify challenges in the design of such policies, including informational, political economy and efficiency concerns. This opens up new possibilities for theoretical and empirical work on the multi-dimensionality of identity.

Key words: Identity; education; labor force participation; inequality; underrepresentation; multi-dimensional.

JEL classification: D10; D63; D71; I24; J2; Z12 ; Z13

[*] *Carvalho*: Department of Economics and Institute for Mathematical Behavioral Sciences, University of California, Irvine, 3151 Social Science Plaza, Irvine, CA 92697, United States of America, jpcarv@uci.edu. *Pradelski*: Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France, bary.pradelski@cnrs.fr.
The paper has benefited from comments by George Akerlof, Robert Akerlof, Alberto Bisin, Rob Boyd, Isabelle Brocas, Paul Collier, Vitalii Emelianov, Paul Klemperer, Natalia Komarova, Rachel Kranton, Avital Livny, Dilip Mookherjee, Roger Myerson, Louis Narens, Andrew Newman, Jonathan Newton, John Quah, Phil Reny, Michael Sacks, Stergios Skaperdas, Dennis Snower, Cole Williams, Peyton Young and participants in the London Summer Colloquium on Identity Economics, the 2019 ASREC conference, the 2019 Southwest Economic Theory Conference, the 2019 ThReD Conference, the 2019 IMBS Conference on Cultural Evolution and Social Norms, and seminar participants at Oxford, USC, Warwick, Oregon and UC Irvine. Nishtha Sharma, Nikhil Addleman and Diego Nuñez Duran provided excellent research assistance. This paper incorporates material that previously appeared in a working paper titled 'Resisting Education'. All errors are ours.

1 Introduction

Inequality is back at the forefront of economics (e.g. Piketty & Saez 2003, Piketty 2014). Attention has recently turned to the socio-demographic structure of inequality, including variation in economic outcomes based on race, gender, class, and age cohort (e.g. U.S. Department of Education 2017, Daly et al. 2017, Bertrand 2020). An outstanding example is underrepresentation in the economics profession. While about 30 percent of the US population is Black or Hispanic, these groups comprise 11.3 percent of Assistant Professors, 9.0 percent of Associate Professors and 6.3 percent of full professors at PhD granting economics departments in the United States. Women make up only 29.1 percent of Assistant Professors, 25.6 percent of Associate Professors and 14.0 percent of full professors at the same institutions (Scott & Siegfried 2019). This has been the subject of panels titled "How Can Economics Solve its Gender Problem?" and "How Can Economics Solve its Race Problem?" at the American Economic Association's annual conferences in 2019 and 2020 respectively.

Building on the work of Akerlof & Kranton (2000, 2010) and Bordalo, Coffman, Gennaioli & Shleifer (2016, 2019), we show that extreme inequality and underrepresentation can arise even under symmetric conditions, through identity-specific norms and stereotypes. Moreover, underrepresentation is deeply connected across identity dimensions, such as race and gender. Ignoring the interaction between race and gender can lead to significant analytical and policy mistakes. It is common, and often necessary, to reduce a complex problem to several parts. This produces errors when the connections between the parts are neglected (Saari 2015, 2018). Existing theoretical, empirical and policy work in economics adopts this reductionist approach. Each identity dimension (e.g. race and gender) is treated in isolation, to identify causes of and solutions to underrepresentation. Statistics are reported in this way. Describing the European Union's gender policies, Skjeie (2015) states: "The dominant equality notion is mainly one-dimensional. What have recently been termed 'gender+' equality policies - i.e. policies which address gender inequalities in relation to other inequalities - are rather few and far between" [p. 79]. As multiple dimensions of identity are bundled in each person, there are connections between race, gender and other identity dimensions which must be accounted for. This has long been the focus of the literature on intersectionality (Crenshaw 1989, Cooper 2016), to which we will return.

In our model, group representation shapes economic participation through a process we call *the representation dynamic*, which is depicted by Figure 1. Through the feedback between representation

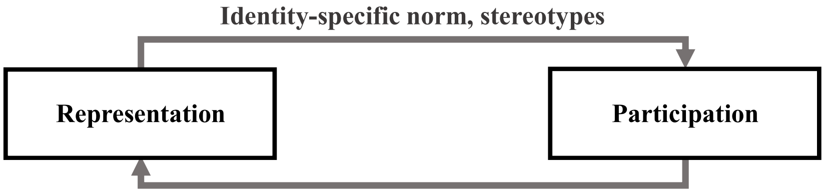


Figure 1: **The Representation Dynamic.** Group representation shapes economic participation due to identity-specific norms or stereotypes. Participation rates taken together determine representation. Thus a feedback loop exists between representation and participation.

and participation, underrepresentation becomes self-reinforcing. This is consistent with current policy debates that focus on achieving greater representation of women and minorities in education and elite professions (Reuben et al. 2015, Leslie et al. 2015, Cheryan et al. 2017). The push for safe spaces for racial and sexual minorities on college campuses indicates that representation, and not just participation, is important to feelings of belonging. Speaking of the wall of portraits of scientists at her university who have won either a Nobel Prize or the Lasker Award, Leslie Vosshall, neurobiologist at Rockefeller University, notes that all of them are male:

> "It just sends the message, every day when you walk by it, that science consists of old white men [...] I think every institution needs to go out into the hallway and ask, 'What kind of message are we sending with these oil portraits and dusty old photographs?' " (Greenfieldboyce 2019)

Accordingly, we suggest that an individual's economic participation (e.g. education, labor force) can depend on their groups' representation in that activity through (i) *identity-specific norms* that determine what is 'normal' or 'appropriate' for each identity (Akerlof & Kranton 2000, 2010) and (ii) *stereotypes* that affect self confidence and social expectations (Bordalo, Coffman, Gennaioli & Shleifer 2016, 2019). Our model is consistent with both of these psychological mechanisms, described in greater detail in Section 2. They can be interpreted as endogenous forms of discrimination, which are overcome by rising representation of the underrepresented group. A version of this dynamic is found in the study of mentorship by Athey, Avery & Zemsky (2000), which is discussed below.

First, we show that large and persistent differences in economic outcomes can arise even under symmetry—two equally sized, equally productive groups. The group with greater past representation enjoys better outcomes today, as the representation feedback locks in inequality and negative group shocks. Thus institutions such as slavery, segregation, and labor market discrimination against women can cast a long shadow. The problem for racial minorities is, however, more severe. For a minority group, equality of opportunity never results in equality of outcome, even when groups have the same productivity. The minority tends not only to have lower representation, but also to be underrepresented: its representation is less than its population share. Temporary big-push interventions might eliminate underrepresentation for women, but not for racial minorities.

Second, we study the interaction between race and gender. 'Sterilized interventions' along a single dimension are impossible. For example, if a policymaker treats representation along the gender dimension, she will also affect representation along the race dimension. In fact, we show that interventions that reduce underrepresentation along one identity dimension can increase underrepresentation along another. Thus the conventional approach of treating underrepresentation dimension by dimension can be counterproductive. We show underrepresentation can be eliminated along every dimension through (a) self-financing subsidies and (b) role models. In the case of self-financing subsidies (e.g. public investments in productivity), the policy must be intersectional, based not on the dimensions such as race and gender, but on the intersectional groups. For example, the subsidy to a black woman is not equal to the sum of the subsidies to black men and white women. Due to interactions between identity dimensions, introducing a new identity dimension changes the entire system of subsidies/taxes. In contrast, the policy of using role models to reduce underrepresentation need not be exactly intersectional, and can be computed using a sequential procedure. Hence it is less complex, requires less information, and is more robust to errors. It turns out, however, that role-model policies result in lower economic output than self-financing subsidies.

Finally, we describe further challenges in the design of such policies, including informational, political economy and efficiency concerns.

1.1 Related Literature

We make two contributions to the economics of identity. In the seminal model by Akerlof & Kranton (2000, 2010), each social category has its own behavioral prescription, which can be costly to violate because it is internalized and/or socially enforced. Hence identity-specific norms regarding education, labor force participation, and occupational choice create differences in economic outcomes among groups. Our first contribution is to endogenize these identity-specific norms based on representation. Alternative approaches to making identity-specific norms endogenous include Shayo (2009), Bénabou & Tirole (2011), Carvalho (2013), Kranton (2016), Snower & Bosworth (2016) and Akerlof (2017). Second, to our knowledge, ours is the first analytic model of underrepresentation with multi-dimensional identities. Existing work in economics examines inequality and underrepresentation based on race, gender and other identity dimensions independently (see reviews by Croson & Gneezy 2009, Bertrand 2011, Altonji & Blank 1999, Fang & Moro 2011). The interaction between identity dimensions has been largely ignored, with notable exceptions such as Akerlof (2017) on how individuals choose to value different dimensions of their identity and Elu & Loubert (2013) on how earnings and returns to schooling in sub-Saharan Africa depend on the interaction between gender and ethnicity.[1]

Brewer et al. (2002) call for economists to devote more attention to this issue, which has received substantial attention in other disciplines. According to the intersectional approach, pioneered by Crenshaw (1989), an individual's experience is not reducible to their race or gender, e.g. discrimination against black women is "greater than the sum of racism and sexism" [p. 140]. Policy should also be defined in intersectional terms. This has spawned a large literature (e.g. Cooper 2016, Collins & Bilge 2016). We are able to provide an analytic characterization of the interaction between race and gender, as well as the systemic consequences of failing to account for this interaction. For self-financing subsidies, we show that policy must be intersectional even when individuals care independently about each identity dimension, and not their intersectional group *per se*. A policy of promoting role models need not be intersectional, but interventions along each identity dimension are still connected.

In our dynamic analysis, inequality is connected over time and across identity dimensions through representation. The representation dynamic is distinct from existing participation-based theories of intergroup inequality. In participation models, an individual's payoff from economic participation is increasing in her group's participation rate. Theories of statistical discrimination (Arrow 1973, Coate & Loury 1993), intergenerational transfers of human capital (Becker & Tomes 1979, Loury 1977, 1981, Borjas 1992), social norms (Young 1998, 2015, Bertrand et al. 2015), learning (Chung 2000, Fernández 2013), peer effects, and local complementarities in education (Borjas 1992, Benabou 1993, Chaudhuri & Sethi 2008) all fall into this category. Our representation model retains increasing returns within groups and adds to it rivalry between groups: representation is a rival good (see Figure 1). This rivalry rules out Pareto improvements and produces more robust and severe forms of inequality than found in previous work. In particular, underrepresentation emerges in *every* equilibrium, not just some Pareto-dominated equilibria.[2]

[1] See also Sen (2006) on how multi-dimensional identities can be used to reduce conflict.

[2] Rivalries can arise in statistical discrimination models due to general equilibrium effects (Moro & Norman 2004). However, there is always a symmetric equilibrium in these models, which is generically not the case in ours.

The first example of a *representation dynamic* of which we are aware is analyzed by Athey, Avery & Zemsky (2000). When mentorship is subject to ingroup bias and firms are myopic, members of a group with low representation in the senior ranks of a firm have fewer mentoring opportunities, making it harder for them to advance. In independent work, Muller-Itten & Öry (2019) extend Athey et al. to account for differential group sizes, examining steady-state outcomes and policy under majority bias. Our work differs in two important ways. First, mentorship is distinct from the psychological mechanisms we study because the number of ingroup mentors available for a group k member depends on the participation rate of group k in the prior period, rather than the group's representation.[3] Second, and most importantly, these papers analyze uni-dimensional identities and do not address the interaction between race and gender, or other dimensions of identity.[4]

The remainder of the paper is structured as follows. Section 2 introduces the baseline representation model with uni-dimensional identities, illustrating differences between inequality based on race and gender. Section 3 contains the main analysis of multi-dimensional identities and in particular the interaction between race and gender. It derives policies that eliminate underrepresentation along every dimension and discusses design considerations including complexity and efficiency. Section 4 concludes.

2 The Representation Dynamic

In this section, we introduce the representation dynamic beginning with the case of uni-dimensional identities, before the analysis of two-dimensional identities in Section 3.

Consider a large, but finite population of agents N indexed by i. Time is discrete and indexed by $t = 0, 1, 2 \ldots$ Each period is a different cohort.

