Emergence of Stereotypes Under Group Competition

Dan McGee*

March 17, 2021

Abstract

How and why do racial stereotypes arise, and how do negative stereotypes harm the stereotyped group? I study a model of stereotypes as motivated reasoning when social groups interact. Even though group differences are intrinsically meaningless, agents develop 'racial' beliefs where they think these differences mark differences in ability when there are sufficient incentives to stereotype. These beliefs lead to asymmetric equilibria with biased first-order and higher-order beliefs, where one group chooses to denigrate the other to achieve a competitive advantage, while believing that this other group 'sees us as we see ourselves.' Furthermore, stereotypes arise from competitive incentives despite the absence of 'inherent' animosity between groups. (JEL: D74, D84, D91, Z13)

*Princeton University. I would like to thank Avidit Acharya, Roland Bénabou, Sylvain Chassang, Octavia Ghelfi, Faruk Gul, Alice Hsiaw, Stephen Morris, Salvatore Nunnari, and Pietro Ortoleva for their discussions and suggestions during this project and seminar participants at Princeton and Columbia for their helpful comments.

1 Introduction

> *"What often appears to be an eruption of 'traditional hatreds' on closer examination... [involves] economic issues that are real and immediate.*
>
> - S. Steinberg, *The Ethnic Myth.*

As soon as people perceive a distinction between 'them' and 'us,' it becomes natural to wonder how we should see them and how they see us. These beliefs often misrepresent the groups in question, let alone any individual within those groups, hardening from exaggeration to outright stereotype. As many of the most protracted, vicious conflicts in the world occur along the lines of identity: nation, race, and class, among others, it is natural to think that hostility between groups leads to competition between them. Yet as the famous 'Robbers Cave' experiment of Sherif and Sherif (1953) showed, groups put in conflict can rapidly develop deep enmity towards each other, even when the group categorization is quite shallow. Furthermore, when the groups' incentives changed, rewarding intergroup cooperation, rather than competition, these groups' hostility vanished as rapidly as it had appeared. Thus, it is not the mere fact of social division that generates negative beliefs about social groups, but the nature of the interactions between groups that determines the content of these beliefs. When groups come into conflict over resources or social position, members of one group face an incentive to develop certain negative stereotypes about the competence of the opposing group, so that they view this conflict as winnable and worth investing effort into, especially when conflict is highly rewarding for the victors (Spears et al., 1997; Leonard et al., 2011). Therefore, stereotyping and stigmatization need not arise from inherent animus between groups; in fact, the presence of intergroup competition and the possibility of gains by discriminating can itself create prejudicial beliefs.[1]

Earlier work in economics has examined stereotyping from two perspectives. First, the

[1]See also Kendi (2016) and Darity et al. (2017) on discrimination as a cause of prejudice. A separate literature argues that stereotypes also serve to morally legitimize the consequences of discrimination, which I do not address here, e.g., Allport (1954), Glaser (2005), Almås et al. (2010), and Crandall et al. (2011).

classical view of stereotyping in economics, starting with Phelps (1972) and Arrow (1973), and expanded upon by Coate and Loury (1993) and others, treats stereotypes as accurate summaries of group-level differences in relevant characteristics. These stereotypes can be self-fulfilling prophesies if they affect individuals' willingness to invest in improving skills, but they do not encode false information, as beliefs are formed based on rational expectations. A newer approach argues that stereotypes arise due to cognitive schemas that produce useful, albeit biased, representations of the world, thus saving scarce cognitive resources (Schneider, 2004). For example, Bordalo et al. (2016) develop a model of stereotyping based on the representativeness heuristic (Tversky and Kahneman, 1983): agents overweight the prevalence of a trait in a group when that trait is highly representative of the group in question.[2] Similarly, Chauvin (2019) examines stereotypes arising from the 'fundamental attribution error,' where agents ignore the role of beliefs in determining actions, attributing all differences in choices to differences in personal traits. Frick et al. (2019) study 'assortativity neglect,' where agents assume their social circle represents an accurate cross-section of society, despite the existence of assortative matching. Although this approach explains neutral exaggerations well, such as 'Irish people have red hair,' it seems less fitting for the most damaging, negative stereotypes of stigmatized groups, which more closely resemble motivated reasoning.[3]

As described in Bénabou (2015), there are two key distinctions between biased beliefs arising from cognitive errors and those arising from purposefully incorrect reasoning: the latter are asymmetrically held across groups and have asymmetric effects and the beliefs they create are emotionally charged, so individuals are loath to have their beliefs corrected. For example, Gneezy et al. (2003) find that women underperform in mixed-gender tournaments, but perform much better in single-gender tournaments, while the same effect is not

[2]See also Gennaioli and Shleifer (2010) for a formal model of the representativeness heuristic and Bordalo et al. (2019), who study stereotypes about gender in a laboratory experiment.

[3]For example, many Americans hold negative stereotypes of immigrants as lazy and criminal, despite evidence to the contrary (Fuligni, 2007; Pew Research, 2015). Bohren et al. (2019) find evidence of statistical discrimination driven by inaccurate stereotypical beliefs and Gorski (2008) discusses inaccurate stereotypes about the poor. Heidhues et al. (2020) advance a theory of stereotyping as misspecified learning in the face of dogmatic overconfidence in one's own ability (i.e., hedonic motivated reasoning).

present for men. Furthermore, Günther et al. (2010) show that this effect is present only for tasks classified as masculine and not for tasks classified as feminine, which suggests that social beliefs affect behaviour, as Coffman (2014) and Bordalo et al. (2019) show beliefs about relative ability differ substantially according to the gendered nature of the task. Likewise, the broader phenomena of 'stereotype threat' and 'stereotype lift' are asymmetrically present for favoured versus unfavoured groups in society. Per Steele and Aronson (1995) and Walton and Spencer (2009), stigmatized groups perform worse in settings where they are stereotypically believed to perform poorly. In response, unstigmatized groups compete more aggressively and perform better when competing against a member of a stigmatized group, which researchers have termed stereotype lift (Walton and Cohen, 2003). Furthermore, these effects are present despite differences in individuals' agreement with the stereotypes. Walton and Cohen (2003) show that stereotype lift arises because members of unstigmatized groups believe stereotypes about stigmatized groups and think the outgroup accepts these stereotypes. By contrast, in Huguet and Régner (2009), women experienced stereotype threat, despite disagreeing with gender stereotypes about their ability, provided they believed that others stereotyped them. Finally, participants in these studies frequently claimed to see themselves as representatives of their group, in competition with a comparison group, even when the experiment evaluates them alone (Steele, 2010).

Hence, I develop a model of stereotyping as motivated reasoning, where stereotypes arise in response to underlying economic and strategic incentives. I embed a parsimonious model of strategic interaction into a setting of cultural transmission as in Tabellini (2008) and Dessi (2008). Parents choose belief distortions for their children to influence their children's behaviour in strategic interactions with other agents. Although, parents care about their children's payoffs, a social externality creates a wedge between parents' and children's preferences, so parents wish to distort beliefs to encourage higher effort than the child would prefer.[4] Given the role of beliefs about the stereotypes of others discussed above and in

[4]Equivalently, the parents could be imperfectly altruistic as in Bisin and Verdier (2001) or children could have limited willpower (Bénabou and Tirole, 2002).

Section 2, I specifically focus on distortions to higher-order beliefs, asking whether agents correctly perceive the beliefs and biases of others. I study two well-documented patterns of higher-order belief distortions, '*stereotype awareness*' and '*belief projection*.'[5] Under the former, children correctly identify how the others in society stereotypes them, but may hold inaccurate stereotypes of their own (Sigelman and Tuch, 1997).[6] Under the latter, children believe that other individuals agree with the children's stereotypes about them, a form of false consensus bias (Ross et al., 1977). These generate novel interactions with distorted first-order beliefs and thus, novel patterns of behaviour that would not be present if only first-order beliefs were distorted. Specifically, although all agents hold distorted beliefs in equilibrium, not all agents distort their beliefs in the same way; different patterns of belief distortions are best responses to each other.

I assume all individuals have one of two physical 'markers,' which are observable to other members of society. Even though the markers are unrelated to individuals' true ability, parents can choose to teach their children that individuals with one marker are more productive than individuals with the other marker, in the sense that their return to effort is higher. By doing so, parents add social meaning to the intrinsically meaningless physical markers. As such, parents' choices construct 'race' as a belief that physical markers denote meaningful underlying differences in ability and that individuals with a different marker are a member of a different 'racial' group. When children believe that physical markers denote meaningful group differences, this shapes how they interpret information they receive from their parent. If their parent denigrates opponents from the child's own perceived group, the child interprets that message as a negative signal of their own ability. Likewise, if the parent increases a child's confidence in their own ability, the child also believes that other members of her perceived group have high ability. By contrast, children do not see messages about the ability of individuals with a different marker as informative about their

[5]See Section 2 for empirical evidence on these patterns of higher-order beliefs.

[6]Related is the 'bias blind spot, where agents perceive biases in others' beliefs, but fail to recognize biases in their own (Pronin et al., 2002).

own ability. As a result, distorted higher-order beliefs arise specifically for interactions with individuals bearing a different marker, alongside negative first-order beliefs about the other group's competence (stereotypes). When a child interacts with a member of her own group, her parent is constrained by her child's belief that her same-group opponent has the same ability as she does, which prevents her parent from stereotyping her opponent. Furthermore, in the absence of stereotypes, parents have no incentive to distort their child's higher-order beliefs. Thus, higher-order beliefs are endogenously correct for within-group competition. By contrast, stereotypes can arise in between-group competition, creating an incentive to distort higher-order beliefs.

I examine when individuals will construct 'racial' beliefs by attaching meaning to the physical markers. Since an agent holds a racial belief when they think that physical markers denote underlying differences in ability, they can only arise because children are not perfectly informed about individuals' ability. If it were common knowledge that ability does not differ across groups, then racial beliefs could not form. However, limited knowledge is not enough; parents also need strong enough incentives to stereotype. These incentives arise when the 'wedge' generated by the externality is large, when the payoffs to winning the competition is much larger than the payoff to losing, and when children's strategies are very sensitive to beliefs. When these incentives are not present, the unique outcome is a symmetric equilibrium: no child holds stereotypes about another individual and all children behave identically in competitive settings. However, when the incentives are sufficiently strong, this symmetric equilibrium becomes unstable and there instead arise stable asymmetric equilibria.

In these asymmetric equilibria, the two groups distort their first and higher-order beliefs in different ways. One 'favoured' group adopts an aggressive strategy, holding a negative stereotype about the ability of individuals from the other 'unfavoured' group and incorrectly believing that this group shares this stereotype. This stereotype is a more negative view of the unfavoured group than either group's self-image. Furthermore, as the stakes of the competition rise, the stereotype becomes more negative, so that it eventually falls below the

group's true ability. When competing, since they expect the unfavoured group to have low ability and think that this group internalizes this stereotype, the favoured group chooses to compete more aggressively. Although the unfavoured group disagrees with the stereotype, they choose to be aware of how the favoured group sees them, so that they can respond to the favoured group's treatment of them. Since the unfavoured group is aware of the stereotype, even as they reject it, they anticipate the favoured group's aggressive behaviour and compete less. Thus, the distortions in higher-order beliefs emerge in response to the biases in first-order beliefs. In these equilibria, the favoured group benefits from holding negative stereotypes about the unfavoured group and achieve higher payoffs from competition than the unfavoured group does. Members of the favoured group incorrectly believe that the unfavoured group is of lower ability than they are and engage in discriminatory behaviour due to those beliefs. However, those incorrect beliefs arise because competitive interactions create an incentive for parents to teach their children stereotypes.

In developing this model, I draw on a literature which examines the strategic benefits of distorted beliefs. Several papers have documented mechanisms by which distorted beliefs can help agents overcome principal-agent conflicts, including intrapersonal or intergenerational conflict (Brocas and Carrillo, 2000; Carrillo and Mariotti, 2000; Bénabou and Tirole, 2002; Compte and Postlewaite, 2004). Furthermore, recent experimental and theoretical work has highlighted that incorrect beliefs about others' strategies can improve agents' strategic positions (Charness et al., 2018; Heller and Winter, 2020). In line with this literature, distorted beliefs about groups can improve the welfare of some agents. However, it is not the case that all agents distort their beliefs in the same way; different higher-order beliefs are best responses to each other.[7] I also draw on the broader literature on cultural transmission, which focuses on the incentives for parents to consciously transmit certain beliefs, behaviours, and preferences to their offspring. As shown by Bisin and Verdier (2001), parents may not want

[7]There is also a literature which examines hedonic motives for biased beliefs (Caplin and Leahy, 2001; Bénabou and Tirole, 2002; Kőszegi, 2006, 2010; Brunnermeier and Parker, 2005). I abstract from these motives to develop an instrumental case for stereotype formation; however, adding belief utility would not qualitatively change the results.

their children to have the same preferences as they do, even when parents evaluate their children's outcomes according to the parents' own preferences. Bisin and Verdier (1998) study cultural transmission of status-seeking preferences, where status-seeking behaviour exhibits similar anti-coordination incentives, but do not consider the role of beliefs. Also related is Guiso et al. (2008) who discuss social capital as a belief in the trustworthiness of others.