At $t = 0$, the population N is partitioned into two groups, N_A and N_B. The share of group $k \in \{A, B\}$ in the population is $m_k > 0$. We assume $m_A \geq \frac{1}{2}$ without loss of generality. Group shares are fixed for all time. Hence group identity can be gender, race, caste or any form of social categorization in which group sizes can be treated as exogenous.

In every period $t \geq 1$, each individual i chooses to participate in economic activity ($e_i = 1$) or

[3] The rivarly in Athey et al.'s model comes from the number of senior positions being fixed plus random assignment of mentors. The number of positions is not fixed in our model, as we are interested not in a single firm, but more broadly in the education system and labor market.

[4] A recent simulation-based analysis by O'Connor et al. (2019) examines multi-dimensional identities in a bargaining context.

not participate ($e_i = 0$). More generally, we can think of $e_i = 0$ and $e_i = 1$ as being two different types of occupations, with $e_i = 0$ being an occupation in which individuals interact primarily with members of their own group or in which identity is not important, while $e_i = 1$ is a socialized occupation in which mixing occurs and identity is important. For example:

- *Education choice.* $e_i = 1$ denotes college education.
- *Labor force participation.* $e_i = 1$ denotes joining the labor force.
- *Occupational choice.* For example, suppose i is female and j male. $e_i = 0$ denotes i's choice of a traditional female occupation and $e_j = 0$ denotes j's choice of a traditional male occupation. $e_i = 1$ ($e_j = 1$) denotes i's (j's) choice of an *ex ante* neutral profession (e.g. STEM field), where i's (j's) payoff from this choice depends on female (male) representation in the profession.
- *Competition.* $e_i = 1$ denotes participation in a highly competitive task, e.g. in a mixed-sex environment, whereas $e_i = 0$ is choosing not to compete or to compete in a single-sex environment.[5]

Participation choice is driven not only by standard economic incentives, but also by identity. Each agent has a two-dimensional (economic and social) type, (y, θ).

Economic. Participation $e = 1$ produces an economic benefit of y (net of costs) for an economic type y. For each individual, y is an i.i.d. draw from the c.d.f. F with associated p.d.f. f. We assume F is the same for each group, i.e. groups have identical productivity.[6] We refer to the component y as the economic return to participation. We normalize the economic benefit of non-participation ($e = 0$) to 0.

Social. Individuals belong to one of two groups A or B which are defined by a one-dimensional, binary identity characteristic.

The representation of group A in period t among individuals choosing to participate ($e = 1$) is

$$R^t = \frac{\sum_{i \in N_A} e_i^{t-1}}{\sum_{i \in N_A} e_i^{t-1} + \sum_{i \in N_B} e_i^{t-1}}, \tag{1}$$

[5] Gneezy et al. (2003) present experimental evidence that women tend to be less competitive than men, especially in mixed-sex settings. Consistent with identity-specific norms based on representation, Niederle & Vesterlund (2008) show that a quota which ensures women are equally represented induces women to compete more.

[6] In an earlier version of the paper, we included results for groups with different levels of productivity and/or discrimination. The identity-based effects we study amplify group differences.

which is the proportion of active individuals in the previous period who are from group A. Group B's representation is $1 - R^t$.

Representation can affect participation through two channels:

(i) *Identity-specific norms.* Building on Akerlof & Kranton (2000, 2010), the greater a group's representation in an activity, the more the activity is deemed 'normal' or 'appropriate' for its members.[7] For example, if historically breadwinners have been men and homemakers women, then women bear a higher (psychic and/or social) cost of labor force participation than men. This is consistent with evidence from social psychology and economics on the formation of gender-specific norms, as well as the quote in the introduction. See Wood & Eagly (2012) for a review, as well as Fernández et al. (2004) and (Oh 2019) for outstanding examples relating to gender and caste.

(ii) *Stereotypes.* Based on Bordalo, Coffman, Gennaioli & Shleifer (2016, 2019), greater representation in an economic activity can produce stereotypes which exaggerate a group's advantage in that activity. These stereotypes can create implicit bias and stereotype threat. There is evidence that stereotypes affect representation across scientific fields (Miller et al. 2015, Carli et al. 2016). As such, exposure to female role models can reduce gender gaps (Porter & Serra 2020), especially when role models violate personality stereotypes (Cheryan et al. 2011).

We account for these motivations by assuming an individual's identity-based payoff from participation is increasing in her group's representation. Specifically, the identity-based payoff at time t is αR^t for group A members and $\alpha(1 - R^t)$ for group B members, where $\alpha > 0$ is the (common) level of group identification, i.e. how much an individual cares about her group's representation relative to standard economic incentives.[8] This is consistent with both internalized and socially enforced identity-specific norms and stereotypes that depend on the group's representation. We assume that an individual's identity-based payoff from non-participation is $\alpha > 0$.

To summarize, the total payoffs to not participating ($e = 0$) and to participating ($e = 1$) are:

$$\begin{aligned} \text{not participating } (e = 0) &: \alpha \\ \text{participating } (e = 1, \text{group A}) &: y + \alpha R^t \\ \text{participating } (e = 1, \text{group B}) &: y + \alpha(1 - R^t). \end{aligned} \tag{2}$$

[7] This could also be thought of as a form of endogenous discrimination. Alternative theories of how identity-specific norms are determined include Shayo (2009), Akerlof (2017), Carvalho (2013), and Snower & Bosworth (2016). See Kranton (2016) for a review.

[8] The results here do not depend qualitatively on the linearity of this payoff. It only needs to be strictly increasing in the representation of one's group.

Following participation choice, each individual receives the corresponding payoff and is replaced by a successor with the same identity.

2.1 Participation & Representation

At time t, a group A member participates $e = 1$ if the total payoff difference is positive, that is:

$$y + \alpha R^t > \alpha \quad \Longleftrightarrow \quad y - \alpha(1 - R^t) > 0. \tag{3}$$

Hence the expected proportion of group A members who participate in period t (defined by the expectation with respect to F) is

$$p_A^t = 1 - F\big(\alpha(1 - R^t)\big). \tag{4}$$

We refer to this as group A's *participation rate.*[9]
A group B member participates $e = 1$ if:

$$y + \alpha(1 - R^t) > \alpha \quad \Longleftrightarrow \quad y - \alpha R^t > 0. \tag{5}$$

Hence, group B's participation rate in period t is

$$p_B^t = 1 - F\big(\alpha R^t\big). \tag{6}$$

To focus on interior solutions, we assume throughout that F is strictly increasing on $\big[0, \alpha\big]$ and $F(\alpha) < 1$. Note that the participation rate for each group is increasing in that group's representation: p_A^t is increasing in R^t and p_B^t is increasing in $1 - R^t$.

2.2 Dynamics & Equilibrium

Let us turn to the dynamics of representation starting from arbitrary initial conditions. Rather than working with the exact stochastic process from now on we study the expected (with respect to the distribution F) motion of R^t which we denote by r^t. The initial condition is $r^1 = R^1$. Thenceforth, r is determined by the following recurrence relation:

$$r^{t+1} = \frac{m_A\big[1 - F\big(\alpha(1 - r^t)\big)\big]}{m_A\big[1 - F\big(\alpha(1 - r^t)\big)\big] + m_B\big[1 - F\big(\alpha r^t\big)\big]} \equiv G\big(r^t\big). \tag{7}$$

[9] That individuals do not compare their group's representation to its base population share is consistent with the representativeness heuristic (Kahneman & Tversky 1972). Otherwise, individuals from a group with two percent population share would feel equally included in a profession in which their representation is 2 percent as a majority group with 98 percent representation. All our results apply when individuals partially account for their group's base population share.

By the law of large numbers, the stochastic dynamic converges to its deterministic approximation r^t as the population size becomes large. The law of large numbers applies since the process has finite expectation ($R^t \in [0,1]$).

An absorbing state r^* is a fixed point of G. We refer to this as an equilibrium (i.e. a steady state). $G : [0,1] \to [0,1]$ is continuous, so there exists at least one fixed point by Brouwer's fixed point theorem. Since $G(r)$ is strictly between zero and one, every fixed point is interior: $r^* \in (0,1)$. Finally, as G is strictly increasing and continuous, the dynamic process has nice convergence properties:

Proposition 1 *The process r^t converges to an equilibrium from any initial state r^1. Every equilibrium is interior, $r^* \in (0,1)$.*

The proof of this and all other results is in the Appendix.

Before proceeding, we note that if r^t were to be redefined as group A's share in the participating subpopulation in period t (rather than $t-1$), then there would no longer be intertemporal externalities from participation and the setting would reduce to a one-shot game. The pure-strategy Nash equilibria of this one-shot game are approximated by the fixed points of G, i.e., the equilibria of the dynamic process. Hence the steady state analysis below applies to the Nash equilibria of the one-shot game under the alternative definition of r^t.

2.3 Gender: Path-dependence

Let us define terms. Group A is less represented than group B if $r^t < \frac{1}{2}$. Group A is underrepresented (overrepresented) if $r^t < m_A$ ($r^t > m_A$). The definitions are symmetric for group B. We shall say that state r^t is representative if $r^t = m_A$.

We first consider the case of gender and assume equally sized groups: $m_A = m_B = \frac{1}{2}$. In this case, underrepresentation and inequality coincide. This is the case examined by Athey et al. (2000). Despite differences in the setup of our model, we attribute the type of path dependence we study here to their paper. Once we introduce majority bias in the next subsection, we depart substantially from their work.

Comparing Equations (4) and (6), the group with a higher historical participation rate has an immediate advantage. Now we see that identity-based effects lock in inequality over finite horizons.

Proposition 2 *Suppose the groups are of equal size,* $m_A = m_B = \frac{1}{2}$*. If* $r^T \overset{>}{\underset{<}{=}} \frac{1}{2}$ *at some time* T*, then* $r^t \overset{>}{\underset{<}{=}} \frac{1}{2}$ *for all finite* $t \geq T$*.*

Does inequality dissipate over time? If there is a unique equilibrium r^*, then the process is ergodic: the long-run mix r^* is independent of the initial condition r^1. If, however, there are multiple equilibria, then the process is path dependent. Whether persistent inequality can arise from path dependence depends on the precise distribution of economic returns F.

Consider the following example with normally distributed returns and equal group sizes.

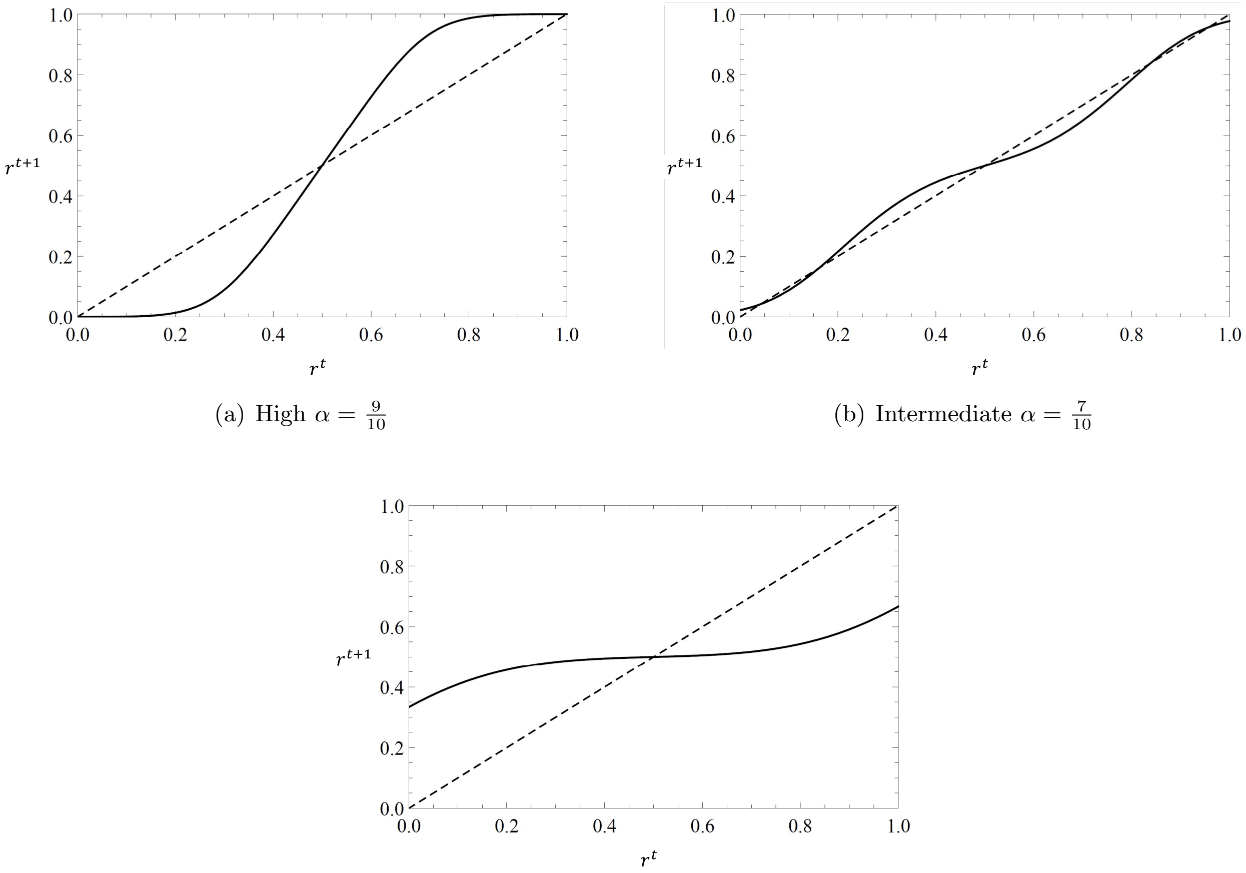
Example 1. $Y \sim N(0.5, 0.1)$, $m_A = \frac{1}{2}$.