Theories of discrimination in economics emphasize statistical discrimination, based on accurate stereotypes (Arrow, 1973), or taste-based discrimination, based on individuals' personal prejudices (Becker, 1957). However, neither theory applies in this setting: agents do not inherently prefer those with the same identity marker and productivity is identical across groups. Instead, this setting represents discrimination for advantage.[8] Favoured group parents want their children to compete more aggressively, benefiting themselves at the unfavoured group's expense. Since doing so is costly, stereotypes induce the desired behaviour when children would prefer to not discriminate when fully informed. Drawing on a similar idea of discrimination for advantage, Harbaugh and To (2014) shows how minorities are more vulnerable to opportunism by members of the majority, Lagerlöf (2020) finds that firms can strategically discriminate to segment the labour market and lower wages, and Dewan and Wolton (2020) show that workers from a majority group may support symbolic policies that raise social identities' salience to improve their own labour market outcomes.[9] The notion that 'race' represents social meaning imputed to intrinsically irrelevant characteristics is closely related to Darity et al. (2006), who distinguish between one's exogenous physical appearance and the endogenous social norms which determine whether agent choose to respond differently to others based on appearance.

The outline of the paper is as follows. Section 2 presents evidence on higher-order beliefs and stereotyping. Section 3 analyzes children's behaviour under competition with distorted beliefs. Section 4 examines the baseline model of stereotyping as motivated reasoning through parents' choices. Section 5 discusses some extensions. Section 6 concludes.

[8]See also Darity et al. (2015) and references therein.

[9]See also Eeckhout (2006) and Pęski and Szentes (2013) on discriminatory equilibria in dynamic settings.

2 Psychology and Stereotyping

To discipline the model, I focus on two patterns of higher-order belief distortions that empirical evidence suggests are commonly found in real-world environments.

The first pattern is 'Stereotype Awareness,' which draws from extensive research in social psychology showing that while individuals can often identify biases and errors in others' reasoning, they fail to realize that they are subject to these same biases and errors (Pronin et al., 2002).[10] In particular, social psychologists have examined 'meta-stereotypes,' which they define as the belief that members of a group have about the stereotypes held by other social groups about their group, and which economists would recognize as an individual's second-order belief. Since researchers can simultaneously survey a group about their meta-stereotypes and the other group about their stereotypes, they can directly verify whether the meta-stereotypes are accurate. This literature was started by Sigelman and Tuch (1997), who found that Black Americans hold largely accurate beliefs about the stereotypes that White Americans have about them. Later research by Vorauer et al. (1998) and Vorauer et al. (2000) on White Canadians and Aboriginals, Torres and Charles (2004) on Black and White American university students, and Schmitt and Wirth (2009) and Martinez et al. (2010) on men and women, find that the lower-status group (Aboriginals, Black Americans, and women, respectively) tend to hold broadly accurate meta-stereotypes, while the higher-status groups hold inaccurate meta-stereotypes.

The second structure is 'Belief Projection,' which represents the tendency for individuals to think and act as though others share their beliefs, a form of false consensus (Ross et al., 1977). In particular, Watt and Larkin (2010) find that highly prejudiced individuals perceive more support for their views than truly exists in the population. Since social norms influence individual behaviour, misperceptions about the beliefs of others strongly predict behaviour in social settings (Botvin et al., 1992; Bauman and Geher, 2002; Bursztyn et al., 2020). Indi-

[10] Earlier work by Griffin and Ross (1991), and Armor (1999) similarly find an 'illusion of objectivity' for individual biases.

viduals who project incorrectly assume that others share their introspection and thus reach similar conclusions. In particular, Ross et al. (1977) argue that instances of agreement are more easily recalled than instances of disagreement, so individuals overestimate consensus due to availability bias (Tversky and Kahneman, 1973; Marks and Miller, 1987).[11] Ross et al. (1977) also find that false consensus extends to higher-order beliefs: individuals who project tend to overestimate the degree to which others perceive a false consensus.[12]

Differential awareness of other groups' beliefs starts in childhood and parental teaching helps develop these beliefs (McKown and Weinstein, 2003; Hughes et al., 2006; Stevenson and Arrington, 2009). Hughes et al. (2006) find that racial minority parents explicitly teach their children how the racial majority views them as a way of teaching them to cope with discrimination. A sequence of studies examined the effects of accurate and inaccurate higher-order beliefs and found various mechanisms through which these beliefs affect well-being. Finchilescu (2005, 2010), Wakefield et al. (2013), van Leeuwen et al. (2014), and Fowler and Gasiorek (2020) find evidence that higher-order beliefs affect social groups' willingness to interact with and help members of their ingroup versus members of an outgroup. Among lower-status groups, negative meta-stereotypes lead Black students to perform worse academically in settings with White students by diverting effort away from study (Torres and Charles, 2004), lead women to compete with men less for high-status positions at work (Schmitt and Wirth, 2009), and lead women and ethnic minorities to search less intensively for employment (Owuamalam and Zagefka, 2014).[13] However, these negative effects were balanced by better readiness among these low-status groups for how high-status groups would treat them (Hughes et al., 2006; Jerald et al., 2017). By contrast, native Italians who incorrectly believed African immigrants shared their negative stereotypes about immigrants experienced greater willingness to work and higher overall well-being (Matera et al., 2015).

[11] See also Frick et al. (2019), who find false consensus arising from assortativity neglect.

[12] Also related are Madarász (2012, 2016) and Danz et al. (2018), who study information projection. To the extent that beliefs depend on information, even if misremembered or favourably interpreted, belief projection and information projection should be related phenomena.

[13] See also discussions by Pratto et al. (1997, 2006) about status-seeking behaviour in various settings.

3 Competition Under Motivated Reasoning

I assume the population contains a unit mass of children, each of whom has a parent, and that each child has exactly one physical marker, denoted W and B, which can be easily identified. Hence, when two individuals interact, both immediately know which marker the other has. All parents know the share of the population with each marker and the likelihood their children will interact with an individual with the same marker versus a different marker. I refer to all individuals with the same marker as a child as that child's group and denote by group W and group B, the set of all individuals with those respective markers.[14]

In the baseline model, the game proceeds in two stages. In the second stage, each child is matched with another child; these children then decide whether to exert high or low effort competing with their match.[15] Let $c_i \sim F(c)$ represent the cost of exerting high effort with full support on $[\underline{c}, \bar{c}] \subset \mathbb{R}_+$ and I normalize the cost of low effort to zero. Exerting high effort serves two purposes. First, separately from the opponent's choice, high effort raises the probability of receiving a benefit of known value V by π. Hence, I interpret the ratio π/c_i as the child's intrinsic ability, where the component c_i is an idiosyncratic component privately observed by the child when making their effort choice. I assume that while intrinsic ability varies idiosyncratically across individuals in the population, it does not vary systematically between groups:

$$F(c|W) = F(c|B) \tag{1}$$

Hence, group membership is not an intrinsically useful category for predicting an individual's ability. Second, if only one competitor exerts high effort while the other does not, the child who exerts high effort receives b_1, while the child who does not pays b_2 with $b_1, b_2 > 0$. To distinguish this from V, I interpret V as the return to effort in absolute terms, while b_1 and b_2 represent the return on effort relative to others in society, such as tournament-style

[14] Although I use race as my motivating example, the results are equally applicable to other social groupings, where individuals build social meaning upon intrinsically irrelevant differences, such as gender or nationality, as well as to organizations such as firms, armies, or political parties.

[15] I consider cooperative interactions in Section 5.1.

pay (Lazear and Rosen, 1981), receipt of status positions, social approval, and internal self-satisfaction. Finally, I assume that the density $f(c)$ is single-peaked and bounded above by $1/|b_1 - b_2|$. In particular, I assume $f(c)$ has no mass points.

However, when deciding whether to exert high effort, children do not know the true value of π. Instead, they must rely on beliefs imparted to them by their parents. Additionally, in contrast to the true structure of the environment, children treat group membership as potentially relevant to ability, allowing parents to impart beliefs that π varies between groups. When this occurs, children hold 'racial' beliefs, where the physical differences that identify the groups are seen as signals of deeper differences in ability.[16] Although I model belief formation through cultural transmission, the results are equally applicable to a 'neoclassical' principal-agent setting and also to a 'psychological' model of self-deception. Furthermore, Hughes et al. (2006), Ritterhouse (2006), and DuRocher (2011) demonstrate that parental socialization plays a critical role in developing racial stereotypes for both White and Black children. For example, the latter two authors highlight segregated schools and public lynchings as institutions that enabled Whites in the Jim Crow South to instruct their children in racist beliefs and practices.[17]

In the first stage, parents choose beliefs to transmit to their children. Specifically, the parent chooses three objects: a belief about their own group's ability, $\tilde{\pi}^s$, a belief about the other group's ability, $\tilde{\pi}^o$, and a hierarchy of higher-order beliefs. As I define formally in Section 3.1, a hierarchy of higher-order beliefs is a sequence of beliefs $n = 2, 3...$, where the n^{th}-order beliefs are that child's beliefs about the other group's $(n-1)^{th}$-order beliefs. These beliefs answer the questions, "How do they see us?" and "How do they think we see them?" In the baseline model, I assume that parents are fully informed about the returns to effort in society; parents know π and that there are no systematic differences in ability by group.

[16]Morning (2009) shows that a significant majority of university students understand group differences as indicating meaningful underlying differences; Williams and Eberhardt (2008) find that individuals who believe in biological racial differences are more likely to endorse negative stereotypes about racial outgroups.

[17]As DuRocher documents, parents frequently brought their children to lynchings as a community event. Young boys were also encouraged to take part in the lynching as they entered their youth.

Furthermore, I assume that parents are fully convincing when they transmit beliefs to their children, so the child's beliefs are exactly those which the parent chose, and abstract from explicitly modeling the signalling process between parent and child.

When parents choose beliefs to transmit, there are three important sources of friction. First, parents cannot observe the child's idiosyncratic cost c_i when choosing beliefs. Although this is a strong assumption, the motivation is that parents must begin enculturating their children early in life and it is difficult to drastically change social messaging later on. This implies that all parents from the same group transmit the same beliefs to their child. Relatedly, parents do not know the identity of their child's future match. As a result, stereotypes arise to convey usable information; rather than distorting beliefs about individuals, parents provide distorted information about the groups their child could encounter. Second, since children believe that ability varies across groups, they treat information about opponents from their own group as informative about their own ability and vice versa. As children may face within-group matches, this prevents parents from making their children too overconfident about their own ability because they also believe their same-group opponents have equally high ability. Thus, when beliefs about the child's own group are pinned down by within-group compoetition, beliefs about the other group serve as an additional degree of freedom to encourage effort in between-group matches.

Third, the parents' utility function is not perfectly aligned with the child's. I assume that the child's action produces a social externality, so that the benefits, V, b_1, and b_2 are multiplied by $(1+\theta) > 1$ from the parent's perspective, while the cost, c, is not, as in Dessi (2008) and Adriani and Sonderegger (2012).[18] For example, it could represent the returns to a public good, parental pride in their children's accomplishments, or a desire to maintain social arrangements that benefit the parents, but don't directly benefit the children. Importantly, parents seek to internalize this externality, which implies that for some realizations of

[18]See also Adriani and Sonderegger (2009) who study an informational externality in cultural transmission. Two alternative interpretations of θ are (i) imperfect altruism of parents towards their children as in Bisin and Verdier (2001) and Tabellini (2008), and (ii) limited willpower as in Bénabou and Tirole (2006). My results are qualitatively unchanged if I allow θ to vary between groups.

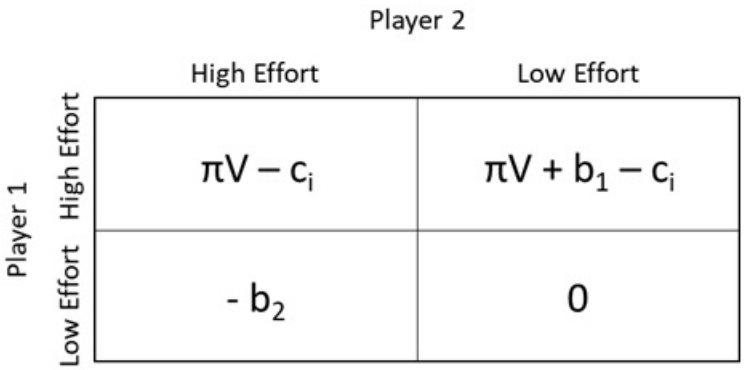

Figure 1: Payoff Matrix for Player 1 (Child)

ability, the parent would want the child to act while the child herself would not. To illustrate how the child and parent's payoffs differ, suppose the child decides to exert high effort and her opponent does not. The child's realized payoff is:

$$\pi V + b_1 - c_i$$

while the parent values this outcome at:

$$(1 + \theta)(\pi V + b_1) - c_i$$

I summarize the payoffs to a child for each possible combination of choices by herself and the opponent with which she's matched in Figure 1; the payoffs to her opponent are symmetric.

Suppose, for illustrative purposes, that costs were common knowledge. From Figure 1, it is straightforward to observe that if $b_1 > b_2$, the strategy pairs (High Effort, Low Effort) and (Low Effort, High Effort) are both pure strategy Nash Equilibria, like in the classic 'Game of Chicken' or 'Hawk-Dove' game. Likewise, when $b_2 > b_1$, the points (High Effort, High Effort) and (Low Effort, Low Effort) are both Nash equilibria, like in the classic coordination game. Hence, the case where $b_1 > b_2$ exhibits a form of strategic substitutability, while the other case exhibits strategic complementarities. Given my focus on intergroup competition, I focus primarily on the case where $b_1 > b_2$, which I call 'the competitive setting,' as opposed

to $b_2 > b_1$, which I call 'the cooperative setting,' discussed in Section 5.1. Finally, I provide an analysis of 'the non-strategic setting,' where $b_1 = b_2$ in Section 4.4.

I now formally define an equilibrium in this setting. For every child, strategies must be optimal given their beliefs about their match. Additionally, parents' choice of beliefs must be optimal given the beliefs chosen by other parents and the matching structure of society. Finally, I impose a symmetry requirement: all parents of the same group choose identical strategies. In what follows, I denote by U_c the child's utility and U_p the parent's utility.

Definition 1. An equilibrium is a set of strategies chosen by children, a_{xy}, where x indicates the child's group and y indicates the opponent's group $(x, y = W, B)$; a set of strategies that children of group x believe that children of group y follow, $\tilde{a}_{xy}$; and structures of beliefs, both first-order and higher-order, $\tilde{\pi}_i$, $i = 1, 2$ chosen by parents, such that:

1. For each child, her actual strategy is a best response to the strategy she believes her opponent follows:

$$\forall a',\ U_c(a_{xy}, \tilde{a}_{yx}) \geq U_c(a', \tilde{a}_{yx}) \quad x, y = W, B.$$

2. For each parent, the structure of beliefs is optimal given beliefs chosen by other parents and strategies chosen by children under those beliefs:

$$\forall \tilde{\pi}_i',\ U_p(\tilde{\pi}_i, \tilde{\pi}_{-i}) \geq U_p(\tilde{\pi}_i', \tilde{\pi}_{-i}) \quad i = 1, 2.$$

3. Beliefs chosen by parents depend only on their group. Let i and j be two parents and G_x denote the group of x:

$$G_i = G_j \implies \tilde{\pi}_i = \tilde{\pi}_j \quad \forall i\, \forall j$$

This symmetry requirement arises since parents from the same group face an ex-ante

identical decision because they cannot observe their child's idiosyncratic ability when choosing beliefs. Consider an individual parent from a group G and hold fixed the beliefs chosen by all other members of that group. Suppose that this parent would like to deviate and transmit a different set of beliefs to their child. Since the parents are ex-ante identical, the other parents from group G would also choose to deviate. However, if this is true, the initial set of beliefs cannot be an equilibrium. Thus, in equilibrium, each parent's best response must be to choose the same beliefs as the other parents of their group.[19]

3.1 Stereotype Awareness and Belief Projection

Before proceeding, I formally define a hierarchy of beliefs. The underlying state is the systematic component of ability, π, which does not truly vary between groups. However, parents can impart beliefs that there are systematic differences in ability by group. The agent i's first-order belief is a probability distribution $\Delta(\pi^W, \pi^B)$, and the agent's second-order beliefs are her beliefs about her opponent's first-order beliefs: $\Delta_i(\Delta_j(\pi^W, \pi^B))$. Let the agent's n^{th}-order belief be written compactly as $\Delta_{i,n}(\pi^W, \pi^B)$. Then, given each player's first-order beliefs, I define higher-order beliefs recursively:

$$\Delta_{i,n+1}(\pi^W, \pi^B) = \Delta_i(\Delta_{j,n}(\pi^W, \pi^B)) \quad \forall n = 1, 2...$$

Then, a hierarchy of beliefs for an agent i is a sequence of n^{th}-order beliefs for all n: $\{\Delta_{i,n}(\pi^W, \pi^B)\}_{n=1}^{\infty}$. For simplicity, I assume that agents hold point beliefs at each level of the hierarchy, i.e., $\Delta_{i,n}(\pi^W, \pi^B) = (\tilde{\pi}^W_{i,n}, \tilde{\pi}^B_{i,n})\ \forall n$.

As stated above, I examine two biases in higher-order beliefs: stereotype awareness and belief projection. First, when an agent projects her beliefs, she incorrectly believes that her opponent shares her first-order belief about her opponent's group's ability. Following the evidence, I then extend belief projection to higher-order beliefs in the following way: At the

[19] I show in the Appendix that mixed-strategy equilibria for parents or children do not exist in my setting.

$(n+1)^{th}$ level of the hierarchy, she incorrectly believes that her opponent shares her n^{th}-order belief about his group. Hence:

Definition 2. Player 1 is subject to belief projection if her beliefs about group G_2 are such that:

$$\tilde{\pi}_{1,n}^{G_2} = \tilde{\pi}_{1,1}^{G_2} \quad \forall n = 1, 2...$$

Likewise, Player 2 is subject to belief projection if his beliefs about group G_1 satisfy:

$$\tilde{\pi}_{2,n}^{G_1} = \tilde{\pi}_{2,1}^{G_1} \quad \forall n = 1, 2...$$

As such, at each step in the hierarchy, a player who projects their beliefs thinks that both players agree about the other player's ability. However, this is not necessarily true, which is the false consensus effect arising from belief projection. If an agent is not subject to belief projection, I assume that they hold correct higher-order beliefs about their opponent's self-image:

$$\tilde{\pi}_{1,n+1}^{G_2} = \tilde{\pi}_{2,n}^{G_2} \quad \forall n = 1, 2...$$

An agent is subject to stereotype awareness when they hold incorrect first-order beliefs about their own group and the other player's group, but hold correct higher-order beliefs about the other player's perception of the agent's group. In particular, if her beliefs about herself differ from how her opponent sees her, she perceives him as holding a biased view of her, while thinking her own self-image is correct. Furthermore, as awareness extends to higher-order belief distortions, I assume that if Player 1 is subject to stereotype awareness:

$$\tilde{\pi}_{1,n+1}^{G_1} = \tilde{\pi}_{2,n}^{G_1} \quad \forall n = 1, 2...$$

When an individual has stereotype awareness, her higher-order beliefs depend on the biases she thinks her opponent holds; in particular, whether his beliefs about her group are subject

to belief projection or not. In equilibrium, since her opponent's identity indicates the belief biases he holds, her perception is correct, even as she thinks her own beliefs are unbiased. Applying the equations above and iterating for each step in the hierarchy, I obtain the following two definitions:

Definition 3. When Player 1 believes Player 2 is subject to belief projection, Player 1 has stereotype awareness if her belief hierarchy satisfies:

$$\tilde{\pi}_{1,n}^{G_1} = \begin{cases} \tilde{\pi}_{1,1}^{G_1} & \text{if } n = 1 \\ \tilde{\pi}_{2,1}^{G_1} & \text{if } n > 1. \end{cases}$$

If instead, Player 1 believes that Player 2 is not subject to belief projection, Player 1 has stereotype awareness if her belief hierarchy satisfies:

$$\tilde{\pi}_{1,n}^{G_1} = \begin{cases} \tilde{\pi}_{1,1}^{G_1} & \text{if } n \text{ is odd} \\ \tilde{\pi}_{2,1}^{G_1} & \text{if } n \text{ is even.} \end{cases}$$

Both when her opponent projects and when he does not, the agent correctly realizes that her opponent's beliefs are biased, but fails to realize that her own are as well. That means that an individual is aware that her opponent is stereotyping her, while potentially remaining unaware that she is stereotyping him in return, which is a form of the bias blind spot and illusion of objectivity. If Player 1 is not aware, she fails to recognize stereotypes that her opponent may hold of her, so I define:

Definition 4. Player 1 has stereotype unawareness if her belief hierarchy satisfies:

$$\tilde{\pi}_{1,n}^{G_1} = \tilde{\pi}_{1,1}^{G_1}.$$

Since both correct beliefs and stereotype awareness imply correct higher-order beliefs about the agent's own group, I distinguish the two as follows: if an agent holds correct

beliefs about her own ability and correctly perceives her opponent's beliefs, I will say she has correct beliefs; if an agent holds distorted beliefs about her own ability, but correctly perceives her opponent's beliefs about her, I will say she has stereotype awareness.

Hence, an agent can be subject to two distinct belief biases: stereotype awareness / unawareness, which determine how she thinks outgroups view her, and belief projection, which determines how she thinks outgroups view themselves. Furthermore, if an agent has stereotype awareness, her beliefs also depend on whether she thinks her opponent is subject to belief projection or not. Finally, since higher-order beliefs depend directly on the first-order beliefs, I suppress the n subscript in what follows.

3.2 Competition Between Children

Since competition proceeds in two stages, I solve backwards, starting with competition between matched children and then address parents' optimal choice of beliefs. Consider a child (Player 1) facing competition with their match, given the beliefs she inherited from her parent and let $b_1 > b_2$ (i.e., the competitive setting). The child's optimal choice depends both on her cost realization and what she believes about her opponent. When $c_i < \tilde{c}_L^{G_1} = \tilde{\pi}^{G_1} V + b_2$, the child wants to exert high effort irrespective of her opponent's choice. Likewise, for $c_i > \tilde{c}_U^{G_1} = \tilde{\pi}^{G_1} V + b_1$, the child's dominant strategy is to exert low effort. I assume that $\underline{c} < b_2 < V + b_1 < \bar{c}$, so there always exist some agents with dominant strategies of low and high effort, irrespective of beliefs. In the interim region, $c_i \in [\tilde{c}_L^{G_1}, \tilde{c}_U^{G_1}]$, the child's optimal strategy depends on whether Player 2 exerts high effort. Importantly, these boundary points depend on the child's first-order beliefs about her own group's ability, $\tilde{\pi}^{G_1}$, but not on her beliefs about the other group's ability, nor on her higher-order beliefs, because these boundaries define the points at which Player 1 has a dominant strategy and thus does not need to consider Player 2's behaviour.

When c_i are children's private information, I can show that the optimal strategy for each child is a threshold rule. To see why, suppose that Player 1 draws a cost in the interim

region (i.e., $c_i \in [\tilde{c}_L^{G_1}, \tilde{c}_U^{G_1}]$) and suppose that given her beliefs about Player 2's likelihood of exerting high effort, Player 1 prefers high effort. Then, it follows that Player 1 also prefers high effort for any lower cost. Likewise, if she prefers low effort for her cost, then she also prefers low effort at any higher cost.

Lemma 1. *There exists a threshold $c^*(\tilde{\pi})$, which depends on the child's first-order and higher-order beliefs, $\tilde{\pi}$, such that a child exerts high effort when matched another child if and only if $c \leq c^*(\tilde{\pi})$.*

All proofs are in the Appendix. Furthermore, I can characterize this threshold directly in terms of the parents' choice of beliefs, Player 1's perception of Player 2's strategy, and model primitives.

Lemma 2. *Let $P(H|\tilde{\pi})$ denote Player 1's subjective probability that Player 2 chooses high effort according to Player 1's hierarchy of beliefs. The optimal threshold for Player 1 satisfies:*

$$c_1^*(\tilde{\pi}) = \tilde{\pi}^{G_1} V + b_2 P(H|\tilde{\pi}) + b_1(1 - P(H|\tilde{\pi})) \tag{2}$$

where $\tilde{c}_L^{G_1} < c_1^*(\tilde{\pi}) < \tilde{c}_U^{G_1}$.

The optimal threshold is a weighted average of the upper and lower thresholds, $\tilde{c}_U^{G_1}$ and $\tilde{c}_L^{G_1}$, where the weight on the lower threshold equals the perceived probability that Player 1 thinks Player 2 will choose high effort. This threshold rule is the player's best response to any strategy she thinks Player 2 follows. Since the same results hold for Player 2, both players follow threshold rules. As any perceived threshold rule $\tilde{c}$ defines a perceived probability of acting, $F(\tilde{c})$, I can define a player's optimal decision rule in terms of the threshold she believes her opponent follows using (2). Let $\tilde{c}_{j,n}^*$ denote Player 1's n^{th}-order conjecture of Player j's strategy. Then:

$$c_1^*(\tilde{\pi}) = \tilde{\pi}^{G_1} V + b_2 F(\tilde{c}_{2,1}^*) + b_1(1 - F(\tilde{c}_{2,1}^*)) \tag{3}$$

Furthermore, as players share common knowledge of rationality, the threshold that Player 1 believes Player 2 follows must be a best response to the threshold that Player 1 believes Player 2 thinks she follows. Thus $\tilde{c}^*_{2,1}$ must satisfy:

$$\tilde{c}^*_{2,1} = \tilde{\pi}_1^{G_2} V + b_2 F(\tilde{c}^*_{1,2}) + b_1(1 - F(\tilde{c}^*_{1,2}))$$

In turn, this second-order conjecture must be a best response to the third-order conjecture and so on. That is:

$$\tilde{c}_{j,n} = \tilde{\pi}^{G_j} V + b_2 F(\tilde{c}^*_{i,n+1}) + b_1(1 - F(\tilde{c}^*_{i,n+1}))$$

Hence, a strategy must be consistent with the full hierarchy of conjectured threshold strategies. To pin down the agent's optimal threshold, I must specify which belief biases she holds, which determine her hierarchy of higher-order beliefs.[20] Therefore, a Nash Equilibrium between children is a pair of thresholds, one for each child, and perceived thresholds, to which each child best responds. Existence and uniqueness follows from a similar argument to the global games literature (Morris and Shin, 2000). Under full information, this second stage could have multiple equilibria; however, the added randomness from idiosyncratic ability allows me to define a unique threshold.

As a benchmark, suppose that π was common knowledge. Since costs are private knowledge, an analogous argument to Lemma 1 proves that the optimal strategy remains a threshold. Then, given that $F(c)$ satisfies the assumptions above, I can show that the unique equilibrium under common knowledge is symmetric.[21]

Lemma 3. *Suppose that $f(c)$ satisfies $f(c) < 1/|b_1 - b_2|$ for all c and is single-peaked. Then, there is a unique equilibrium under common knowledge of π where all agents follow*

[20] Formally, there is a surjective function from the space of hierarchies of beliefs, a subset of the universal type space as in Mertens and Zamir (1985) and Ely and Pęski (2006), onto the set of rationalizable threshold strategies.