Figure 2 plots the function G in this case. As one can see, there are multiple stable fixed points in Figures 2(a)-(b). Hence the positive feedback via identity representation can create path-dependent underrepresentation in this case. If group A starts out being overrepresented it remains so forever.

Path dependence emerges from the interaction of standard economic returns and group identification. Recall that the two groups are *ex ante* identical: equal size, same distribution of economic returns and same level of group identification α. It is also evident from Equation (7) that the equilibria are determined by the strength of group identification α. From Figure 2(a), when α is large, there are three fixed points, one equal $r^* = \frac{1}{2}$ and two (highly) unequal. The first and third (unequal) fixed points are asymptotically stable ($|G'(r^*)| < 1$), while the second (equal) fixed point is unstable ($|G'(r^*)| > 1$). Thus, path dependence will lead to *extreme* long-term underrepresentation of one group or another from almost every state, despite the groups having identical productivity.

Now observe Figure 2(b). When α is intermediate, there are five fixed points. The first, third and fifth are asymptotically stable. Hence there is an open set of initial conditions leading to equal representation. Of course, there is also an open set of initial conditions leading to persistent underrepresentation.

Finally, consider Figure 2(c). When α is small, the unique fixed point is equal representation, $r^* = \frac{1}{2}$. Hence lower group identification eliminates the path-dependence in inequality.



(a) High $\alpha = \frac{9}{10}$

(b) Intermediate $\alpha = \frac{7}{10}$

(c) Low $\alpha = 1/2$

Figure 2: Fixed points for Example 1, varying the strength of group identity α. Fixed points are strictly between 0 and 1.

In sum, identity-specific norms which depend on representation can generate path dependence which locks in initial inequalities even without current differences in ability, resources, initial human capital or discrimination. Thus historical discrimination and other negative shocks to a group can lead to permanent underperformance. For some time, representation of the group that suffered historical discrimination improves. But eventually the process can get stuck and never come close to equal representation. Figure 2 suggests that low group identification α mitigates such path dependence. Hence there is a tension between raising group identification to eliminate structural inequality through collective action and the exacerbation of the intergroup inequality that it targets.

What are the implications for gender inequality?

Though many factors are at work, these results are consistent with changes in women's labor force participation (Bureau of Labor Statistics 2018). In 1950, 33.9% of U.S. women 16 or older worked for pay compared to around 54.6% in 2017. The increase over this period was monotonic. As Goldin (2006) describes, rising participation was driven not only by economic factors but changes in identity, including the rising importance women attached to recognition at work. Also, consistent with our theory, family became more important for men during this period and their labor force participation rate declined from 86.4% in 1950 to 66% in 2017. Hence exogenous changes, such as wartime work and social movements (Fernández et al. 2004), could have initiated a positive feedback process between rising women's representation in the labor force and the breakdown of traditional gender norms. We have shown, however, that *this does not necessarily lead to equal representation.* Instead the process can get stuck in an unequal state. In fact, women's labor force participation in the U.S. has stagnated below 60% since the 1990s. In addition, while women's representation in formerly male-dominated fields has risen dramatically since the 1960s (Goldin 2006), women still comprise less than 7% of Fortune 500 CEOs and 14% of U.S.engineers (Jones 2017, U.S. Census Bureau 2016).

2.4 Race: Majority Bias

When it comes to racial/ethnic minorities, there is an additional factor contributing to underrepresentation, which we call majority bias.[10]

Let the majority group be A, so that $m_A > \frac{1}{2}$. We rely on methods for comparing (possibly many) equilibria (Milgrom & Roberts 1994). Let $\underline{r}^*$ and $\overline{r}^*$ be the smallest and largest fixed points of G respectively. Since $G(0) > 0$ and $G(1) < 1$, $\underline{r}^*$ and $\overline{r}^*$ are always asymptotically stable. We find that concerns over identity representation bias long-run participation rates in favor of the majority group.

Proposition 3 *The equilibrium structure depends on the majority share m_A:*

(i) *$\underline{r}^*$ and $\overline{r}^*$ are strictly increasing in m_A.*

(ii) *If $m_A > m_B$, there exists an equilibrium $r^* > \frac{1}{2}$. For all such equilibria, $r^* > m_A$.*

(iii) *There exists a representative equilibrium $r^* = m_A$ if and only if $m_A = \frac{1}{2}$.*

[10] Bruner (2017) and O'Connor & Bruner (2017) study majority bias in evolutionary bargaining games.

(iv) *For m_A sufficiently large, there exists a finite time T such that $r^t > m_A$ for all $t > T$, regardless of the initial condition r^1.*

(v) *For $m_A > m_B$ and $r^1 \geq \frac{1}{2}$, $r^t > m_A$ for all $t \geq 2$.*

According to Proposition 3, as group A's population share grows, equilibrium participation rates increasingly favor group A.[11] When groups are equal in size, there exists an equilibrium in which their participation rates are equalized and both groups have equal representation: $r^* = \frac{1}{2}$. As seen in Figure 2, this equilibrium may not be stable or unique. When group A is a strict majority, there exists no representative equilibrium $r^* = m_A$. There does, however, exist an equilibrium $r^* > m_A$. *That is, not only does the majority group have greater representation, but it is overrepresented.* To understand the identity-based multiplier, suppose that $r^t = m_A$. Because $m_A > \frac{1}{2}$, identity-based concerns mean group A has a higher participation rate in period t than group B. Therefore, $r^{t+1} > r^t = m_A$. Iterating this reasoning, we have $r^T > m_A$ for all $T > t$. Finally, if the majority group's share m_A is sufficiently large, then it is overrepresented in the long run regardless of initial conditions. That is, a large enough majority bias renders the process ergodic and eliminates path dependence.

To illustrate the interaction between majority bias and path dependence, examine Figure 3. This is the same case as Figure 2(b) except that group A's share of the population is 0.7, not 0.5. In Figure 2(b) there were five equilibria with the equal equilibrium being among the three asymptotically stable ones. When $m_A = 0.7$, there is a unique equilibrium and in this equilibrium group A is overrepresented: $r^* > 0.95$. Hence equality of opportunity leads to extreme underrepresentation of the racial/ethnic minority. We can show that the same forces that lead to underrepresentation of minorities under homogeneoity can lead to minority overrepresentation, and even dominance ($r^* < \frac{1}{2}$), when the minority has lower identification (α) and/or is more productive.

Thus majority bias makes the problem of racial/ethnic underrepresentation more severe than the problem of gender underrepresentation. This is one possible explanation for why economic underrepresentation has been reduced far more for women than minorities subject to historical discrimination. For example, of the doctoral degrees in economics awarded by U.S. institutions to US citizens and permanent residents in 2014, 31.4% were awarded to women, compared to 8.4% in total for blacks, Hispanics and native Americans who comprise around 30% of the U.S. population (Bayer & Rouse 2016).

[11] A similar result to Proposition 3(i) is produced by Muller-Itten & Öry (2019).

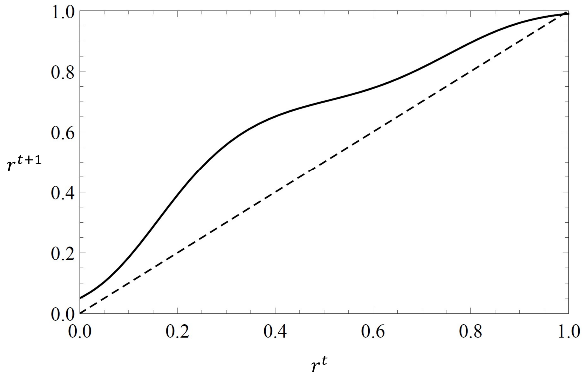


Figure 3: Group A is a strict majority, $m_A = 0.7$, and group identification is intermediate, $\alpha = \frac{7}{10}$. All else is as in Example 2.

Majority bias also has implications for policy. Suppose a policymaker imposes a quota $r = m_A$, so that each group is represented according to its population share? Can the temporary imposition of such a quota permanently eliminate underrepresentation? By Proposition 3(iii), the answer is yes if and only if $m_A = \frac{1}{2}$. Hence gender inequality is more easily mitigated through temporary interventions. Many proponents of affirmative action in the United States in the 1960s and 1970s saw it as a temporary measure. In 1978, in *Regents of the University of California vs. Bakke*, Supreme Court justice Harry Blackmun who supported the legality of affirmative action stated that “I yield to no one in my earnest hope that the time will come when an affirmative action program is unnecessary and [...] a relic of the past” and hoped that “within a decade at most, American society must and will reach a stage of maturity where acting along this line is no longer necessary” (quoted in Fang & Moro 2011, p. 164). Our results suggest that if a representative state ($r = m_A$) is the objective, ongoing subsidization of a racial/ethnic minority is necessary.

3 Multi-Dimensional Identity

Now that we understand how the representation dynamic produces inequality and underrepresentation in one dimension, we turn to our main analysis of multi-dimensional identity. We first present analogs of the uni-dimensional results on path dependence and majority bias, before examining the interaction between identity dimensions, especially race and gender.

The analysis focuses on two-dimensional identities, but our framework is suited to any finite number of dimensions. Let A, B denote the two identity groups along the first dimension and a, b

denote the two identity groups along the second. The population N is thus split into four different (intersectional) groups $N_{Aa}, N_{Ab}, N_{Ba}, N_{Bb}$. Again, the share of group $k \in \{Aa, Ab, Ba, Bb\}$ in the population is $m_k > 0$.

The representation of group N_k, $k \in \{Aa, Ab, Ba, Bb\}$, is

$$R_k^t = \frac{\sum_{i \in N_k} e_i^{t-1}}{\sum_{i \in N} e_i^{t-1}}. \tag{8}$$

Let $R^t = (R_{Aa}^t, R_{Ab}^t, R_{Ba}^t, R_{Bb}^t)$. Also define representation along the two identity dimensions as $R_A^t = R_{Aa}^t + R_{Ab}^t$, $R_B^t = 1 - R_A^t$ and $R_a^t = R_{Aa}^t + R_{Ba}^t$, $R_b^t = 1 - R_a^t$.

How does an individual evaluate their representation when identity is multi-dimensional?

Do they (i) care only about the representation of their intersectional group (R_k) or (ii) do they also care about their representation along each identity dimension (R_A and R_a)? This is an important empirical question, and we encourage work on it. We can, in this paper, say something about the consequences of each possibility.