[21] I examine the case

the threshold c_0 defined by:

$$c_0 = \pi V + b_2 F(c_0) + b_1(1 - F(c_0))$$

for all matches, irrespective of the identity of their match.

In this equilibrium, group identity is irrelevant to behaviour as both group B and group W children choose the same strategy for both within-group and between-group matches. Hence, although discrimination could arise when children lack information about ability, this benchmark equilibrium shows that if group's average ability was common knowledge, discrimination would not occur.

To understand how higher-order beliefs shift the threshold rule when children do not know π, consider Player 1's optimal threshold. Since $b_1 > b_2$, this threshold decreases with Player 2's optimal threshold, consistent with strategic substitutability. Therefore, the effects of higher-order belief distortions on the strategies that players choose depend on whether those belief distortions lead them to think their opponents are more likely or less likely to exert high effort. Suppose that Player 1 holds a pessimistic belief about group B's ability. When Player 1 projects her beliefs, she believes Player 2 is unlikely to exert high effort because she thinks he shares her pessimism about his group. By (2), this encourages Player 1 to exert high effort because she believes her opponent will acquiesce to her treatment of him. Now consider Player 2's response if he is aware of Player 1's pessimistic belief. This alone is not enough to discourage him because if he thinks that Player 1 does not project her beliefs, he will not expect Player 1 to act on her stereotype of him. By contrast, if he thinks that Player 1 is subject to belief projection, then he expects her to act as though the stereotype was commonly held and choose a higher threshold. When this is true, his best response is to choose a lower threshold.

Proposition 1. *(Stereotype Threat & Stereotype Lift) Suppose that $\tilde{\pi}_2^{G_2} > \tilde{\pi}_1^{G_2}$. Then:*

1. *$c_1^*(\tilde{\pi})$ increases when player 1 projects her beliefs, irrespective of player 2's higher-order*

beliefs.

2. $c_2^*(\tilde{\pi})$ *decreases when player 2 is aware of stereotypes and player 1 projects her beliefs.*

3. *If (i) the players are from different social groups, (ii) player 1 projects her beliefs, and (iii) player 2 is aware of stereotypes:*

$$\frac{\partial c_1^*(\tilde{\pi})}{\partial \tilde{\pi}_1^{G_2}} < 0 \quad \frac{\partial c_2^*(\tilde{\pi})}{\partial \tilde{\pi}_1^{G_2}} > 0$$

This proposition demonstrates two effects of stereotyping, both of which increase in strength as one group's stereotype of the other gets worse:

- **Stereotype Lift:** Player 1's threshold increases if she holds a negative view of group B's ability and projects her beliefs, so she thinks Player 2 shares her negative opinion.

- **Stereotype Threat:** Player 2's threshold decreases exactly when his opponent holds a negative view of his ability, Player 1 projects her beliefs, and Player 2 has stereotype awareness, so Player 2 perceives that his opponent thinks the stereotype is widely held.

Importantly, the presence and intensity of stereotype threat and stereotype lift depends on relative overconfidence and stereotyping, comparing group W's belief about group B's ability, $\tilde{\pi}_W^B$, to group B's self-image, $\tilde{\pi}^B$, rather than absolute stereotyping, comparing it to the truth, π. Hence, it is possible for stereotype threat and stereotype lift to be present even when members of group W overestimate group B's ability relative to the truth, provided they hold less optimistic views than group B does about themselves. Additionally, a player experiencing stereotype lift is not sufficient to conclude that her opponent experiences stereotype threat. Specifically, if Player 1 projects her beliefs, but Player 2 is not aware of her stereotypes, then Player 1 experiences stereotype lift, even though Player 2 does not face stereotype threat. Hence, the mere perception that Player 2 is dissuaded from high effort suffices to increase Player 1's threshold, even if Player 2 is not actually affected. However,

stereotype lift is a necessary condition for stereotype threat; if a player experiences stereotype threat, then it must be the case that her opponent experiences stereotype lift.

Here, stereotype threat and stereotype lift arise due to the interaction between distorted first-order and higher-order beliefs. Player 1 experiences stereotype lift because she holds a negative stereotype about group B and also believes that her opponent shares that stereotype. Likewise, Player 2 experiences stereotype threat because he recognizes that Player 1 holds a negative stereotype of him, even if he does not agree with that stereotype, because he recognizes that Player 1 will act on her beliefs and optimally responds to her treatment of him. Additionally, these effects are only present because beliefs do not truly coincide. If Player 2 held the same pessimistic beliefs about himself as Player 1 holds about him, then it would not matter whether he recognized Player 1's beliefs as his own beliefs would already lead him to choose a low threshold. Similarly, if Player 1 was equally optimistic about group B as Player 2, then projecting her beliefs would not change either player's strategy. As such, differences in first-order beliefs are critical for generating a meaningful interaction between distorted first and higher-order beliefs.

4 Endogenous Stereotypes

In the previous section, players' equilibrium strategies depend on the first and higher-order beliefs that their parents transmit to them. However, each parent's choice is determined in equilibrium, taking as given the beliefs chosen by all other parents from both groups. Specifically, each parent chooses first order beliefs $(\tilde{\pi}_i^W, \tilde{\pi}_i^B)$, whether their child will have stereotype awareness, and whether their child will project their beliefs, where the latter two decisions determine the child's higher-order beliefs.

Critically, I assume parents must choose beliefs for their child before the identity of the child's match is revealed. As a result, 'racial' beliefs can emerge when this limitation interacts with children's perceptions of the world. Children perceive groups as potentially

relevant for understanding ability; they think they share the common component of ability, π, with all members of their social group, but think that this may differ for those not from their group. This cognitive frame allows for 'racial' beliefs since parents can disparage the outgroup without reducing their own child's motivation. Furthermore, since parents do not know the exact identity of their child's match, they must prepare their child for both within-group and between-group matches.

As Lemma 2 implies that each child follows the threshold given by:

$$c^*(\tilde{\pi}) = \tilde{\pi}^G V + b_2 P(H|\tilde{\pi}) + b_1(1 - P(H|\tilde{\pi})) \tag{4}$$

parents can influence their child's strategy either by directly distorting beliefs about their group's ability, $\tilde{\pi}^G$, or by distorting their child's belief about her opponent's strategy, $P(H|\tilde{\pi})$. As discussed, by choosing a specific hierarchy of beliefs, the parent can induce a specific threshold strategy by their child. Since a child views all members of the same group identically, she will follow the same strategy for any individual that she matches within that group. However, she may view the groups differently and follow different strategies depending on the group identity of her match. Thus, her parent must induce two threshold strategies: one for within-group competition and one for between-group competition.

Let c_{G_2} represent the strategy used by Player 2's group in equilibrium. Then, by the same argument as Lemma 2, the parent wants to induce the threshold:

$$c_1 = (1 + \theta)[\pi V + b_2 F(c_{G_2}) + b_1(1 - F(c_{G_2}))] \tag{5}$$

In equilibrium, c_{G_2} must be best response to the strategy chosen by the rest of Player 1's group, c_{G_1}, which implies:

$$c_{G_2} = (1 + \theta)[\pi V + b_2 F(c_{G_1}) + b_1(1 - F(c_{G_1}))] \tag{6}$$

Let the function $G(c_{G_1})$ be defined by substituting (6) into (5). This function defines the threshold that Player 1's parent wants to induce, given that the rest of Player 1's group induces c_{G_1}. Any potential equilibrium is thus a fixed point of this function, i.e.,, $c_{G_1} = G(c_{G_1})$. At a fixed point, when the rest of Player 1's group is following the strategy c_{G_1}, Player 1's parent wants her to do the same. For any c_{G_1} that is not a fixed point of this function, Player 1's parent would want to deviate and so, it cannot be an equilibrium. I show in Appendix A that at least one fixed point exists but is not necessarily unique. However, I can show that there exists a unique fixed point, c_S, which is best response to itself and thus defines the unique symmetric equilibrium:

$$c_S = (1+\theta)[\pi V + b_2 F(c_S) + b_1(1 - F(c_S))] \tag{7}$$

Intuitively, c_S is unique because actions are strategic substitutes. If the opposing group followed a threshold strategy higher than c_S, the best response would be a threshold lower than c_S. Hence, there can be at most one threshold that is best response to itself. Additionally, the symmetric equilibrium involves more effort than the benchmark equilibrium: $c_S > c_0$, where c_0 is defined as in Lemma 3. Since all parents value effort more highly than due their children due to θ, they want to induce their children to choose higher thresholds and be more likely to choose high effort.

I make the following assumption on the cost distribution, $F(c)$:

Assumption 1. $(1+\theta)(b_1 - b_2)f(c_S) > 1$

This assumption summarizes the three forces that affect parents' willingness to distort their children's beliefs: the size of the externality, θ, the strength of competitive incentives, $b_1 - b_2$, and the sensitivity of a player's probability of high effort to their threshold rule, $f(c_S)$. Collectively, these determine how sensitive a parent's desired threshold is to the threshold strategy chosen by other parents. When this is sufficiently low, parents will not have enough incentive to stereotype because strategies respond too weakly. However, when the incentive

is sufficiently strong, there can exist asymmetric equilibria, which I call 'group-favoured' equilibria where the 'favoured' group chooses a high threshold and the 'unfavoured' group chooses a low threshold.

Lemma 4. *There exists a unique symmetric equilibrium, c_S, defined by:*

$$c_S = (1+\theta)[\pi V + b_2 F(c_S) + b_1(1 - F(c_S))]$$

Suppose Assumption 1 holds. Then, there are also two asymmetric equilibria with strategies:

$$c_U = (1+\theta)[\pi V + b_2 F(c_L) + b_1(1 - F(c_L))]$$
$$c_L = (1+\theta)[\pi V + b_2 F(c_U) + b_1(1 - F(c_U))]$$

where $c_L < c_S < c_U$. In these equilibria, one group follows c_L and the other group follows c_U. If Assumption 1 is reversed, the unique equilibrium is c_S.

4.1 Within-Group Competition

I start by examining what happens when a child competes with a member of their own group. Since children believe that they have the same ability, $\tilde{\pi}^G$, as other members of their group, if their parents make them more confident about their ability, they also believe their opponents have higher ability, which dampens the motivational effect of overconfident beliefs. In equilibrium, children from the same group hold the same beliefs, which implies by Lemma 2 that they follow the same threshold strategy. I construct a restricted version of the function $G(c_{G_1})$, using (5) and imposing that $c_{G_1} = c_{G_2}$. Hence:

$$c_1 = G(c_{G_1} | c_{G_1} = c_{G_2}) = (1+\theta)[\pi V + b_2 F(c_{G_1}) + b_1(1 - F(c_{G_1}))]$$

Since this function is decreasing in c_{G_1}, consistent with strategic substitutes, it is straightforward to observe that there is a unique solution, which is the symmetric equilibrium c_S,

defined by (7).

Since the unique equilibrium is c_S, parents must choose beliefs that induce this threshold strategy. To identify these beliefs, observe that since group members hold the same beliefs in equilibrium, higher-order beliefs are constrained to be correct, meaning that neither stereotype lift, nor stereotype threat, are present. Furthermore, since higher-order beliefs are correct, children correctly anticipate the probability that their opponent chooses high effort. Hence:

$$c_s = \tilde{\pi}_i^{G_i} V + b_2 F(c_S) + b_1(1 - F(c_S)) \tag{8}$$

(7) and (8) pin down a unique level of overconfidence in the child's own group:

$$\tilde{\pi}_i^{G_i} = (1+\theta)\pi + \theta \frac{b_2 F(c_S) + b_1(1 - F(c_S))}{V} > \pi \tag{9}$$

- **Overconfidence:** As in other models of motivated reasoning, such as Bénabou and Tirole (2002), overconfidence in one's own ability serves to increase effort, which here serves to internalize the externality.

To show that this is an equilibrium, suppose that a parent deviated and made their child more confident about their ability so that the child would follow a higher threshold. However, since their child treats information about their own ability as informative about other members of their group, they also think every child in their group has higher ability. When this occurs, they incorrectly think that others members of their group follow a higher threshold, which reduces their motivation to choose high effort. Hence, the unique symmetric equilibrium is stable for within-group matches because children hold a 'racial' belief about their own group: that all children in their group share a common component to their ability, $\tilde{\pi}^G$. Thus, within-group matches pin down the optimal level of self-confidence, but do not determine the parent's choice of beliefs about the other group, nor the choice of higher-order beliefs.

Theorem 1. *There is a unique symmetric equilibrium for within-group matches, where parents in group G choose $\tilde{\pi}^G$ according to (9) and children follow the threshold strategy defined*

implicitly by (7).

In equilibrium, the child and her opponent hold the same beliefs, which implies by Proposition 1 that her strategy is not affected by higher-order belief distortions. Without differences in first-order beliefs, parents lack an incentive to distort higher-order beliefs. As such, within-group competition does not determine distorted higher-order beliefs.

4.2 Between-Group Competition

Consider now a match between two child from different social groups. Unlike when a child interacts with a member of her own group, the child may believe that her opponent has a different level of ability and follows a different threshold strategy than she does. If the parent could only distort her child's self-image, the parent would have to trade off suboptimal behaviour when competing within group versus between groups, depending on the relative likelihood of each. However, the parent can instead distort the child's belief about the other group and distort her higher-order beliefs, which gives the parent an extra degree of freedom to influence the child's strategy. Given the threshold in (4), the parent distorts $P(H|\tilde{\pi})$, allowing $\tilde{\pi}^G$ to be pinned down by within-group matches as given by (9). From (5), the strategy that Player 1's parent induces must be best response to the strategy chosen by Player 2's group according to the parent's preferences. In equilibrium, this threshold must be the same threshold induced by all other parents in Player 1's group. However, when Player 1 and Player 2 are from different groups, they do not need to choose the same strategy in an equilibrium. Hence, there are three possible equilibrium strategies that Player 1's group could follow: c_L, c_S, and c_U, depending on what Player 2's group follows.

Suppose first that parents in Player 2's group choose beliefs that induce c_S. Since c_S is best response to itself, Player 1's parent also wants her to choose c_S. As mentioned, Player 1's first-order belief about her own group is pinned down by within-group matches, according to (9). Then, given this self-belief, it follows from (8) that Player 1 will choose c_S when she correctly anticipates that her opponent will also choose c_S. So that Player 1

correctly anticipates Player 2's strategy, her parent must endow her with correct beliefs. Thus, $\tilde{\pi}_1^{G_2} = \tilde{\pi}_2^{G_2}$ and higher-order beliefs are correct. By an identical argument, Player 2's parent will endow him with correct higher-order beliefs about Player 1's group. This shows that in the symmetric equilibrium, stereotyping does not occur; for both groups, their self-image equals the opposing group's image of them. Although all agents are overconfident about their ability, no agent believes that ability differs between the two groups, and so players treat group markers as meaningless. However, if Assumption 1 holds, this equilibrium is unstable in the following sense:

Definition 5. An equilibrium $\hat{c}$ is stable if there exists $\epsilon > 0$ such that, after perturbing the equilibrium strategy of group G_i by ϵ, and iterating on best responses, starting with group G_j for $i \neq j$, strategies converge to $\hat{c}$.

The symmetric equilibrium, c_S, is always stable for within-group matches, but potentially unstable for between-group matches. Suppose that Player 1's group chooses a higher threshold than the equilibrium c_S. In between-group competition, this will lead Player 2's group to respond by adopting a lower threshold. Thus, Player 1's parent will want her to choose a higher threshold herself. If Assumption 1 is satisfied, then Player 1's parent will induce an even higher threshold than the rest of her group and thus, every other parent in that group would also want to deviate to a higher threshold. This will cause Player 2's group to deviate to an even lower threshold. Hence, the equilibrium is unstable. When Assumption 1 fails, Player 1's parent will choose a higher threshold than c_S, but lower than the rest of the group, and so iterated best response will converge back to c_S. By contrast, for within-group matches, if the rest of Player 1's group chooses a higher strategy, Player 1's parent will induce a lower strategy. This difference reflects the difference in how a player thinks about her own group compared to the opposing group. Since she believes that the other group could have a different level of ability, this allows her to expect them to follow a different strategy. As such, shifts in her own group's strategy affect her only indirectly through how they affect the other group's strategy.

Now consider the asymmetric equilibria and without loss, assume that Player 2's group induces threshold c_L (i.e., group 2 is the unfavoured group). Player 1's parent would like her to follow c_U, where these thresholds are defined by:

$$c_U = (1+\theta)[\pi V + b_2 F(c_L) + b_1(1 - F(c_L))] \tag{10}$$

$$c_L = (1+\theta)[\pi V + b_2 F(c_U) + b_1(1 - F(c_U))] \tag{11}$$

To do so, they need to encourage their child to exert more effort as $c_U > c_S$, but cannot do so by giving them more optimistic beliefs about their own ability, as that would cause inefficiently high effort in within-group matches. However, by Proposition 1, Player 1's parent can also distort her effort choice upward by giving her a negative stereotype of Player 2's group and teaching her to project her beliefs. By holding a negative stereotype and thinking that Player 2 agrees with that stereotype, Player 1 experiences stereotype lift and competes more aggressively. Hence, $\tilde{\pi}_1^{G_2} < \tilde{\pi}_2^{G_2}$, so Player 1 engages in relative stereotyping.

Now suppose that all parents in Player 1's group teach these stereotypes and belief projection, so that their children follow c_U. Parents in Player 2's group will want their children to choose c_L, which requires distorting their children's threshold choices downward. As before, parents do not want to impart a different self-image than $\tilde{\pi}_2^{G_2}$ as defined by (9), since that would cause inefficiently low effort in within-group matches. However, since Player 1's group holds negative stereotypes about Player 2's group and projects their beliefs, if Player 2 is aware of those stereotypes, he adopts a less aggressive strategy due to stereotype threat. Thus, stereotype awareness allows Player 2 to respond optimally without harming his motivation for within-group matches, as stereotype threat is only present in between-group matches. However, stereotype threat alone reduces Player 2's threshold by more than his parent would like, so his parent would like to create a compensating force to reduce the

impact of stereotype threat:

$$\tilde{\pi}_2^{G_2} V + b_2 F(c_U) + b_1(1 - F(c_U)) < c_L$$

To do so, Player 2's parents will teach him a slightly negative stereotype about Player 1's group and teach him to project his beliefs. By doing so, they temper the effects of stereotype threat so that his strategy is not distorted downward too much. Finally, since Player 2's group stereotypes Player 1, her parents do not want their child to be aware of these stereotypes so that Player 1 continues to choose c_U.

Hence, there is an asymmetric equilibrium, where parents in the favoured group teach negative stereotypes about the unfavoured group and ensure that their children do not know the stereotypes that the unfavoured group holds about them. Simultaneously, the unfavoured group teaches negative stereotypes, but does teach their children about the stereotypes the favoured group holds about them. Importantly, although both groups hold negative stereotypes about each other, the favoured group holds more negative stereotypes about the unfavoured group than the unfavoured group holds in return:

$$\tilde{\pi}_1^{G_2} < \tilde{\pi}_2^{G_1} < \tilde{\pi}_i^{G_i}$$

Hence, both groups are equally overconfident about their ability, but underestimate the other group's self-image, so both groups engage in relative stereotyping. An identical argument applies when Player 2's group is the favoured group with the labels reversed. Additionally, these asymmetric equilibria are stable according to Definition 5. If Player 1's group deviates to a strategy $c_U + \epsilon$, Player 1's parent will want to induce a strategy $c_1 \in (c_U, c_{U\epsilon})$, since there is not sufficient incentive to stereotype beyond c_U, and the same argument applies for deviations downward and for group 2's deviations from c_L. Thus, the asymmetric equilibria are stable.

Theorem 2. *Suppose Assumption 1 holds. There exist two stable asymmetric equilibria,*

a W-favoured equilibrium and a B-favoured equilibrium, and an unstable symmetric equilibrium.

In the W-favoured equilibrium (B-favoured equilibrium), parents from both groups teach negative stereotypes, with group W holding more negative stereotypes: $\tilde{\pi}_W^B < \tilde{\pi}_B^W < \tilde{\pi}_G^G$. *Both groups teach their children to project their beliefs, but only group B (group W) teach stereotype awareness. Group W (group B) children follow* c_U *and group B (group W) children follow* c_L*, as defined by (10) and (11).*

In the symmetric equilibrium, neither group holds stereotypes about the other group $\tilde{\pi}_B^W = \tilde{\pi}_W^B = \tilde{\pi}^G$ *and children in both groups follow threshold* c_S*, as defined by (7).*

In the asymmetric equilibrium, both groups engage in relative stereotyping. Additionally, by Proposition 1, the more that one group's parents want to distort their children's strategy upwards, the more negative of a stereotype about the opposing group they impart. Consider the W-favoured equilibrium. As the competitive incentives grow stronger, or the externality grows larger, the value of stereotyping increases, so group W parents choose more negative stereotypes about group B. When these incentives are sufficiently strong, group W will hold stereotypes that underestimate group B's ability compared to the truth: $\tilde{\pi}_W^B < \pi$.

Corollary 3. *Given* b_1, b_2*, there exists* $\theta^*(b_1, b_2)$ *such that if* $\theta > \theta^*(b_1, b_2)$, $\tilde{\pi}_W^B < \pi$ *in the W-favoured equilibrium and* $\tilde{\pi}_B^W < \pi$ *in the B-favoured equilibrium. Likewise, there exists* $\Delta^B(\theta)$ *such that if* $b_1 - b_2 > \Delta^B(\theta)$*, then* $\tilde{\pi}_W^B < \pi$ *in the W-favoured equilibrium and* $\tilde{\pi}_B^W < \pi$ *in the B-favoured equilibrium. Finally:*

$$\frac{\partial \tilde{\pi}_W^B}{\partial (b_1 - b_2)} < 0 \quad \frac{\partial \tilde{\pi}_B^W}{\partial (b_1 - b_2)} < 0$$

for $\tilde{\pi}_W^B, \tilde{\pi}_B^W > 0$.

Suppose that the W-favoured equilibrium is played for between-group matches. Agents'

behaviours demonstrate additional features of stereotyping:

- **Asymmetries in Meta-Stereotypes:** Group W falsely believes that the unfavoured group shares their stereotypes and are unaware of how group B views them. By contrast, group B correctly perceives group W's stereotypes.

- **Asymmetric Competitiveness:** Group B competes more readily when matched with members of their own group than when matched across groups as stereotype threat is only present when interacting with a member of the opposing group.

- **Outgroup Favouritism:** Although both groups view themselves more favourably than they view the opposing group, group B holds less negative stereotypes of group W than group W holds of them (Jost and Banaji, 1994; Jost, 1997).

The possibility of within-group matches disciplines groups' self-image for between-group matches. If there were no possibility of within-group matches, parents could implement their optimal strategies merely by making their children more or less self-confident, without using stereotypes or distorting higher-order beliefs. But since there are both within-group and between-group matches, parents choose to use stereotypes and distorted higher-order beliefs to implement optimal strategies in both types of matches.

Finally, in the asymmetric equilibria, children hold 'racial' beliefs that group identity is informative about players' ability, contrary to the truth. Furthermore, individuals' behaviour varies systematically when interacting with a member of their own group compared to when interacting with someone from outside their group, even though individuals express identical beliefs in their own ability across settings. However, children incorrectly believe that ability differs between groups and partially misattribute differences in strategies to differences in ability, similarly to the 'fundamental attribution error' discussed by Chauvin (2019). Additionally, although children correctly recognize that strategies will differ between groups, they mis-estimate the strategy that the other group will follow because their stereotypes lead them to underestimate their opponents. Importantly, even though beliefs about ability and

strategies are incorrect, these inaccurate beliefs influence the strategies that children choose. Comparing the unique benchmark equilibrium to the asymmetric equilibria demonstrates that by inducing distorted beliefs, parents can lead their children to discriminate, even when the children would not want to do so if they were fully informed. For the parents, the possibility of inducing discrimination leads them to teach their children stereotypes and in turn, these stereotypes cause the children to discriminate.

4.3 Equilibrium Selection

In the baseline model, there are two asymmetric equilibria, and as the groups are ex-ante identical, there is no force selecting between these equilibria. However, there are some features of real-world environments that may help select a particular favoured group.

I assume that parents can freely choose whether to make their children aware of stereotypes that others hold about them. However, if one group has greater control over public communications, that group can broadcast their beliefs to both groups, making awareness of this group's stereotypes unavoidable by the other group. If the receiving group must be aware of the broadcasting group's stereotypes, the unique stable equilibrium will be the group-favoured equilibrium favouring the broadcasting group. For example, suppose that one group disproportionately controls popular media production and representations of the opposing group replicate common negative stereotypes that the media-controlling group holds about them. This should support a group-favoured equilibrium which favours the group in charge of the media as the opposing group cannot avoid being aware of the stereotypes about them.[22] Similarly, if one group were in a superior social position, with access to greater political and economic resources, these resources can give them the power to influence the beliefs transmitted by the other group. For example, in the United States, the Southern planter elite stoked anti-Black prejudice against newly freed slaves to protect their social position, whereas poor Blacks had little ability to shape Whites' views of them (Acharya et al., 2016).

[22]See e.g., Entman and Gross (2000), Martins and Harrison (2012), and Ross (2019).

Finally, parents of both groups choose beliefs simultaneously. However, if instead, one group chooses their beliefs before the other, then the unique outcome would be the equilibrium favouring the group that chooses their beliefs first. For example, if one group is the existing citizens of a country and the other group represents new immigrants to this country. Then, the existing citizens would have the opportunity to set their beliefs about themselves and about immigrants before the immigrants arrived. When new immigrants represent competition for the existing citizens over limited resources and status positions, my results imply that existing citizens should respond to this threat by denigrating immigrants without a concomitant tendency for immigrants to denigrate the existing citizens.[23]

4.4 Overconfidence Without Stereotyping

In Heidhues et al. (2020), stereotypes about other groups arise due to misspecified learning when agents attempt to maintain stubbornly overconfident beliefs about their own ability. In their setting, if agents are overconfident, they always develop stereotypes about other groups. By contrast, in my setting, it is possible to observe overconfidence without stereotyping. If the inequality in Assumption 1 is reversed, there is not enough incentive for parents to develop stereotypes of the opposing group as stereotype threat and lift do not sufficiently affect strategies. This is easiest to see in the extreme case of the 'non-strategic' setting $b_1 = b_2 = b$, where Assumption 1 never holds. By Lemma 2, a child's optimal threshold reduces to:

$$c_i^*(\tilde{\pi}) = \tilde{\pi}_i^{G_i} V + b$$

This threshold rule depends only on the child's first-order belief about her own group's average ability and not on her belief about the other group or on her higher-order beliefs. The player ignores her opponent's strategy because in the non-strategic setting, each child always has a strictly dominant strategy conditional on her cost realization c_i, except for the measure zero case at the threshold, where she is indifferent.

[23] See e.g., Stephan et al. (1999), Meuleman et al. (2009), and Matera et al. (2015).