An example of case (i) is when a black woman cares only about the representation of black women, and not the representation of black men or white women (with whom they match along one identity dimension). This is an extreme interpretation of the claim that an individual's experience is not the sum of their race and gender (Crenshaw 1989). Case (i) produces less interesting interaction between race and gender. In the 2 × 2 case we analyze, there would be four independent (intersectional) groups which is equivalent to the analysis in Section 2 with four uni-dimensional identities.

The interesting connections between race and gender emerge in case (ii), when individuals do not solely care about their intersectional group. This is the case we explore. The possibility that a black woman, for example, also cares about the representation of black men and white women, does not necessarily mean that an individual's experience is precisely the sum of their race and gender, except in the additively separable case. Even in this case, however, we shall see that interesting interactions between race and gender emerge.

The economic incentives for participation are as before. But now identity representation concerns are defined over both dimensions of identity. Let $X \in \{A, B\}$ and $x \in \{a, b\}$. The identity-based payoff from participation is then

$$H\left(R_X^t, R_x^t\right), \tag{9}$$

where H is a function aggregating an individual's representation concerns across the two dimensions of their identity.

To avoid manufacturing an interaction between identity dimensions, we assume H is additively separable with positive weights α and β:

$$\alpha R_X^t + \beta R_x^t. \tag{10}$$

In this case, any relationship between race and gender is emergent.[12]

To summarize, the total payoff to not participating ($e = 0$) and to participating ($e = 1$) for a member of group Xx, where $X \in \{A, B\}$ and $x \in \{a, b\}$, is

$$\begin{aligned} \text{not participating } (e = 0): &\quad \alpha + \beta \\ \text{participating} (e = 1): &\quad y + \alpha R_X^t + \beta R_x^t. \end{aligned} \tag{11}$$

3.1 Participation & representation

At time t a group Xx member participates $e = 1$ if

$$y + \alpha R_X^t + \beta R_x^t > \alpha + \beta \quad \iff \quad y > \alpha(1 - R_X^t) + \beta(1 - R_x^t). \tag{12}$$

Hence the participation rate of group Xx members in period t is:

$$p_{Xx}^t = 1 - F\big(\alpha(1 - R_X^t) + \beta(1 - R_x^t)\big). \tag{13}$$

To focus on interior solutions, we assume F is strictly increasing on $[0, \alpha + \beta]$ and $F(\alpha + \beta) < 1$.

3.2 Dynamics & equilibrium

As before we will study the expected (with respect to the distribution F) motion of R_k^t given by the following recurrence relation for $k \in \{Aa, Ab, Ba, Bb\}$:

$$r_k^{t+1} = \frac{m_k p_k^t}{\sum_{g \in \{Aa, Ab, Ba, Bb\}} m_g p_g^t} \equiv G_k\big(r^t\big). \tag{14}$$

Define $r^t = (r_{Aa}^t, r_{Ab}^t, r_{Ba}^t, r_{Bb}^t)$, as well as $r_A^t = r_{Aa}^t + r_{Ab}^t$ and $r_a^t = r_{Aa}^t + r_{Ba}^t$. Also let $G = (G_{Aa}(\cdot), G_{Ab}(\cdot), G_{Ba}(\cdot), G_{Bb}(\cdot))$.

Note that the same observations as in Section 2.2 apply and in particular the analog of the convergence result in Proposition 1 holds.

[12] One potentially fruitful way of extending our work is to explore the effect of complementarity and substitutability across dimensions.

3.3 Path Dependence

To focus on path dependence, consider the symmetric case of equal group sizes.

Proposition 4 *Suppose $m_{Aa} = m_{Ab} = m_{Ba} = m_{Bb} = \frac{1}{4}$. For arbitrary α, if $r_A^T \overset{>}{\underset{<}{=}} \frac{1}{2}$ at some time T, then $r_A^t \overset{>}{\underset{<}{=}} \frac{1}{2}$ for all finite $t \geq T$. The same applies to r_a^t.*

Thus path dependence produces inequality irrespective of the behavior along the other dimension, the strength of group identity, and the distribution F.[13]

3.4 Majority Bias

Now consider the effect of unequal group sizes. Population shares across each dimension are $m_A = m_{Aa} + m_{Ab}$ and $m_a = m_{Aa} + m_{Ba}$. Let $\underline{r}_k^*$ and $\overline{r}_k^*$ be the smallest and largest fixed points of G in the respective element G_k for $k \in \{Aa, Ab, Ba, Bb, A, B, a, b\}$, where $G_A = r_{Aa} + r_{Ab}$, $G_a = r_{Aa} + r_{Ba}$, and so forth. Since $G_k(0) > 0$ and $G_k(1) < 1$, $\underline{r}_k^*$ and $\overline{r}_k^*$ are asymptotically stable for all k. Again, we find that concerns for identity representation bias long-run participation rates in favor of the majority group.

Proposition 5 *Independent of α and β:*

(i) *Suppose the ratio $\frac{m_{Aa}}{m_{Ab}}$ remains constant. $\underline{r}_A^*$ and $\overline{r}_A^*$ are strictly increasing in m_A.*

(ii) *If $m_A > m_a \geq \frac{1}{2}$ and $\alpha \geq \beta$, then there exists an equilibrium in which $r_A^* > m_A$.*

(iii) *There exists a representative equilibrium, $r_k^* = m_k$ for each $k \in \{Aa, Ab, Ba, Bb\}$, if and only if $m_A = m_a = \frac{1}{2}$.*

As in the uni-dimensional case, as group A's population share grows (holding all other proportions equal), equilibrium representation increasingly favors group A. The same applies to group a, along the other dimension. The other two results, however, diverge from the uni-dimensional case. Majority bias emerges only under certain conditions when identity is multi-dimensional. Suppose there is a strict majority along at least one identity dimension. Take the dimension with the largest majority and set the weight on this dimension no lower than the weight on the other dimension. For example, if $m_A > m_a \geq 1/2$, set $\alpha \geq \beta$. Then there exists an equilibrium in which the majority

[13] In addition, the results for path dependence and majority bias hold for any strictly increasing H function, not just the additively separable case.

is overrepresented along this dimension: $r_A^* > m_A$. Otherwise, there exist type distributions and population shares with $m_A > 1/2$ such that there is no equilibrium in which $r_A^* > m_A$, due to spillovers from the other identity dimension. In addition, there exists an equilibrium in which each intersectional group is represented according to its population share if and only if population shares are equal along each identity dimension ($m_A = m_a = \frac{1}{2}$). This is another connection between identity dimensions. Suppose the first dimension is gender. Even if women comprise half the population ($m_A = \frac{1}{2}$), there may be no equilibrium in which women are equally represented due to size effects along the other identity dimension. Therefore, unlike the uni-dimensional case, it is generally impossible to achieve equal gender representation with a temporary quota, due to connections between the identity dimensions.

3.5 Interaction between Race and Gender

Let us examine more closely the interaction between different dimensions of an individual's identity. The common reductionist approach, which treats outcomes along each dimension as independent, ignores such interactions. Interventions are designed as if they are 'sterilized', in the sense that policies targeting inequality and underrepresentation along one dimension have no affect on other dimensions. We find that such interventions are not possible. We focus here on a simple subsidy, a lump-sum payment of s to all individuals with trait B along the first dimension who choose to participate. We show that such a policy almost always changes outcomes along the second dimension. Similar results could be provided for other policies, including the quotas and role-model interventions we describe below.

Proposition 6 *Let* $\Omega = \{(m_{Aa}, m_{Ab}, m_{Ba}, m_{Bb}) \in [0,1]^4 : m_{Aa} + m_{AB} + m_{Ba} + m_{Bb} = 1\}$. *Given a state* r^t *and a subsidy* $s > 0$ *to* B *types along the first dimension, let* $\hat{\Omega} \subseteq \Omega$ *be the set of population shares for which the second dimension is unchanged by the subsidy:* $r_a^{t+1}(s) = r_a^{t+1}(0)$. *Then* $\hat{\Omega}$ *has measure zero.*

Hence sterilized interventions are generically impossible. This connection between economic outcomes along multiple identity dimensions adds to the body of examples provided by Saari (2015, 2018), ranging from voting to the calculation of dark matter, in which reductionism produces analytical errors. Not only is there non-trivial interaction between identity dimensions such as race and gender, but failing to account for these interactions can produce significant policy mistakes.

In fact, as we shall now show, piecemeal interventions which reduce underrepresentation along one identity dimension can increase underrepresentation along the other dimension.

For concreteness, consider an example based on the US population. Let the first dimension be gender (men and women) and the second be race (non-black and black). The population shares are based on the United States Census (2019):[14]

Example 2. $Y \sim N(0.5, 0.1)$, $(m_{Aa}, m_{Ab}, m_{Ba}, m_{Bb}) = (0.426, 0.066, 0.440, 0.068)$, and $(\alpha, \beta) = (0.7, 0.3)$.

The equilibrium with maximal representation of men (A types) is $r^* = (0.866, 0.133, 0.001, 0.000)$. While men are overrepresented ($r^*_A = 0.999 > m_A = 0.492$), the black and non-black populations are represented roughly according to their population shares ($r^*_a = 0.867 \approx m_a = 0.866$). Black representation, however, is almost entirely due to participation by black men ($r^*_{Ab} = 0.133$), who benefit from the representation of white men. Black women, who are doubly disadvantaged, hardly participate ($r^*_{Ab} = 0.000$).

Next, following Proposition 6, we introduce a subsidy for women (B types) of $s = 0.05$. This has important consequences. The equilibrium with maximal representation of men is now $r^*(s = 0.05) = (0.440, 0.003, 0.532, 0.025)$. While men are no longer overrepresented ($r^*_A(0.05) = 0.443 < m_A = 0.492$), the black population is now severely underrepresented ($r^*_b(0.05) = 0.028 < m_b = 0.134$), due to a decline in participation by black men ($r^*_{Ab}(0.05) = 0.003$). By leveling representation between men and women, black representation declines from 13.3% to 2.8% (with a population share of $m_b = 13.4\%$).

Hence eliminating underrepresentation of women, without accounting for the connection between race and gender, can increase underrepresentation along the race dimension. Of course, under different conditions, the effect could go the other way. That is, eliminating underrepresentation along one dimension could reduce underrepresentation along other dimensions. Identifying identity dimensions that offer such positive trickle-down effects resembles the concept of protecting *umbrella species* in ecology. Such species are selected for conservationist action as their protection leads indirectly to the protection of many other species in their habitat (Roberge & Angelstam 2004).

In sum, policies are required that consider all relevant dimensions of identity. One obvious candidate that we could analyze is identity-based quotas. Quotas are straightforward when there are a fixed

[14] Retrieved via `https://www.census.gov/quickfacts/fact/table/US/PST045219` on 20 August 2020.

number of slots to be filled, for example in a single firm. Ours is a larger-scale analysis of an education system or labor market, so the number of participating agents is endogenous. In this case, moving to a representative state through identity-based quotas requires restricting hiring from overrepresented groups, which is inefficient. Hence we examine how underrepresentation can be eliminated along both identity dimensions using two other policy instruments: (I) self-financing subsidies and (II) role models. Later on, we will examine the efficiency properties of these policies.

3.6 Policy I: Self-financing Subsidies

Underrepresented groups could be subsidized through public investments in their productivity, including educational subsidies. Let s_{Xx} be the lump-sum subsidy for each participating member of intersectional group Xx ($X \in \{A, B\}$ and $x \in \{a, b\}$). If $s_{Xx} < 0$, it is a lump-sum tax. Given the system of subsidies/taxes, the participation rate for group Xx is:

$$p_{Xx}(s_{Xx}) \quad = \quad 1 - F\left(\alpha(1 - r_X) + \beta(1 - r_x) - s_{Xx}\right). \tag{15}$$

To reach a representative state $r_{Xx} = m_{Xx}$ for all $X \in \{A, B\}$ and $x \in \{a, b\}$, participation rates must be equalized across intersectional groups. One way to achieve this is through a system of self-financing subsidies.

Proposition 7 *From any state r^t, a representative state $r^{t+1} = (m_{Aa}, m_{Ab}, m_{Ba}, m_{Bb})$ can be reached in the next period through self-financing subsidies/taxes, $s^t_{Aa}, s^t_{Ab}, s^t_{Ba}, s^t_{Bb}$. The subsidies are given by the following set of equations and are generically unique:*

$$\begin{aligned}
s^t_{Aa} &= -[(1 - m_A) \cdot \alpha \cdot (2r^t_A - 1) + (1 - m_a) \cdot \beta \cdot (2r^t_a - 1)] \\
s^t_{Ab} &= s^t_{Aa} + \beta \cdot (2r^t_a - 1) \\
s^t_{Ba} &= s^t_{Aa} + \alpha \cdot (2r^t_A - 1) \\
s^t_{Bb} &= s^t_{Aa} + \alpha \cdot (2r^t_A - 1) + \beta \cdot (2r^t_a - 1).
\end{aligned} \tag{16}$$

Proposition 7 characterizes how to reach a representative state and maintain it once there. The steady state is reached in two periods, with the steady-state subsidies given by replacing representation with the population shares in (16), e.g. by replacing r^t_A by m_A. The policy is easy to compute and, apart from the identification parameters (α, β), requires only readily available data on population shares (m_A, m_a) and representation in the prior period (r^t_A, r^t_a). In particular, it is independent of the distribution F. Notice that the greater the strength of group identity (α, β), the larger the subsidy to each underrepresented group. Hence, while strong identification exacerbates underrepresentation, it also creates incentives for collective action to reduce it.

Most importantly, the policy contrasts with the standard reductionist approach of treating underrepresentation independently along each dimension. Formally, the system (16) cannot be reduced to two independent equations, one for each dimension. If that were true, underrepresentation could be eliminated through a subsidy/tax of s_b along the race dimension and a separate subsidy/tax for women/men s_B. Then the subsidy to a black woman would be $s_{Bb} = s_B + s_b$. According to Proposition 7, underrepresentation cannot be eliminated in this way. The interaction between identity dimensions must be accounted for.

The system of subsidies/taxes (16) tells us precisely how to do so. The policy is holistic in that each group's subsidy/tax takes into account data (population shares, past representation) from both dimensions. The policy is also intersectional: each intersectional group Xx has a different subsidy/tax. Notice that this intersectional policy emerges from non-intersectional primitives. That is, by (11), each type's payoff is an additively separable function of representation along the two identity dimensions. Nevertheless, the subsidy to a black woman (e.g. s_{Bb}) is not equal to the sum of the subsidies to black men (s_{Ab}) and white women (s_{Ba}). Despite arising from different primitives, this result is consistent with the intersectional view of race and gender:

> "These [multidimensional] problems of exclusion cannot be solved simply by including Black women within an already established analytical structure. Because the intersectional experience is greater than the sum of racism and sexism, any study that does not take intersectionality into account cannot address the particular manner in which Black women are subordinated." (Crenshaw 1989, p. 140).

To further illustrate the connection between identity dimensions, suppose for the moment that the race dimension is not taken into account when calculating subsidies. This is equivalent to assuming that all individuals have the same race ($m_a = 1$). Then the subsidy to (white) women from (16) is $s_{Ba} = \alpha m_A(2r_A - 1)$. By inspection of Equations (16), this does not generally equal the subsidy s_{Ba} when there is also variation along the race dimension ($m_a < 1$). Consider the following numerical example:

Example 3. Let $m_A = 0.5$, $r_A = 0.7$, $\alpha = \beta = 1$. Suppose that the race dimension is not taken into account. Then the subsidies are $(s_{Aa}, s_{Ba}, s_{Ab}, s_{Bb}) = (-0.2, 0.2, -0.2, 0.2)$. Now suppose that both identity dimensions are taken into account and let, for example, $m_a = 0.7$, $r_a = 0.7$. Then $(s_{Aa}, s_{Ba}, s_{Ab}, s_{Bb}) = (-0.32, 0.08, 0.08, 0.48)$.

Consider the subsidy to white women. When only the gender dimension is considered, white women belong to the underrepresented group and thus receive a subsidy of $s_{Ba} = s_{Bb} = 0.2$. Now when the second identity dimension, race, is also taken into account it is black women who need to be subsidized the most to reach a representative state. They are underrepresented on the gender dimension and in the minority on the race dimension. Though white women are also underrepresented on the gender dimension, they are in the majority on the race dimension ($m_a = 0.7$). As such, the subsidy to white women falls to $s_{Ba} = 0.08$, while black women receive a subsidy of $s_{Bb} = 0.48$. Notice that this occurs even when there is no underrepresentation along the race dimension ($r_a = m_a = 0.7$). This is another (non-obvious) interaction between multiple identity dimensions.

3.7 Policy II: Role Models

Another instrument for unwinding identity-specific norms and stereotypes is the promotion of role models. For example, Porter & Serra (2020) show that the number of women majoring in economics rises significantly when introductory classes are exposed to "successful and charismatic" women who majored in economics at the same university. The American Economic Association conducts similar initiatives through the Committee on the Status of Women in the Economics Profession [CSWEP] and the Committee on the Status of Minority Groups in the Economics Profession [CSMGEP].

We model the effect of role models as follows.[15] We have so far assumed that an individual's identity-based payoff from participating in an economic activity is increasing in her representation. We suggest that role models can alter perceived representation and thereby eliminate actual underrepresentation. In particular, suppose that perceived representation is a convex combination of actual representation and role-model effects, with weight $\delta \in [0, 1]$ on role-model effects. Role models have identities $k \in \{Aa, Ab, Ba, Bb\}$. A policymaker can choose to promote certain role models to alter perceived representation and encourage participation by that group. Let $\gamma_k \in [0, 1]$ be the share of 'exposure' of role models from group k, with $\sum_{k \in \{Aa,Ab,Ba,Bb\}} \gamma_k = 1$. Define $\gamma = (\gamma_{Aa}, \gamma_{Ab}, \gamma_{Ba}, \gamma_{Bb})$. Then a role-model policy is a pair (δ, γ).

[15] In existing theoretical work, role models do not shape norms/stereotypes but rather play an informational role, helping successors learn about the net benefits to economic participation (e.g. Chung 2000).

Such a policy alters perceived representation $\hat{r}_X$ and $\hat{r}_x$ along each dimension as follows:

$$\begin{aligned} \hat{r}_X &= \delta(\gamma_{Xa} + \gamma_{Xb}) + (1-\delta) r_X & (17) \\ \hat{r}_x &= \delta(\gamma_{Ax} + \gamma_{Bx}) + (1-\delta) r_x. & (18) \end{aligned}$$

We can then redefine the identity-based payoff in (10) with perceived representation replacing actual representation. Thus identity-specific norms/stereotypes, and participation decisions, could be decoupled from actual representation.

Proposition 8 *From any state r^t, a representative state $r^{t+1} = (m_{Aa}, m_{Ab}, m_{Ba}, m_{Bb})$ can be reached in the next period through a role-model policy (δ, γ) that satisfies the following system of equations:*

$$\begin{aligned} \gamma_{Aa} + \gamma_{Ab} &= \tfrac{1}{2} - \tfrac{1-\delta}{\delta}\left(r_A - \tfrac{1}{2}\right) & (19) \\ \gamma_{Aa} + \gamma_{Ba} &= \tfrac{1}{2} - \tfrac{1-\delta}{\delta}\left(r_a - \tfrac{1}{2}\right). & (20) \end{aligned}$$

In particular, this implies:

$$\delta \geq \max\left\{\frac{r_A^t - \frac{1}{2}}{r_A^t}, \frac{r_a^t - \frac{1}{2}}{r_a^t}\right\}. \qquad (21)$$

First, the role-model policy can be computed using a sequential procedure and hence is simpler than the case of self-financing subsidies. For example, a common degree of exposure $\gamma_{Aa} = \gamma_{Ab} = \frac{1}{4} - \frac{1-\delta}{2\delta}\left(r_A - \frac{1}{2}\right)$ can be set along the first dimension without considering the second dimension. Then role-model exposure can be set appropriately along the second dimension accounting for the exposure chosen along the first dimension. Hence the role-model strategy need not be exactly intersectional. It must, however, be holistic in the sense that interventions along the two identity dimensions are connected.

Second, by (19)-(20), when the role-model effect δ is low (due for example to low total investment in the policy), role models from underrepresented groups will have to be promoted more heavily to overcome their disadvantage from low actual representation. Inequality (21) can be interpreted as requiring investment in the role-model policy to be sufficiently large. If δ is too low (such that (21) is violated), the exposure of the overrepresented groups may have to be negative to reach a representative state. One can think of this as stigmatizing the overrepresented groups, which we have made infeasible.

Third, the role-model policy requires even less information than self-financing subsidies. The policy is independent of the population shares m_k, the distribution F, and the strength of group identity (α, β). All a policymaker needs to know is the relevant identity dimensions and the prior period's representation along each dimension.

The role-model approach is not, however, unambiguously better. By (19), in a representative state $r_k = m_k$:

$$\gamma_{Ba} + \gamma_{Bb} = \tfrac{1}{2} + \tfrac{1-\delta}{\delta} \left(m_A - \tfrac{1}{2} \right). \tag{22}$$

Hence to equalize perceived representation, very small minorities (e.g. m_A close to zero) need a huge boost in exposure. As a consequence, it turns out that self-financing subsidies produce higher economic output. This is true even when ignoring the direct costs of a role-model policy (of setting $\delta > 0$).

Let y_{Xx} be the threshold such that an Xx type participates ($e = 1$) if $y \geq y_{Xx}$. Then we can define economic output as

$$Y = \sum_{Xx \in \{Aa, Ab, Ba, Bb\}} m_{Xx} \int_{y_{Xx}}^{\infty} y \, dF(y). \tag{23}$$

Proposition 9 *For $m_A \neq \frac{1}{2}$ and/or $m_a \neq \frac{1}{2}$, a representative state stabilized through self-financing subsidies yields strictly higher economic output than a representative state stabilized through role models. (Otherwise, they yield the same output.)*

The reason is that, in the case of self-financing subsidies, the cost of subsidizing a minority is divided over a large number of majority group members. With role models, the full effect of a large boost in exposure for a minority is felt by each majority group member, and not divided among them. We have not seen this distinction made before.

3.8 Further Policy Considerations

The degree to which identity-based policies have been tried varies across national and subnational units, including U.S. states. India, for example, operates a large-scale system of affirmative action based on caste (Bagde et al. 2016). In some places, such identity-based policies are prohibited by law. They also violate the liberal principle of general (identity-free) rules (Johnson & Koyama

2019). An alternative approach to eliminating underrepresentation would be to reduce group identification (α, β) to zero. This would eliminate the rivalry produced by concerns about representation. However, it could be that group identity is a stable part of human preferences, at least in certain contexts (Tajfel & Turner 1986, Chen & Li 2009). In addition, when a group faces discrimination, some level of group identification may be necessary for collective action (Hwang et al. 2016). Our analysis uncovers additional considerations. Identity-based policymaking is made complex by the connections between race, gender and other characteristics. Such policies can also come at a cost in terms of economic output. These insights are summarized as follows:

1. *Knowledge.* In a larger world than the 2 × 2 identity configuration we analyze, many other identity dimensions could be relevant (e.g. class, age, sexual orientation). A policymaker requires perfect knowledge of which dimensions are important to individuals and how strongly they identify along each dimension. Otherwise, the same problem described in Proposition 6 can arise. That is, interventions along some identity dimensions can increase underrepresentation along other (neglected) dimensions.

2. *Conflict.* The fact that representation is a rival good means the structure of identity might be politically contested. In particular, (i) the set of salient identity dimensions and (ii) the categorization scheme within each dimension are subject to social/political construction. It is not clear whether there is a stable set of identities, as there will always be an incentive to lobby for recognition of a new identity category. In this domain, political conflict might be complex and costly, because the introduction of a new identity category affects the entire system (see Example 3), as does political distortion along an existing identity dimension. In addition, the number of intersectional identities grows rapidly in the number of dimensions, making intersectional policies cumbersome.