Nevertheless, overconfidence is still potentially present. When $\theta > 0$, parents would like to make their children more optimistic about the return to effort to encourage them to exert high effort (Bénabou and Tirole, 2006). Parents want their children to exert high effort whenever:

$$c_i \leq (1+\theta)(\pi V + b)$$

Hence, parents make their children overconfident about their own group:

$$\tilde{\pi}^G = (1+\theta)\pi + \theta\frac{b}{V} \tag{12}$$

Since Assumption 1 does not hold, the unique between-group equilibrium is a stable symmetric equilibrium with $\tilde{\pi}^W = \tilde{\pi}^B$, defined by (12) and without stereotyping.

5 Extensions and Discussion

5.1 Beliefs Under Strategic Complementarity

The baseline model considers competitive matches; however, some interactions may exhibit strategic complementarities, which I can analyze by assuming that $b_2 > b_1$. This implies that when agents draw interim costs, they wish to choose high effort if and only if their match also chooses high effort. As such, the incentives to distort beliefs about the outgroup work in the reverse direction. Since players adopt higher thresholds when they believe their matches are more likely to choose high effort, parents want their children to adopt optimistic beliefs about the outgroup's ability. Hence, cooperation creates the possibility of positive stereotyping, as has been found by Sherif and Sherif (1953) and Matera et al. (2015). The formal analysis is very similar to the baseline model. In particular, Lemmas 1 and 2 continue to hold, meaning a child's optimal threshold can be described by (3). However, since $b_2 > b_1$, the child's optimal threshold is increasing in the threshold that she believes her opponent follows. Similarly to the classic 'coordination game,' there can exist multiple equilibria and unlike the

competitive setting, equilibria will be symmetric, rather than asymmetric. If there exists a fixed point c_F of best responses such that Assumption 1 holds at c_F, then there are multiple stable symmetric equilibria. When this is true, there is a symmetric low-effort equilibrium and a symmetric high-effort equilibrium for both within and between-group matches.

In within-group matches, parents can use a child's confidence in their own group to encourage higher effort. If c^G denotes the strategy chosen by all other members of a child's group in equilibrium, that child's parent choose to give them self-image:

$$\tilde{\pi}^G = (1+\theta)\pi + \theta\frac{F(c^G)b_2 + (1-F(c^G))b_1}{V} \tag{13}$$

as this belief leads their children to also choose c^G. Hence, children are more overconfident in the high-effort equilibrium, where c^G is greater, than in the low-effort equilibrium. However, even in the low-effort equilibrium, children are overconfident about their own group, $\tilde{\pi}^G > \pi$. Thus, there is a pure coordination failure among parents if the low-effort equilibrium occurs for within-group matches.

In between-group matches, there are likewise multiple stable equilibria when Assumption 1 holds, a symmetric low-effort equilibrium and a symmetric high-effort equilibrium. However, the beliefs in an equilibrium depend both on what equilibrium is played for between-group matches and on what equilibrium is played for within-group matches, as that determines players' optimal self-image. However, I can show that analogous effects to stereotype threat and lift exist in the cooperative setting. Suppose that Player 1 projects her beliefs and Player 2 is aware of Player 1's stereotypes. If Player 1 holds more optimistic beliefs about Player 2, she increases her threshold because she thinks he has higher ability. In response, Player 2 also increases his threshold because he anticipates Player 1's behaviour. As this is true even when Player 1 holds more optimistic beliefs about Player 2 than Player 2 holds about himself, I interpret this as a form of leadership. Player 2's group knows that others expect more of them than they do of themselves, and anticipate that others will choose higher

thresholds based on that belief. In response, they also choose to exert high effort to 'live up to' Player 1's group's expectations. The notion that leaders' expectations can affect their followers' performance has been previously explored in teacher-student relationships (Chauvin, 2019; Papageorge et al., 2020) and for manager-employee relationships (Bolton et al., 2013; Glover et al., 2017); however, I document a novel channel through which distorted higher-order beliefs can serve to improve performance.

Proposition 2. *(Stereotype Boost) Suppose that Player 1 (group W) projects her beliefs and Player 2 (group B) has stereotype awareness. Then, both players' thresholds are increasing in $\tilde{\pi}^B_W$.*

Since the interaction involves cooperation, both players adopt higher thresholds when one group holds a positive stereotype about the other:

- **Stereotype Boost** Player 2's threshold increases when Player 1 holds an optimistic view of his ability, projects her beliefs, and Player 2 is aware of this stereotype (Shih et al., 2013).

Player 2's threshold is higher due to this stereotype boost effect, just as their threshold is lower in the W-favoured competitive equilibrium due to stereotype threat. In both competitive and cooperative interactions, the effect which increases Player 1's threshold are forms of stereotype lift: exerting more effort due to stereotypes about other individuals.

5.2 Reducing Stereotype Threat

Steele (1997) and Spencer et al. (1999) discuss strategies that reduce stereotype threat in individuals subject to it, and these strategies have natural interpretations through the lens of my model. First, they remark that individuals only suffer from stereotype threat when they care about the outcome; individuals who are indifferent to how they perform do not suffer from stereotype threat, but also do not perform well in general. Here, that could be

straightforwardly modeled as an agent for whom $b_1 = b_2 = 0$. By Lemma 2, the agent's optimal threshold would then be:

$$c^*(\tilde{\pi}) = \tilde{\pi}^G V$$

Since this does not depend on the opponent's strategy, the agent following this threshold is immune to stereotype threat and cannot benefit from stereotype lift. Furthermore, their threshold is lower than an agent for whom b_1, $b_2 > 0$, irrespective of whether that agent is subject to stereotype threat, matching the general poor performance of indifferent agents.

Another strategy they discuss is to convince the agent that the challenge they face depends on a non-stereotyped trait rather than a stereotyped one. For example, telling students that a test measures puzzle-solving ability instead of general intelligence tends to reduce stereotype threat. In the context of this model, if there were multiple skills, each may have different degrees to which they are stereotyped. Additionally, by Corollary 3, stereotypes are stronger in high-stakes competitions, those with $b_1 - b_2$ large, compared to low-stakes competitions. Since parents choose beliefs before observing their child's match, a reasonable assumption is that they choose stereotypes for each skill according to the expected stakes, but any individual competition may have higher or lower stakes than the average for that skill. Hence, if puzzle-solving competitions are primarily low stakes, while intelligence competition are primarily high stakes, Corollary 3 implies that intelligence will be more stereotyped than puzzle-solving ability. As such, convincing an individual that a particular competition tests puzzle-solving ability, holding fixed that competition's stakes, will reduce stereotype threat and improve performance. However, while this strategy can reduce individual instances of stereotype threat, it cannot eliminate it at the population level: if sufficiently many puzzle-solving competitions become high stakes, favoured-group parents will optimally respond by stereotyping puzzle-solving skill more strongly. My results also suggest a third way to reduce stereotype threat. First, comparing Proposition 1 to Proposition 2, competitive interactions incentivize negative stereotypes, but cooperative interactions promote positive beliefs. Therefore, if a greater fraction of intergroup interactions

occur in cooperative settings, as opposed to competitive settings, these incentives should push stereotypes to become more positive.

5.3 Exploitation and Avoidance

Another force affecting behaviour is the incentives to direct one's matching. Implicitly in the baseline model, matches are random, so players cannot choose to pursue or avoid matching within or across groups. However, if between-group matches played the W-favoured equilibrium, members of group B would have an incentive to avoid interacting with group W to avoid being exploited, while members of group W would want to seek out intergroup matches. It may be that agents only have partial control over their matches. In this case, the group favoured in a group-favoured equilibrium faces a trade-off between earning exploitative gains, but having the other group try to avoid them, versus a more equitable distribution in between-group matches without the outgroup avoiding them.

It may also be the case that agents face both competitive and cooperative matches, which affects the incentives to match in two ways. Between-group matches may be intrinsically more productive than within-group matches.[24] The possibility of productive cooperation should reduce the temptation to exploit during competitive matches as exploitation will cause the outgroup to exert effort avoiding intergroup contact. However, if the benefits of cooperation to the unfavoured group are sufficiently large, they may choose not to avoid intergroup contact, even though they are exploited in competitive matches. Additionally, the possibility of cooperative matches creates a tension between holding negative views of the unfavoured group, which are beneficial for competitive matches, and holding positive views of them for cooperative matches. In practice, this may lead the favoured group to create arbitrary exceptions, allowing them to hold positive views of the low-status individuals with whom they cooperate, while maintaining negative views of individuals they face in competition.

In my setting, group W would not engage in segregation, reducing the frequency with

[24]See e.g., Phillips et al. (2009) on the benefits to diverse teams for decision-making processes.

which they match with group B. Since the overall share of competitive versus cooperative matches is independent of the share of within-group versus between-group matches, engaging in segregation does not reduce the amount of competition an agent faces. Instead, it only reduces the share of competitive matches that are with the opposing group, where they can earn exploitative gains.[25]

To formalize these intuitions, I consider a deliberately simplified game of within vs. between-group matching. In this game, I assume members of group W choose their beliefs first, breaking the symmetry in the baseline model and allowing them to select the W-favoured equilibrium. In response, members of group B can attempt to avoid intergroup matches at some cost.[26] The game proceeds as follows:

1. Group W parents choose $\tilde{\pi}^B_W$, for brevity I will write as $\tilde{\pi}$.[27]

2. Each group B player chooses an avoidance level a with cost function $c(a)$, so that $\lambda(a)$ fraction of matches are between-group, $1 - \lambda(a)$ are within-group.

3. Children are matched and play either a competitive or cooperative game, where α represents the share that are competitive and $1 - \alpha$ are cooperative.

Since the payoff of within-group matches does not vary with $\tilde{\pi}$, I normalize it to zero. I define the payoffs to between-group matches in reduced-form, letting γ represent a productivity shifter for the relative value of between-group matches compared to within-group matches:

- Let $V^G_{cm}(\tilde{\pi}, \gamma)$ be the value of competitive intergroup matches to group G under the W-favoured equilibrium.

[25] C.f. Lagerlöf (2020) and Dewan and Wolton (2020) where segregation affects the degree of competition agents face, so segregation can improve payoffs for the segregating agents. Additionally, group W agents have no intrinsic dislike for group B agents, which would be an additional reason to segregate.

[26] The results are qualitatively unchanged if group B can specifically avoid competitive matches, provided they cannot perfectly do so.

[27] Importantly, agents are constrained to hold the same beliefs across competitive and cooperative matches; however, as long as beliefs in one type of match partially constrain beliefs for the other type of match, my results are unchanged (i.e., if an agent holds extremely negative views during competition, there is a binding upper limit on how positive their views can be during cooperation).

- Let $V_{cp}^G(\tilde{\pi}, \gamma)$ be the value of cooperative intergroup matches to group G under the W-favoured equilibrium.

I assume that all V_x^G are increasing in γ and over the relevant range of values:

Assumption 2. The following assumptions hold:

1. $\frac{\partial V_{cm}^W}{\partial \tilde{\pi}} < 0$
2. $\frac{\partial V_{cp}^W}{\partial \tilde{\pi}} > 0$.
3. $\frac{\partial V_{cm}^B}{\partial \tilde{\pi}} > 0$
4. $\frac{\partial V_{cp}^B}{\partial \tilde{\pi}} > 0$
5. $\lambda'(a) < 0$
6. $c'(a) > 0$ and $c''(a) > 0$

In particular, the game described in Sections 3-5 satisfies the first four assumptions. The fifth implies avoidance effort reduces the likelihood of between-group matches and the sixth implies that the cost of doing so is convex. As is intuitive, groups' interests diverge in competitive matches and align in cooperative matches. First, suppose that group B cannot avoid matching (i.e., $a \equiv 0$). Then, the optimal belief about group B, $\tilde{\pi}^*$, satisfies:

$$\alpha \frac{\partial V_{cm}^W(\tilde{\pi}^*, \gamma)}{\partial \tilde{\pi}} + (1 - \alpha) \frac{\partial V_{cp}^W(\tilde{\pi}^*, \gamma)}{\partial \tilde{\pi}} = 0 \tag{14}$$

From this condition, it is possible to derive some comparative statics for group W's optimal beliefs about group B. I make the following additional assumptions:

Assumption 3.

$$\frac{\partial^2 V}{\partial \tilde{\pi}^2} < 0 \quad \frac{\partial^2 V_{cm}}{\partial \tilde{\pi} \partial \gamma} \leq 0 \quad \frac{\partial^2 V_{cp}}{\partial \tilde{\pi} \partial \gamma} \geq 0$$

The first is implied by the game analyzed above, while the second and third assumptions imply that as intergroup matches become more valuable, group W chooses to become more exploitative in competitive matches and more encouraging in cooperative matches, and would be satisfied if, for example, an increase in γ increased b_1 and b_2 proportionately.

Lemma 5. *Suppose that $a = 0$ and assume that Assumptions 2 and 3 are satisfied. Then:*

1. $\frac{\partial \tilde{\pi}^*}{\partial \alpha} < 0$

2. There exists $\hat{\alpha}$ such that $\frac{\partial \tilde{\pi}^*}{\partial \gamma} > 0$ if and only if $\alpha < \hat{\alpha}$.