3. *Efficiency.* We have seen that self-financing subsidies are more efficient than role models. Nevertheless, subsidies can reduce economic output compared to non-intervention, under certain conditions. Consider the following example:

Proposition 10 *Suppose $m_A > \frac{1}{2}$, $m_a = \frac{1}{2}$ and $\alpha = \beta$. If $y \cdot f(y)$ is decreasing (thus including power law distributions), a state r such that $r_A > m_A$ and $r_a = m_a$ yields higher economic output than $r'_A = m_A$, $r'_a = m_a$ with self-financing subsidies.*

Hence, for example, if an equilibrium is already representative along the gender dimension, then eliminating underrepresentation along the race dimension can reduce output. This is not obvious. A representative policy raises participation by B types and lowers participation by A types. If the density of economic returns f is decreasing, two countervailing forces are at play. On the one hand, the marginal member of the minority B is more productive (higher y) than the marginal member of the majority A. On the other hand, more members of the majority are marginal. For a number of distributions, including power law distributions, the second effect dominates. That does not mean that interventions should be abandoned. The correct policy is a matter of distributive justice and a host of other considerations beyond the scope of this study.

4 Conclusion

In this paper, we introduce a model of economic participation and representation with multi-dimensional identity. This allows us to examine the interaction between race and gender, something that is usually neglected. In the uni-dimensional case, we show that temporary quotas can eliminate underrepresentation of women, but not racial minorities. In the multi-dimensional case, permanent interventions are needed even along the gender dimension, due to the interaction between race and gender. The consequences of neglecting this connection can be severe. Piecemeal interventions that reduce underrepresentation along one dimension can increase underrepresentation along another. Hence we encourage economists to depart from the standard reductionist approach and devote more attention to the interaction between identity dimensions, both in analysis and policymaking. We show how a system of self-financing subsidies or role models can eliminate underrepresentation along every dimension. In the subsidy case, the policy must be intersectional, but not necessarily in the role model case. Our modeling framework is simple enough to permit extension in a number of directions, including (i) asymmetries in productivity, discrimination and identification, (ii) the introduction of qualitatively different dimensions such as class, and (iii) identity choice, including choice of gender identity and the possibility of racial/ethnic ‘passing’. While the multi-dimensional nature of identity has long been recognized in other disciplines, we hope it becomes a central part of work on identity in economics.

References

Akerlof, G. A. & Kranton, R. E. (2000), ‘Economics and identity’, *Quarterly Journal of Economics* **115**(3), 715–753.

Akerlof, G. A. & Kranton, R. E. (2010), ‘Identity economics: How identities shape our work, wages, and well-being’, *Princeton, NJ: Princeton University Press* .

Akerlof, R. (2017), ‘Value formation: The role of esteem’, *Games and Economic Behavior* **102**, 1–19.

Altonji, J. G. & Blank, R. M. (1999), 'Race and gender in the labor market', *Handbook of Labor Economics* **3**, 3143–3259.

Arrow, K. J. (1973), The theory of discrimination, *in* 'Discrimination in Labor Markets', Princeton University Press, pp. 3–33.

Athey, S., Avery, C. & Zemsky, P. (2000), 'Mentoring and diversity', *American Economic Review* **90**(4), 765–786.

Bagde, S., Epple, D. & Taylor, L. (2016), 'Does affirmative action work? Caste, gender, college quality, and academic success in India', *American Economic Review* **106**(6), 1495–1521.

Bayer, A. & Rouse, C. E. (2016), 'Diversity in the economics profession: A new attack on an old problem', *Journal of Economic Perspectives* **30**(4), 221–42.

Becker, G. S. & Tomes, N. (1979), 'An equilibrium theory of the distribution of income and intergenerational mobility', *Journal of political Economy* **87**(6), 1153–1189.

Benabou, R. (1993), 'Workings of a city: Location, education, and production', *Quarterly Journal of Economics* **108**(3), 619–652.

Bénabou, R. & Tirole, J. (2011), 'Identity, morals, and taboos: Beliefs as assets', *Quarterly Journal of Economics* **126**(2), 805–855.

Bertrand, M. (2011), New perspectives on gender, *in* 'Handbook of Labor Economics', Vol. 4b, pp. 1543–1590.

Bertrand, M. (2020), Gender in the twenty-first century, *in* 'AEA Papers and Proceedings', Vol. 110, pp. 1–24.

Bertrand, M., Kamenica, E. & Pan, J. (2015), 'Gender identity and relative income within households', *Quarterly Journal of Economics* **130**(2), 571–614.

Bordalo, P., Coffman, K., Gennaioli, N. & Shleifer, A. (2016), 'Stereotypes', *Quarterly Journal of Economics* **131**(4), 1753–1794.

Bordalo, P., Coffman, K., Gennaioli, N. & Shleifer, A. (2019), 'Beliefs about gender', *American Economic Review* **109**(3), 739–73.

Borjas, G. J. (1992), 'Ethnic capital and intergenerational mobility', *Quarterly Journal of Economics* **107**(1), 123–50.

Brewer, R. M., Conrad, C. A. & King, M. C. (2002), 'The complexities and potential of theorizing gender, caste, race, and class', *Feminist Economics* **8**(2), 3–17.

Bruner, J. P. (2017), 'Minority (dis) advantage in population games', *Synthese* pp. 1–15.

Bureau of Labor Statistics (2018), 'Labor force statistics from the current population survey', United States Department of Labor, Washington DC.

Carli, L. L., Alawa, L., Lee, Y., Zhao, B. & Kim, E. (2016), 'Stereotypes about gender and science: Women$\neq$ scientists', *Psychology of Women Quarterly* **40**(2), 244–260.

Carvalho, J.-P. (2013), 'Veiling', *Quarterly Journal of Economics* **128**(1), 337–370.

Chaudhuri, S. & Sethi, R. (2008), 'Statistical discrimination with peer effects: Can integration eliminate negative stereotypes?', *Review of Economic Studies* **75**(2), 579–596.

Chen, Y. & Li, S. X. (2009), 'Group identity and social preferences', *American Economic Review* **99**(1), 431–57.

Cheryan, S., Siy, J. O., Vichayapai, M., Drury, B. J. & Kim, S. (2011), 'Do female and male role models who embody STEM stereotypes hinder women's anticipated success in stem?', *Social Psychological and Personality Science* **2**(6), 656–664.

Cheryan, S., Ziegler, S. A., Montoya, A. K. & Jiang, L. (2017), 'Why are some stem fields more gender balanced than others?', *Psychological Bulletin* **143**(1), 1.

Chung, K.-S. (2000), 'Role models and arguments for affirmative action', *American Economic Review* **90**(3), 640–648.

Coate, S. & Loury, G. C. (1993), 'Will affirmative-action policies eliminate negative stereotypes?', *American Economic Review* **83**(5), 1220–40.

Collins, P. H. & Bilge, S. (2016), *Intersectionality*, John Wiley & Sons.

Cooper, B. (2016), Intersectionality, *in* L. Disch & M. Hawkesworth, eds, 'The Oxford Handbook of Feminist Theory', Oxford University Press.

Crenshaw, K. (1989), 'Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics', *University of Chicago Legal Forum* pp. 139–167.

Croson, R. & Gneezy, U. (2009), 'Gender differences in preferences', *Journal of Economic Literature* **47**(2), 448–74.

Daly, M. C., Hobijn, B. & Pedtke, J. H. (2017), 'Disappointing Facts about the Black-White Wage Gap', Economic Letter, Federal Reserve Bank of San Francisco.

Elu, J. U. & Loubert, L. (2013), 'Earnings inequality and the intersectionality of gender and ethnicity in Sub-Saharan Africa: The case of Tanzanian manufacturing', *American Economic Review* **103**(3), 289–92.

Fang, H. & Moro, A. (2011), Theories of statistical discrimination and affirmative action: A survey, *in* M. O. J. Jess Benhabib & A. Bisin, eds, 'Handbook of Social Economics', Elsevier, pp. 133–200.

Fernández, R. (2013), 'Cultural change as learning: The evolution of female labor force participation over a century', *American Economic Review* **103**(1), 472–500.

Fernández, R., Fogli, A. & Olivetti, C. (2004), 'Mothers and sons: Preference formation and female labor force dynamics', *Quarterly Journal of Economics* **119**(4), 1249–1299.

Gneezy, U., Niederle, M. & Rustichini, A. (2003), 'Performance in competitive environments: Gender differences', *Quarterly Journal of Economics* **118**(3), 1049–1074.

Goldin, C. (2006), 'The quiet revolution that transformed women's employment, education, and family', *American economic review* **96**(2), 1–21.

Greenfieldboyce, N. (2019), 'Academic science rethinks all-too-white 'dude walls' of honor'.
URL: *https://www.npr.org/sections/health-shots/2019/08/25/749886989/academic-science-rethinks-all-too-white-dude-walls-of-honor*

Hwang, S.-H., Naidu, S. & Bowles, S. (2016), 'Social conflict and the evolution of unequal conventions'.

Johnson, N. D. & Koyama, M. (2019), *Persecution & toleration: The long road to religious freedom*, Cambridge University Press.

Jones, S. (2017), 'White men account for 72% of corporate leadership at 16 of the fortune 500 companies'.
URL: *http://fortune.com/2017/06/09/white-men-senior-executives-fortune-500-companies-diversity-data/*

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

Kranton, R. E. (2016), 'Identity economics 2016: Where do social distinctions and norms come from?', *American Economic Review* **106**(5), 405–09.

Leslie, S.-J., Cimpian, A., Meyer, M. & Freeland, E. (2015), 'Expectations of brilliance underlie gender distributions across academic disciplines', *Science* **347**(6219), 262–265.

Loury, G. C. (1977), A dynamic theory of racial income differences, *in* P. Wallace & A. LaMond, eds, 'Women, Minorities, and Employment Discrimination', D.C Heathand Co, Lexington MA, pp. 86–153.

Loury, G. C. (1981), 'Intergenerational transfers and the distribution of earnings', *Econometrica* **49**(4), 843–867.

Milgrom, P. & Roberts, J. (1994), 'Comparing equilibria', *American Economic Review* pp. 441–459.

Miller, D. I., Eagly, A. H. & Linn, M. C. (2015), 'Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations', *Journal of Educational Psychology* **107**(3), 631.

Moro, A. & Norman, P. (2004), 'A general equilibrium model of statistical discrimination', *Journal of Economic Theory* **114**(1), 1–30.

Muller-Itten, M. & Öry, A. (2019), 'Mentoring and the dynamics of affirmative action'.

Niederle, M. & Vesterlund, L. (2008), 'Gender differences in competition', *Negotiation Journal* **24**(4), 447–463.

O'Connor, C., Bright, L. K. & Bruner, J. P. (2019), 'The emergence of intersectional disadvantage', *Social Epistemology* **33**(1), 23–41.

O'Connor, C. & Bruner, J. (2017), 'Dynamics and diversity in epistemic communities', *Erkenntnis* pp. 1–19.

Oh, S. (2019), 'Does identity affect labor supply?', *Job Marker Paper, Columbia University* .

Piketty, T. (2014), *Capital in the Twenty-first Century*, Harvard University Press, Cambridge, MA.

Piketty, T. & Saez, E. (2003), 'Income inequality in the United States, 1913–1998', *Quarterly Journal of*

Economics **118**(1), 1–41.

Porter, C. & Serra, D. (2020), 'Gender differences in the choice of major: The importance of female role models', *American Economic Journal: Applied Economics* **12**(3), 226–54.

Reuben, E., Sapienza, P. & Zingales, L. (2015), Taste for competition and the gender gap among young business professionals, Working Paper 21695, National Bureau of Economic Research.

Roberge, J.-M. & Angelstam, P. (2004), 'Usefulness of the umbrella species concept as a conservation tool', *Conservation biology* **18**(1), 76–85.

Saari, D. G. (2015), 'Social science puzzles: A systems analysis challenge', *Evolutionary and Institutional Economics Review* **12**(1), 123–139.

Saari, D. G. (2018), *Mathematics Motivated by the Social and Behavioral Sciences*, SIAM.

Scott, C. E. & Siegfried, J. J. (2019), 'American Economic Association universal academic questionnaire summary statistics', *AEA Papers and Proceedings* **109**, 590–592.

Sen, A. (2006), *Identity and Violence: The Illusion of Destiny*, W.W. Norton, New York, NY.

Shayo, M. (2009), 'A model of social identity with an application to political economy: Nation, class, and redistribution', *American Political Science Review* **103**(2), 147–174.

Skjeie, H. (2015), Gender equality and nondiscrimination: How to tackle multiple discrimination effectively, *in* F. Bettio & S. Sansonetti, eds, 'Visions for Gender Equality', European Union, Luxembourg, pp. 79–82.

Snower, D. J. & Bosworth, S. J. (2016), 'Identity-driven cooperation versus competition', *American Economic Review* **106**(5), 420–24.

Tajfel, H. & Turner, J. C. (1986), The social identity theory of intergroup behaviour, *in* S. Worchel & W. Austin, eds, 'The Psychology of Intergroup Relations', Nelson-Hall, Chicago, pp. 7–24.

U.S. Census Bureau (2016), 'American Community Survey'.

U.S. Department of Education (2017), 'Consolidated State Performance Report, 2015-16'.

Wood, W. & Eagly, A. H. (2012), Biosocial construction of sex differences and similarities in behavior, *in* 'Advances in Experimental Social Psychology', Vol. 46, Elsevier, pp. 55–123.

Young, H. P. (1998), *Individual Strategy and Social Structure*, Princeton University Press, Oxford, U.K.

Young, H. P. (2015), 'The evolution of social norms', *Annual Review of Economics* **7**(1), 359–387.

Appendix

Proof of Proposition 1. Consider the function G in Equation (7). Suppose $G(r^1) = r^2 \geq r^1$. Since G is strictly increasing, $r^3 \geq r^2$, $r^4 \geq r^3$, and so forth. That is, the entire sequence is increasing. As each element of the sequence is an image of G and G is bounded, the sequence is also bounded and must therefore converge. Because G is continuous it converges to a fixed point $G(r^*) = r^*$.

Suppose instead that $G(r^1) = r^2 \leq r^1$, then the entire sequence is decreasing and again we have convergence to a fixed point. □

Proof of Proposition 2. Suppose $r^T > \frac{1}{2}$ for some $T \geq 1$. By hypothesis, $m_A = m_B = \frac{1}{2}$. As G is strictly increasing in Equation (7), $r^{T+1} = G(r^T) > G(\frac{1}{2}) = \frac{1}{2}$. Iterating, we have $r^t > \frac{1}{2}$ for all $t > T$. The remainder of the results follow immediately. □

Proof of Proposition 3. (i) Recall $m_B = 1 - m_A$. Hence $G(r; m_A)$ in Equation (7) is strictly increasing in m_A. The result then follows from Milgrom & Roberts (1994, Theorem 1).

(ii) Recall that G is continuous, $G\big(\frac{1}{2}\big) = m_A > \frac{1}{2}$ by Equation (24) and $G\big(1\big) < 1$. It follows that there exists a fixed point $r^* \in \big(\frac{1}{2}, 1\big)$. In addition, recall that G is strictly increasing in r. Hence for all such fixed points $r^* = G\big(r^*\big) > G\big(\frac{1}{2}\big) = m_A$.

(iii) First note that by Proposition 3(ii) it follows that for $m_A \neq 1/2$ there is no equilibrium such that $r = m_A$.

Thus it remains to evaluate Equation (7) at $m_A = r = \frac{1}{2}$:

$$\begin{aligned} G\big(\tfrac{1}{2}\big) &= \frac{m_A\big[1 - F\big(\frac{\alpha}{2}\big)\big]}{m_A\big[1 - F\big(\frac{\alpha}{2}\big)\big] + (1-m_A)\big[1 - F\big(\frac{\alpha}{2}\big)\big]} \\ &= \frac{m_A}{m_A + (1-m_A)} \\ &= m_A. \end{aligned} \tag{24}$$

Hence $r = \frac{1}{2}$ is a fixed point of G if and only if $m_A = \frac{1}{2}$.

(iv) Consider an arbitrary r^1. By (7), for m_A sufficiently large $G(r^1; m_A) = r^2 > \frac{1}{2}$. As G is strictly increasing in r and $r^2 > \frac{1}{2}$, $r^3 = G(r^2; m_A) > G(\frac{1}{2}; m_A)$. As $m_A > \frac{1}{2}$, $G(\frac{1}{2}; m_A) > \frac{1}{2}$. Hence $r^3 > \frac{1}{2}$. Iterating this argument, $r^t > \frac{1}{2}$ for all $t > 1$.

Thus r^t must converge (by Proposition 1) to a fixed point $r^* > \frac{1}{2}$. By part (iii), $r^* > m_A$ for all such fixed points.

(v) Consider $r^1 = \frac{1}{2}$. Recall that G is continuous, $G\big(\frac{1}{2}\big) = m_A > \frac{1}{2}$ by Equation (24) and G is increasing. Thus we have $r^t > m_A$ for all $t \geq 2$. □

Proof of Proposition 4. Suppose $r_A^T > \frac{1}{2}$ for some $T \geq 1$. By assumption, $m_{Aa} = m_{Ab} = m_{Ba} = m_{Bb} = \frac{1}{4}$, thus

$$G_{Aa}(r^T) + G_{Ab}(r^T) = \frac{p_{Aa}^T + p_{Ab}^T}{p_{Aa}^T + p_{Ab}^T + p_{Ba}^T + p_{Bb}^T} \tag{25}$$

Given, $r_A^T > \frac{1}{2}$ we have $p_{Aa}^T > p_{Ba}^T$ and $p_{Ab}^T > p_{Bb}^T$ independent of α and β. Consequently $p_{Aa}^T + p_{Ab}^T > p_{Ba}^T + p_{Bb}^T$ and $r_A^{T+1} = G_{Aa}(r^T) + G_{Ab}(r^T) > \frac{1}{2}$. By iterating the result follows immediately. □

Proof of Proposition 5. (i) We are interested in the participation of agents of type $\theta = 1$:

$$r_A = \frac{m_{Aa}p_{Aa} + m_{Ab}p_{Ab}}{m_{Aa}p_{Aa} + m_{Ba}p_{Ba} + m_{Ab}p_{Ab} + m_{Bb}p_{Bb}} \tag{26}$$

We shall define $0 \leq x \leq 1, 0 \leq z \leq 1$ such that $m_{Aa} = x \cdot m_A, m_{Bb} = z \cdot (1 - m_A)$:

$$r_A = \frac{\overbrace{xm_Ap_{Aa} + (1-x)m_Ap_{Ab}}^{\equiv \mathcal{N}}}{\underbrace{xm_Ap_{Aa} + (1-z)(1-m_A)p_{Ba} + (1-x)m_Ap_{Ab} + z(1-m_A)p_{Bb}}_{\equiv \mathcal{D}}} \tag{27}$$

If the derivative $\frac{d}{dm_A} r_A > 0$ then the assertion holds. The numerator of the derivative is given by:

$$\begin{aligned} & \mathcal{D}\big[xp_{Aa} + (1-x)p_{Ab}\big] - \mathcal{N}\big[xp_{Aa} - (1-z)p_{Ba} + (1-x)p_{Ab} - zp_{Bb}\big] \\ = \; & (\mathcal{D} - \mathcal{N})\big[xp_{Aa} + (1-x)p_{Ab}\big] - \mathcal{N}\big[-(1-z)p_{Ba} - zp_{Bb}\big] \\ = \; & \big[(1-z)(1-m_A)p_{Ba} + z(1-m_A)p_{Bb}\big]\big[xp_{Aa} + (1-x)p_{Ab}\big] \\ & + \big[xm_Ap_{Aa} + (1-x)m_Ap_{Ab}\big]\big[(1-z)p_{Ba} + zp_{Bb}\big] \end{aligned}$$

Note that the latter equation is positive since x, z are positive and p_k^t are assumed to be positive for all $k \in \{Aa, Ab, Ba, Bb\}$. This establishes part (i).

(ii) Begin in state $r^t_A = r^t_a = 1/2$. Then participation rates are the same for each group: $p^t_k = \hat{p}$ for all $k \in \{Aa, Ab, Ba, Bb\}$. Hence

$$\begin{aligned} r^{t+1}_k &= \frac{m_k p^t_k}{m_{Aa}p^t_{Aa} + m_{Ab}p^t_{Ab} + m_{Ba}p^t_{Ba} + m_{Bb}p^t_{Bb}} \\ &= \frac{m_k \hat{p}}{m_{Aa}\hat{p} + m_{Ab}\hat{p} + m_{Ba}\hat{p} + m_{Bb}\hat{p}} = m_k. \end{aligned} \tag{28}$$

As $r^{t+1}_A = r^{t+1}_{Aa} + r^{t+1}_{Ab}$ which equals $m_{Aa} + m_{Ab} = m_A > 1/2$, $p^{t+1}_{Aa} > p^{t+1}_{Ba}$ and $p^{t+1}_{Ab} > p^{t+1}_{Bb}$. Further, as $r^{t+1}_a = r^{t+1}_{Aa} + r^{t+1}_{Ba} \geq \frac{1}{2}$, we have that p^{t+1}_{Bb} is minimal and p^{t+1}_{Aa} is maximal. Finally, as $m_A > m_a$ and $\alpha \geq \beta$ we have $p^{t+1}_{Ab} > p^{t+1}_{Ba}$. Thus:

$$\begin{aligned} r^{t+2}_A &= \frac{m_{Aa}p^{t+1}_{Aa} + m_{Ab}p^{t+1}_{Ab}}{m_{Aa}p^{t+1}_{Aa} + m_{Ab}p^{t+1}_{Ab} + m_{Ba}p^{t+1}_{Ba} + m_{Bb}p^{t+1}_{Bb}} \\ &> \frac{m_{Aa}p^{t+1}_{Aa} + m_{Ab}p^{t+1}_{Ab}}{m_{Aa}p^{t+1}_{Aa} + m_{Ab}p^{t+1}_{Ab} + m_{Ba}p^{t+1}_{Ab} + m_{Bb}p^{t+1}_{Ab}} \\ &> \frac{m_{Aa}p^{t+1}_{Aa} + m_{Ab}p^{t+1}_{Ab} - [m_{Aa}p^{t+1}_{Aa} - m_{Aa}p^{t+1}_{Ab}]}{m_{Aa}p^{t+1}_{Aa} + m_{Ab}p^{t+1}_{Ab} + m_{Ba}p^{t+1}_{Ab} + m_{Bb}p^{t+1}_{Ab} - [m_{Aa}p^{t+1}_{Aa} - m_{Aa}p^{t+1}_{Ab}]} \\ &= \frac{m_{Aa}p^{t+1}_{Ab} + m_{Ab}p^{t+1}_{Ab}}{m_{Aa}p^{t+1}_{Ab} + m_{Ab}p^{t+1}_{Ab} + m_{Ba}p^{t+1}_{Ab} + m_{Bb}p^{t+1}_{Ab}} \\ &= m_A. \end{aligned} \tag{29}$$

The last inequality holds as subtracting the same positive amount from the numerator and denominator of a fraction that is smaller than one, decreases the value of the fraction.

Iterating this argument, we have $r^T_A > m_A$ for all $T > t$. Since the process converges to an equilibrium, the result follows.