As a greater share of matches are competitive, the incentive to denigrate the unfavoured group and benefit from competition dominates the incentive to encourage them and benefit from cooperation, so optimal beliefs decrease. More interestingly, an increase in the productivity of between-group matches can actually worsen the favoured group's stereotypes if sufficiently many of those matches are competitive. In line with this result, White American's beliefs about the competence of Black Americans worsened significantly during the Reconstruction era, even as American productivity improved, as newly-freed Blacks increasingly competed with Whites for scarce resources, political power, and social status (Roediger, 2008; Acharya et al., 2016; Gross, 2008; Ore, 2019).[28]

Now consider agents in group B's choice to avoid intergroup matches. They choose a to maximize:

$$\lambda(a)[\alpha V_{cm}^B(\tilde{\pi}, \gamma) + (1-\alpha)V_{cp}^B(\tilde{\pi}, \gamma)] - c(a)$$

taking as given group W's beliefs. Since $a \geq 0$, I obtain:

$$\lambda'(a^*)[\alpha V_{cm}^B(\tilde{\pi}, \gamma) + (1-\alpha)V_{cp}^B(\tilde{\pi}, \gamma)] - c'(a^*) \leq 0 \tag{15}$$

where (15) holds with equality if $a^* > 0$. This occurs only when the term in square brackets is negative; naturally, if between-group matches are more valuable on average than within-group matches, agents will make no effort to avoid between-group matches, even though they are exploited in the competitive matches. However, a sufficient share of matches are competitive, then the cost of exploitation dominates the benefit of productive matches. There exists $\bar{\alpha}$, which depends on model parameters, such that (15) holds with equality for $\alpha > \bar{\alpha}$. When this is true, it is straightforward to obtain the comparative statics for a^*.

Lemma 6. *There exists $\bar{\alpha}$ such that (15) binds with equality for $\alpha > \bar{\alpha}$. When this holds,*

[28] In addition to the strategic motivation, after slavery ended, Whites no longer faced an affective motive to believe that slaves were happily enslaved, although this did not directly affect beliefs about competence.

then:

$$\frac{\partial a^*}{\partial \alpha} > 0 \quad \frac{\partial a^*}{\partial \tilde{\pi}} < 0 \quad \frac{\partial a^*}{\partial \gamma} < 0$$

Each of these results is intuitive: as the share of competitive matches increases, group B engages in more avoidance since the likelihood of an exploitative competitive match when they match across groups is higher; as group W's beliefs are more positive, they exploit less and encourage more, so between-group matches are more attractive; and as the productivity of between-group matches rises, group B wants to engage in more between-group matches.

Returning to group W parents' choice, the threat of group B self-isolating limits how exploitative competitive matches can become.[29]

Proposition 3. *(Safe Spaces) Let $\tilde{\pi}_0^*$ be the optimal belief when no avoidance is possible, $a \equiv 0$, and let $\tilde{\pi}_a^*$ be the optimal belief when avoidance is possible. Then:*

$$\tilde{\pi}_0^* \leq \tilde{\pi}_a^*$$

with strict inequality if (15) binds with equality.

The comparative statics on $\tilde{\pi}$ are qualitatively unchanged from before. When most between-group matches are cooperative, (15) does not bind and group B does not try to avoid matches, so the results are clearly unchanged. However, when group B attempts to avoid between-group matches, changes in α or γ have conflicting effects. An increase in the share of competitive matches increases the incentive to hold negative views, but also increases in the incentive for group B to avoid group W. Likewise, if α is sufficiently high, by Lemma 5, an increase in γ raises the incentive for group W to hold negative beliefs, which should increase avoidance, but the direct effect of an increase in γ reduces avoidance, by Lemma 6. Nevertheless, the direct effects described in Lemma 5 dominate the indirect effect via avoidance in Lemma 6. Another result is that group W is made worse off by the

[29]In an 1865 meeting, Black Baptist minister Garrison Frazier, told Union General Sherman that newly emancipated slaves would prefer 'to live by ourselves [due to] a prejudice against us in the South that will take years to get over' (Kendi, 2016).

possibility of avoidance, strictly if (15) binds with equality, while group B is made better off. Although, I do not model this explicitly, this implies that group W would like to make avoidance harder, while group B would like to make it easier.

5.4 Multiple Generations

The baseline model considers the incentives for parents to distort their children's beliefs across a single generation of cultural transmission. Another question is how beliefs might change across multiple generations. In this setting, the answer depends on how much children are able to learn after participating in competition, but before they choose beliefs for their own children, the initial parent's grandchildren.

For one benchmark, parents could become fully informed before teaching their children, learning that the average ability of both groups is π. When this is true, the model is stationary: each generation of parents teaches their children the same distorted beliefs as the previous generation. At each generation, the incentives to distort beliefs are identical, so parents choose the same incorrect beliefs. This implies that stereotypes persist even if all agents are completely Bayesian and each generation unravels the belief distortions.

If instead, parents did not fully unravel how their parents distorted their beliefs, then beliefs will tend to become more polarized across groups over time as each generation of parents wants to make their children overconfident relative to what the parents believe their group's 'true' ability to be. Many of the results in other models in which agents 'misread' evidence ex post in a dynamic setting to maintain beliefs apply in my setting (Bénabou and Tirole, 2004; Ali, 2011; Zimmermann, 2020; Heidhues et al., 2020). Finally, a fully dynamic theory of prejudice and discrimination would need to explore how choices made in the present shape the balance between competitive and cooperative interactions in the future, including how the incentives facing the children may be different than those their parents faced.

6 Conclusion

I present a theory of stereotyping as motivated reasoning, where parents transmit biased beliefs to their children to help them compete more effectively against another social group. As suggested by Bobo (1999), an analysis of racial prejudice should make explicit the incentives to stereotype. By doing so, prejudice becomes amenable to economic analysis on the basis of costs and benefits. Following this logic, I show that empirically observed patterns of stereotyping and higher-order belief biases can arise from purely instrumental motives in a setting of 'discrimination for advantage,' and investigate the implications of these distorted beliefs on within-group and between-group competition. I document how 'stereotype threat' and 'stereotype lift' affect players' behaviour through the interaction between distorted first-order and distorted higher-order beliefs. Under strategic substitutes, players want to choose high effort exactly when they think their opponent will choose low effort. If a player holds a negative stereotype of her opponent and believes that her opponent agrees, she thinks that opponent will not compete aggressively, to which she replies by increasing her effort. Since actions are substitutes, her opponent's best response is to be aware of this stereotype and reduce effort, thus validating the player's initial stereotype.

These effects create group-favoured equilibria in between-group competition. When one group holds negative stereotypes about the other and teaches their children that these stereotypes are commonly held, the other group's optimal response is to make their children aware of these stereotypes, even as they deny their accuracy. When children match across groups, agents from the stereotype-aware group underperform the other group. In these equilibria, children of both groups are behaving rationally given their beliefs: children in the favoured group compete aggressively, believing they are fated to win against a weaker opponent, while the unfavoured group acquiesces to the favoured group's treatment of them, knowing how the favoured group sees them. In these equilibria, children attach 'racial' meaning to the intrinsically meaningless physical markers, by thinking that ability varies systematically between groups, even though markers are uninformative about ability. When agents are

constrained to believe that all agents have the same average ability, then the only equilibrium is the unique symmetric within-group equilibrium. However, as previous work has shown, even avowed egalitarians tend to implicitly believe that 'they' are different from 'us,' permitting the formation of racial beliefs (Gaertner and Dovidio, 1986; Bonilla-Silva, 2003). Furthermore, dominant groups have many codes of conduct and behaviours that reinforce and reward an 'us-them' distinction among children learning racial behaviour (DuRocher, 2011; Tatum, 2017). Thus, intrinsically meaningless groups become a 'race' because social beliefs lead individuals to treat others differently depending on their group membership.

Furthermore, observed behaviour in equilibrium seems to support these beliefs. Despite having identical intrinsic ability on average, the favoured group will win the competition disproportionately often. Hence, I provide a formal interpretation of how incentives to discriminate precede and cause prejudiced beliefs, rather than prejudice preceding discrimination. Finally, my results suggest that reducing the prevalence and harm of negative stereotyping requires breaking the cycle between competition and stereotypical beliefs. Diminishing stereotypes means changing incentives, not correcting cognitive errors. Competition between groups generates stereotypes, even when abstracting from cognitive errors, such as imperfect memory. Although this paper does not model the choice of whether interactions will be competitive or cooperative, negative stereotypes will lead the favoured group to support political choices that create greater competition, rather than fostering greater cooperation. As I show, rising competition leads to worse stereotypes among the dominant group and more self-isolation by the subordinate group. Even improvements in the productivity of intergroup matches may not necessarily reduce stereotypes if these improvements also make exploitation a more attractive option. If segregation also makes it harder for individuals to make personal observations that break entrenched stereotypes, avoidance makes this cycle harder to break. Thus, reducing racial tension may require simultaneously reducing competition in favour of cooperative interactions, preventing self-isolation to avoid exploitation, and preventing the favoured group from ratcheting competition back up as segregation diminishes.

References

Acharya, A., M. Blackwell, and M. Sen (2016). The political legacy of American slavery. *Journal of Politics 78*(3), 621–641.

Adriani, F. and S. Sonderegger (2009). Why do parents socialize their children to behave pro-socially? An information-based theory. *Journal of Public Economics 93*(11), 1119–1124.

Adriani, F. and S. Sonderegger (2012). Setting the right example: A signaling theory of parental behavior. *Mimeo*.

Ali, N. (2011). Learning self-control. *Quarterly Journal of Economics 126*(2), 857–893.

Allport, G. (1954). *The Nature of Prejudice*. Cambridge: Addison-Wesley Publishing Company.

Almås, I., A. Cappelen, E. Sørensen, and B. Tungodden (2010). Fairness and the development of inequality acceptance. *Science 328*, 1176–1178.

Armor, D. A. (1999). The illusion of objectivity: A bias in the perception of freedom from bias. *Dissertations Abstract International: Section B: The Sciences and Engineering 59*(9), 5163.

Arrow, K. J. (1973). The theory of discrimination. In O. Aschenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*, Chapter 1, pp. 3–33. Princeton: Princeton University Press.

Bauman, K. P. and G. Geher (2002). We think you agree: The detrimental impact of the false consensus effect on behavior. *Current Psychology 21*(4), 293–318.

Becker, G. S. (1957). *The Economics of Discrimination*. Chicago: University of Chicago Press.

Bénabou, R. (2015). The economics of motivated beliefs. *Revue d'économie politique 125*(5), 665–685.

Bénabou, R. and J. Tirole (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics 117*(3), 871–915.

Bénabou, R. and J. Tirole (2004). Willpower and personal rules. *Journal of Political Economy 112*(4), 848–886.

Bénabou, R. and J. Tirole (2006). Belief in a just world and redistributive politics. *Quarterly Journal of Economics 121*(2), 699–746.

Bisin, A. and T. Verdier (1998). On the cultural transmission of preferences for social status. *Journal of Public Economics 70*(1), 75–97.

Bisin, A. and T. Verdier (2001). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory 97*(2), 298–319.

Bobo, L. D. (1999). Prejudice as group position: Microfoundations of a sociological approach to racism and race relations. *Journal of Social Issues 55*(3), 445–472.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2019). Inaccurate statistical discrimination. *NBER Working Paper*.

Bolton, P., M. K. Brunnermeier, and L. Veldkamp (2013). Leadership, coordination, and corporate culture. *Review of Economic Studies 80*(2), 512–537.

Bonilla-Silva, E. (2003). *Racism without Racists*. Lanham: Rowman & Littlefield.

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

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–773.

Botvin, G. J., E. M. Botvin, E. Baker, L. Dusenbury, and C. J. Goldberg (1992). The false consensus effect: Predicting adolescents' tobacco use from normative expectations. *Psychological Reports 70*(1), 171–178.

Brandenburger, A. and E. Dekel (1987). Rationalizability and correlated equilibria. *Econometrica 55*(6), 1391–1402.

Brocas, I. and J. D. Carrillo (2000). The value of information when preferences are dynamically inconsistent. *European Economic Review 44*(4), 1104–1115.

Brunnermeier, M. K. and J. A. Parker (2005). Optimal expectations. *American Economic Review 95*(4), 1092–1118.

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2020). Misperceived social norms: Female labor force participation in Saudi Arabia. *American Economic Review 110*(10), 2997–3029.

Caplin, A. and J. Leahy (2001). Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics 116*(1), 55–79.

Carrillo, J. D. and T. Mariotti (2000). Strategic ignorance as a self-disciplining device. *Review of Economic Studies 67*(3), 529–544.

Charness, G., A. Rustichini, and J. van de Ven (2018). Self-confidence and strategic behavior. *Experimental Economics 21*(1), 72–98.

Chauvin, K. P. (2019). A misattribution theory of discrimination. *Mimeo*.

Coate, S. and G. C. Loury (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review 83*(5), 1220–1240.

Coffman, K. (2014). Evidence on self-stereotyping and the contribution of ideas. *Quarterly Journal of Economics 129*(4), 1625–1660.

Compte, O. and A. Postlewaite (2004). Confidence-enhanced performance. *American Economic Review 94*(5), 1536–1557.

Crandall, C. S., A. J. Bahns, R. Warner, and M. Schaller (2011). Stereotypes as justifications of prejudice. *Personality and Social Psychology Bulletin 37*(11), 1488–1498.

Danz, D. N., K. Madarász, and S. W. Wang (2018). The biases of others: Projection equilibrium in an agency setting. *Mimeo*.

Darity, W. A., D. Hamilton, P. L. Mason, G. N. Price, A. Dávila, M. T. Mora, and S. K. Stockly (2017). Stratification economics. In A. Flynn, S. R. Holmberg, D. T. Warren, and F. J. Wong (Eds.), *The Hidden Rules of Race*, Chapter 2, pp. 35–51. Cambridge: Cambridge University Press.

Darity, W. A., D. Hamilton, and J. B. Stewart (2015). A tour de force in understanding intergroup inequality: An introduction to stratification economics. *Review of Black Political Economy 42*(1), 1–6.