(iii) By Equation (28) it follows that $r^{t+1}_k = m_k$ for all $k \in \{Aa, Ab, Ba, Bb\}$ holds in particular if $p^t_k = \hat{p}$ for all $k \in \{Aa, Ab, Ba, Bb\}$. Since F is strictly increasing the latter holds if and only if $r^t_A = r^t_a = \frac{1}{2}$ and thus $m_A = m_a$. In contrast, by Equations (29) if $m_A > \frac{1}{2}$, $r_A = m_A$ cannot hold in any equilibrium. The result follows. □

Proof of Proposition 6. We shall omit the superscript t for brevity. Given a subsidy $s > 0$ to B types we shall write for the participation rates (remarking that they are unchanged for Aa and Ab):

$$\begin{aligned} p_{Aa} &\equiv p_{Aa}(\alpha(1 - r_A) + \beta(1 - r_a)) \\ p_{Ab} &\equiv p_{Ab}(\alpha(1 - r_A) + \beta r_a) \\ p_{Ba}(s) &\equiv p_{Ba}(\alpha r_A + \beta(1 - r_a) - s) \\ p_{Bb}(s) &\equiv p_{Bb}(\alpha r_A + \beta r_a - s) \end{aligned} \tag{30}$$

We are interested under what condition the representation of a types is independent of the subsidy $s > 0$ to B types, that is (where we write for brevity $r_a = r_a(0)$):

$$r_a(s) = r_a \tag{31}$$

$$\Leftrightarrow \frac{m_{Aa}p_{Aa} + m_{Ba}p_{Ba}(s)}{m_{Aa}p_{Aa} + m_{Ab}p_{Ab} + m_{Ba}p_{Ba}(s) + m_{Bb}p_{Bb}(s)} = \frac{m_{Aa}p_{Aa} + m_{Ba}p_{Ba}}{m_{Aa}p_{Aa} + m_{Ab}p_{Ab} + m_{Ba}p_{Ba} + m_{Bb}p_{Bb}} \tag{32}$$

$$\Leftrightarrow m_{Aa}p_{Aa}m_{Bb}p_{Bb} + [m_{Ab}p_{Ab} + m_{Bb}p_{Bb}]m_{Ba}p_{Ba}(s) = m_{Ab}p_{Ab}m_{Ba}p_{Ba} + [m_{Aa}p_{Aa} + m_{Ba}p_{Ba}]m_{Bb}p_{Bb}(s) \tag{33}$$

Recall that F is strictly increasing and $F(\alpha, \beta) < 1$, thus $p_{Ba}(s) \neq p_{Ba}(0)$, $p_{Bb}(s) \neq p_{Bb}(0)$, and $p_k > 0$ by assumption for all $k \in \{Aa, Ab, Ba, Bb\}$. Therefore Equation (33) is a non-trivial dependency on the population shares (in addition to the condition that $m_{Aa}+m_{Ab}+m_{Ba}+m_{Bb}=1$). Thus, given r^t and a subsidy $s > 0$ the set of population shares such that $r_a^{t+1}(s) = r_a^{t+1}(0)$ has measure zero with respect to all possible population shares. □

Proof of Proposition 7. To simplify notation we omit the superscripts for time. To achieve a representative state $r_k = m_k$, $k \in \{Aa, Ab, Ba, Bb\}$, s_k must satisfy for all k:

$$\begin{aligned} m_k &= \frac{m_k p_k(s_k)}{m_{Aa}p_{Aa}(s_{Aa}) + m_{Ab}p_{Ab}(s_{Ab}) + m_{Ba}p_{Ba}(s_{Ba}) + m_{Bb}p_{Bb}(s_{Bb})} \\ \iff p_k(s_k) &= m_{Aa}p_{Aa}(s_{Aa}) + m_{Ab}p_{Ab}(s_{Ab}) + m_{Ba}p_{Ba}(s_{Ba}) + m_{Bb}p_{Bb}(s_{Bb}) \end{aligned} \tag{34}$$

Thus $p_k =: \hat{p}$ for all $k \in \{Aa, Ab, Ba, Bb\}$ and therefore the arguments of F in Equation (15) need to be all the same. This holds if and only if the following holds:

$$\begin{aligned} s_{Ab} &= s_{Aa} + \beta \cdot (2r_a - 1) \\ s_{Ba} &= s_{Aa} + \alpha \cdot (2r_A - 1) \\ s_{Bb} &= s_{Aa} + \alpha \cdot (2r_A - 1) + \beta \cdot (2r_a - 1) \end{aligned} \tag{35}$$

The additional requirement we posit is that (s_A, s_B, s_a, s_b) are self-financing:

$$\begin{aligned} m_{Aa}\hat{p}s_{Aa} + m_{Ab}\hat{p}s_{Ab} + m_{Ba}\hat{p}s_{Ba} + m_{Bb}\hat{p}s_{Bb} = 0 \\ \iff m_{Aa}s_{Aa} + m_{Ab}s_{Ab} + m_{Ba}s_{Ba} + m_{Bb}s_{Bb} = 0 \end{aligned} \tag{36}$$

Inserting Equations (35) into the latter equation we find:

$$\begin{aligned} & m_{Aa}s_{Aa} + m_{Ab}s_{Ab} + m_{Ba}s_{Ba} + m_{Bb}s_{Bb} = 0 \\ \iff & m_{Aa}s_{Aa} + m_{Ab}\big[s_{Aa} + \beta \cdot (2r_a - 1)\big] + m_{Ba}\big[s_{Aa} + \alpha \cdot (2r_A - 1)\big] + \\ & m_{Bb}\big[s_{Aa} + \alpha \cdot (2r_A - 1) + \beta \cdot (2r_a - 1)\big] = 0 \\ \iff & s_{Aa}(m_{Aa} + m_{Ab} + m_{Ba} + m_{Bb}) + m_B \cdot \alpha \cdot (2r_A - 1) + m_b \cdot \beta \cdot (2r_a - 1) = 0 \\ \iff & s_{Aa} = -[(1 - m_A) \cdot \alpha \cdot (2r_A - 1) + (1 - m_a) \cdot \beta \cdot (2r_a - 1)] \end{aligned} \tag{37}$$

□

Proof of Proposition 8. To achieve a representative state, participation rates must be equalized across groups $k \in \{Aa, Ab, Ba, Bb\}$, which means the following identity-based payoffs must be equalized:

$$Aa \ : \ \alpha\left[\delta(\gamma_{Aa} + \gamma_{Ab}) + (1-\delta)r_A\right] + \beta\left[\delta(\gamma_{Aa} + \gamma_{Ba}) + (1-\delta)r_a\right] \tag{38}$$

$$Ab \ : \ \alpha\left[\delta(\gamma_{Aa} + \gamma_{Ab}) + (1-\delta)r_A\right] + \beta\left[\delta(\gamma_{Ab} + \gamma_{Bb}) + (1-\delta)(1-r_a)\right] \tag{39}$$

$$Ba \ : \ \alpha\left[\delta(\gamma_{Ba} + \gamma_{Bb}) + (1-\delta)(1-r_A)\right] + \beta\left[\delta(\gamma_{Aa} + \gamma_{Ba}) + (1-\delta)r_a\right] \tag{40}$$

$$Bb \ : \ \alpha\left[\delta(\gamma_{Ba} + \gamma_{Bb}) + (1-\delta)(1-r_A)\right] + \beta\left[\delta(\gamma_{Ab} + \gamma_{Bb}) + (1-\delta)(1-r_a)\right]. \tag{41}$$

This can be achieved by equating perceived representation along each dimension as follows:

$$\delta(\gamma_{Aa} + \gamma_{Ab}) + (1-\delta)r_A = \delta(\gamma_{Ba} + \gamma_{Bb}) + (1-\delta)(1-r_A) \tag{42}$$

$$\delta(\gamma_{Aa} + \gamma_{Ba}) + (1-\delta)r_a = \delta(\gamma_{Ab} + \gamma_{Bb}) + (1-\delta)(1-r_a). \tag{43}$$

Using the fact that $\sum_{k\in\{Aa,Ab,Ba,Bb\}} \gamma_k = 1$, policy (δ, γ) is a solution to equations (42)-(43) if and only if:

$$
\begin{aligned}
\gamma_{Aa} + \gamma_{Ab} &= \tfrac{1}{2} - \tfrac{1-\delta}{\delta}\left(r_A - \tfrac{1}{2}\right) && (44)\\
\gamma_{Aa} + \gamma_{Ba} &= \tfrac{1}{2} - \tfrac{1-\delta}{\delta}\left(r_a - \tfrac{1}{2}\right). && (45)
\end{aligned}
$$

By (44)-(45),for $\gamma_k \geq 0$ for each k, in particular when $r_A = r_a = 1$, it is necessary that $\delta \geq \max\left\{\frac{r_A^t - \frac{1}{2}}{r_A^t}, \frac{r_a^t - \frac{1}{2}}{r_a^t}\right\}$. □

Proof of Proposition 9. It is sufficient to compare steady-state participation rates resulting from Propositions 7 and 8. Hence compare the argument of F in Equation (15). For self-financing subsidies, it is:

$$
\begin{aligned}
& \alpha(1-m_A) + \beta(1-m_a) + (1-m_A)\cdot\alpha\cdot(2m_A-1) + (1-m_a)\cdot\beta\cdot(2m_a-1) && (46)\\
= \; & (1-m_A)\cdot\alpha\cdot 2m_A + (1-m_a)\cdot\beta\cdot 2m_a && (47)\\
< \; & \frac{\alpha}{2} + \frac{\beta}{2} && (48)
\end{aligned}
$$

where we used the condition $m_A \neq \frac{1}{2}$ and/or $m_a \neq \frac{1}{2}$ in the latter inequality. For role models, it is:

$$
\begin{aligned}
& \alpha\left[\delta(\gamma_{Aa} + \gamma_{Ab}) + (1-\delta)m_A\right] + \beta\left[\delta(\gamma_{Aa} + \gamma_{Ba}) + (1-\delta)m_a\right] && (49)\\
= \; & \alpha\left[\delta\left(\tfrac{1}{2} - \tfrac{1-\delta}{\delta}\left(m_A - \tfrac{1}{2}\right)\right) + (1-\delta)m_A\right] + \beta\left[\delta\left(\tfrac{1}{2} - \tfrac{1-\delta}{\delta}\left(m_a - \tfrac{1}{2}\right)\right) + (1-\delta)m_a\right] && (50)\\
= \; & \frac{\alpha}{2} + \frac{\beta}{2}, && (51)
\end{aligned}
$$

where the second line follows from Equations (44)-(45).

As F is strictly increasing, the participation rate (which is equal across groups in a representative state) is strictly larger for self-financing subsidies than for role models, thus yielding higher economic output. □

Proof of Proposition 10. We shall assume, for ease of exposition, that $\alpha = \beta = 1$. Inserting Equation (37) into Equation (15) we find that in the representative state the argument of F is given by $\zeta := 2\alpha m_a(1\text{-}m_A) + 2\beta m_A(1\text{-}m_a) = 1$. The assertion then holds if:

$$
\begin{aligned}
& m_{Aa}\int_{1\text{-}r_A+\frac{1}{2}}^{\infty} yf(y)dy + m_{Ab}\int_{1\text{-}r_A+\frac{1}{2}}^{\infty} yf(y)dy + m_{Ba}\int_{r_A+\frac{1}{2}}^{\infty} yf(y)dy + m_{Bb}\int_{r_A+\frac{1}{2}}^{\infty} yf(y)dy > \\
& m_{Aa}\int_{1}^{\infty}(y\text{-}s_{Aa})f(y)dy + m_{Ab}\int_{1}^{\infty}(y\text{-}s_{Ab})f(y)dy + m_{Ba}\int_{1}^{\infty}(y\text{-}s_{Ba})f(y)dy + m_{Bb}\int_{1}^{\infty}(y\text{-}s_{Bb})f(y)dy
\end{aligned}
\qquad (52)
$$

Note that the right-hand side of Inequality (52) is equal to

$$
(m_A + (1-m_A))\int_{\zeta=1}^{\infty} yf(y)dy \qquad (53)
$$

as the taxes are designed to be self-financing. Since $r_A > \frac{1}{2}$ we can rewrite Inequality (52):

$$
\begin{aligned}
& m_A\int_{1\text{-}r_A+\frac{1}{2}}^{\infty} yf(y)dy + (1-m_A)\int_{r_A+\frac{1}{2}}^{\infty} yf(y)dy > (m_A + (1-m_A))\int_{1}^{\infty} yf(y)dy && (54)\\
\iff \; & m_A\int_{1\text{-}r_A+\frac{1}{2}}^{1} yf(y)dy > (1-m_A)\int_{1}^{r_A+\frac{1}{2}} yf(y)dy && (55)
\end{aligned}
$$

By assumption, $y \cdot f(y)$ is decreasing and thus the latter holds, as by assumption $m_A > \frac{1}{2}$. □