Darity, W. A., P. L. Mason, and J. B. Stewart (2006). The economics of identity: The origin and persistence of racial identity norms. *Journal of Economic Behavior & Organization 60*(3), 283–305.

Dekel, E., D. Fudenberg, and S. Morris (2007). Interim correlated rationalizability. *Theoretical Economics 2*(1), 15–40.

Dessi, R. (2008). Collective memory, cultural transmission and investments. *American Economic Review 98*(1), 534–560.

Dewan, T. and S. Wolton (2020). A political economy of social discrimination. *Mimeo*.

DuRocher, K. (2011). *Raising Racists: The Socialization of White Children in the Jim Crow South*. Lexington: University Press of Kentucky.

Eeckhout, J. (2006). Minorities and endogenous segregation. *Review of Economic Studies 73*(1), 31–53.

Ely, J. C. and M. Pęski (2006). Hierarchies of belief and interim rationalizability. *Theoretical Economics 1*(1), 19–65.

Entman, R. M. and K. A. Gross (2000). *The Black Image in the White Mind: Media and Race in America*. Chicago: University of Chicago Press.

Finchilescu, G. (2005). Meta-stereotypes may hinder inter-racial contact. *South African Journal of Psychology 35*(3), 460–472.

Finchilescu, G. (2010). Intergroup anxiety in interracial interaction: The role of prejudice and metastereotypes. *Journal of Social Issues 66*(2), 334–351.

Fowler, C. and J. Gasiorek (2020). Implications of metastereotypes for attitudes toward intergenerational contact. *Group Processes & Intergroup Relations 23*(1), 48–70.

Frick, M., R. Iijima, and Y. Ishii (2019). Dispersed behavior and perceptions in assortative societies. *Mimeo*.

Fuligni, A. J. (2007). *Contesting Stereotypes and Creating Identities: Social Categories, Social Identities, and Educational Participation*. New York: Russell Sage Foundation.

Gaertner, S. L. and J. F. Dovidio (1986). The aversive form of racism. In J. F. Dovidio and S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism*, pp. 61–89. Orlando: Academic Press.

Gennaioli, N. and A. Shleifer (2010). What comes to mind. *Quarterly Journal of Economics 125*(4), 1399–1433.

Glaser, J. (2005). Intergroup bias and inequity: Legitimizing beliefs and policy attitudes. *Social Justice Research 18*(3), 257–282.

Glover, D., A. Pallais, and W. Pariente (2017). Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *Quarterly Journal of Economics 132*(3), 1219–1260.

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

Gorski, P. (2008). The myth of the culture of poverty. *Educational Leadership 65*(7), 32–36.

Griffin, D. and L. Ross (1991). Subjective construal, social inference, and human misunderstanding. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology*, Chapter 24, pp. 319–359. San Diego: Academic Press.

Gross, A. J. (2008). *What Blood Won't Tell: A History of Race on Trial in America*. Cambridge: Harvard University Press.

Guiso, L., P. Sapienza, and L. Zingales (2008). Social capital as good culture. *Journal of the European Economic Association 6*(2), 295–320.

Günther, C., N. A. Ekinci, C. Schwieren, and M. Strobel (2010). Women can't jump? An experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior and Organization 75*(3), 395–401.

Harbaugh, R. and T. To (2014). Opportunistic discrimination. *European Economic Review 66*, 192–204.

Heidhues, P., B. Kőszegi, and P. Strack (2020). Overconfidence and prejudice. *Mimeo*.

Heller, Y. and E. Winter (2020). Biased-belief equilibrium. *American Economic Journal: Microeconomics 12*(2), 1–40.

Hughes, D., J. Rodriguez, E. P. Smith, D. J. Johnson, H. C. Stevenson, and P. Spicer (2006). Parents' ethnic-racial socialization practices: A review of research and directions for future study. *Developmental Psychology 42*(5), 747–770.

Huguet, P. and I. Régner (2009). Counter-stereotypic beliefs in math do not protect school girls from stereotype threat. *Journal of Experimental Social Psychology 45*(4), 1024–1027.

Jerald, M. C., E. R. Cole, L. M. Ward, and L. R. Avery (2017). Controlling images: How awareness of group stereotypes affects Black women's well-being. *Journal of Counseling Psychology 64*(5), 487–499.

Jost, J. T. (1997). Outgroup favoritism and the theory of system justification: A paradigm for studying the effects of socio-economic status on stereotypes. *Mimeo*.

Jost, J. T. and M. R. Banaji (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology 33*(1), 1–27.

Kendi, I. X. (2016). *Stamped from the Beginning: The Definitive History of Racist Ideas in America*. New York City: Nation Books.

Kőszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association 4*(4), 673–707.

Kőszegi, B. (2010). Utility from anticipation and personal equilibrium. *Economic Theory 44*(3), 415–444.

Lagerlöf, J. N. M. (2020). Strategic gains from discrimination. *European Economic Review 122*.

Lazear, E. P. and S. Rosen (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy 89*(5), 841–864.

Leonard, D. J., W. G. Moons, D. M. Mackie, and E. R. Smith (2011). 'We're mad as hell and we're not going to take it anymore': Anger, self-stereotyping, and collective action. *Group Processes & Intergroup Relations 14*(1), 99–111.

Madarász, K. (2012). Information projection: Model and applications. *Review of Economic Studies 79*(3), 961–985.

Madarász, K. (2016). Projection equilibrium: Definition and applications to social investment and persuasion. *CEPR Discussion Paper*.

Marks, G. and N. Miller (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin 102*(1), 72–90.

Martinez, C., C. Paterna, P. Roux, and J. M. Falomir (2010). Predicting gender awareness: The relevance of neo-sexism. *Journal of Gender Studies 19*(1), 1–12.

Martins, N. and K. Harrison (2012). Racial and gender differences in the relationship between children's television use and self-esteem: A longitudinal panel study. *Communication Research 39*(3), 338–357.

Matera, C., C. Stefanile, and R. Brown (2015). Majority-minority acculturation preferences concordance as an antecedent of attitudes towards immigrants: The mediating role of perceived symbolic threat and metastereotypes. *International Journal of Intercultural Relations 45*, 96–103.

McKown, C. and R. S. Weinstein (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development 74*(2), 498–515.

Mertens, J. F. and S. Zamir (1985). Formulation of bayesian analysis for games of incomplete information. *International Journal of Game Theory 14*(1), 1–29.

Meuleman, B., E. Davidov, and J. Billiet (2009). Changing attitudes towards immigration in Europe, 2002-2007: A dynamic group conflict theory approach. *Social Science Research 38*(2), 352–365.

Morning, A. (2009). Toward a sociology of racial conceptualization for the 21st century. *Social Forces 87*(3), 1167–1192.

Morris, S. and H. S. Shin (2000). Rethinking multiple equilibria in macroeconomic modelling. *In Bernanke, B. and Rogoff, K., NBER Macroeconomics Annual 2000, MIT Press*.

Ore, E. J. (2019). *Lynching: Violence, Rhetoric, and American Identity*. Jackson: University Press of Mississippi.

Owuamalam, C. K. and H. Zagefka (2014). On the psychological barriers to the workplace: When and why metastereotyping undermines employability beliefs of women and ethnic minorities. *Cultural Diversity & Ethnic Minority Psychology 20*(4), 521–528.

Papageorge, N. W., S. Gershenson, and K. M. Kang (2020). Teacher expectations matter. *Review of Economics and Statistics 102*(2), 234–251.

Pęski, M. and B. Szentes (2013). Spontaneous discrimination. *American Economic Review 103*(6), 2412–2436.

Pew Research (2015). Modern immigration wave brings 59 million to U.S., driving population growth and change through 2065. Report, Pew Research Center, Washington DC.

Phelps, E. (1972). The statistical theory of racism and sexism. *American Economic Review 62*(4), 659–661.

Phillips, K. W., K. A. Lilijenquist, and M. A. Neale (2009). Is the pain worth the gain? The advantages and liabilities of agreeing with socially distinct newcomers. *Personality and Social Psychology Bulletin 35*(3), 336–350.

Pratto, F., J. Sidanius, and S. Levin (2006). Social dominance theory and the dynamics of intergroup relations: Taking stock and looking forward. *European Review of Social Psychology 17*(1), 271–320.

Pratto, F., L. M. Stallworth, J. Sidanius, and B. Siers (1997). The gender gap in occupational role attainment: A social dominance approach. *Journal of Personality and Social Psychology 72*(1), 37–53.

Pronin, E., D. Y. Lin, and L. Ross (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin 28*(3), 369–381.

Ritterhouse, J. (2006). *Growing Up Jim Crow: How Black and White Southern Children Learned Race*. Chapel Hill: University of North Carolina Press.

Roediger, D. R. (2008). *How Race Survived U.S. History: From Settlement and Slavery to the Eclipse of Post-racialism*. New York: Verso Books.

Ross, L., D. Greene, and P. House (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology 13*(3), 279–301.

Ross, T. (2019). Media and stereotypes. In S. Ratuva (Ed.), *The Palgrave Handbook of Ethnicity*, Chapter 1, pp. 1–17. London: Palgrave Macmillan.

Schmitt, M. T. and J. H. Wirth (2009). Evidence that gender differences in social dominance orientation result from gendered self-stereotyping and group-interested responses to patriarchy. *Psychology of Women Quarterly 33*(4), 429–436.

Schneider, D. (2004). *The Psychology of Stereotyping*. New York: Guilford Press.

Sherif, M. and C. W. Sherif (1953). *Groups in harmony and tension; an integration of studies of intergroup relations*. New York City: Harper & Brothers.

Shih, M. J., T. L. Pittinsky, and G. C. Ho (2013). Stereotype boost: Positive outcomes from the activation of positive stereotypes. In M. Inzlicht and T. Schmader (Eds.), *Stereotype Threat: Theory, Process, and Application*, Chapter 9, pp. 141–156. Oxford: Oxford University Press.

Sigelman, L. and S. A. Tuch (1997). Metastereotypes: Blacks' perceptions of Whites' stereotypes of Blacks. *Public Opinion Quarterly 61*(1), 87–101.

Spears, R., B. Dossje, and N. Ellemers (1997). Self-stereotyping in the face of threats to group status and distinctiveness: The role of group identification. *Personality and Social Psychology Bulletin 23*(5), 538–553.

Spencer, S. J., C. M. Steele, and D. M. Quinn (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology 35*(1), 4–28.

Steele, C. M. (1997). A threat is in the air: How stereotypes shape intellectual identity and performance. *American Psychologist 52*(6), 613–629.

Steele, C. M. (2010). *Whistling Vivaldi: And Other Clues to How Stereotypes Affect Us*. New York: W. W. Norton & Company.

Steele, C. M. and J. Aronson (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology 69*(5), 797–811.

Stephan, W. G., O. Ybarra, and G. Bachman (1999). Prejudice toward immigrants. *Journal of Applied Social Psychology 29*(11), 2221–2237.

Stevenson, H. C. and E. G. Arrington (2009). Racial/ethnic socialization mediates perceived racism and the racial identity of African American adolescents. *Cultural Diversity & Ethnic Minority Psychology 15*(2), 125–136.

Tabellini, G. (2008). The scope of cooperation: Values and incentives. *Quarterly Journal of Economics 123*(3), 905–950.

Tatum, B. D. (2017). *Why are all the black kids sitting together in the cafeteria?* New York: Basic Books.

Torres, K. C. and C. Z. Charles (2004). Metastereotypes and the black-white divide: A qualitative view of race on an elite college campus. *Du Bois Review 1*(1), 115–149.

Tversky, A. and D. Kahneman (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology 5*(2), 207–232.

Tversky, A. and D. Kahneman (1983). Extensional versus intuitive reasoning: The conjuction fallacy in probability judgment. *Psychological Review 90*(4), 293–315.

van Leeuwen, E., J. J. Oostenbrink, and A. Twilt (2014). The combined effects of meta-stereotypes and audience on outgroup and ingroup helping. *Group Dynamics: Theory, Research, and Practice 18*(3), 189–202.

Vorauer, J. D., A. J. Hunter, K. J. Main, and S. A. Roy (2000). Meta-stereotype activation: Evidence from indirect measures for specific evaluative concerns experienced by members of dominant groups in intergroup interaction. *Journal of Personality and Social Psychology 78*(4), 690–707.

Vorauer, J. D., K. J. Main, and G. B. O'Connell (1998). How do individuals expect to be viewed by members of lower status groups? Content and implications of meta-stereotypes. *Journal of Personality and Social Psychology 75*(4), 917–937.

Wakefield, J. R. H., N. Hopkins, and R. M. Greenwood (2013). Meta-stereotypes, social image and help seeking: Dependency-related meta-stereotypes reduce help-seeking behaviour. *Journal of Community & Applied Social Psychology 23*(5), 363–372.

Walton, G. M. and G. L. Cohen (2003). Stereotype lift. *Journal of Experimental Social Psychology 39*(5), 456–467.

Walton, G. M. and S. J. Spencer (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science 20*(9), 1132–1139.

Watt, S. E. and C. Larkin (2010). Prejudiced people perceive more community support for their views: The role of own, media, and peer attitudes in perceived consensus. *Journal of Applied Social Psychology 40*(3), 710–731.

Williams, M. J. and J. L. Eberhardt (2008). Biological conceptions of race and the motivation to cross racial boundaries. *Journal of Personality and Social Psychology 94*(6), 1033–1047.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review 110*(2), 337–361.

Appendix A. Proofs

Proof of Lemma 1:

Proof. Fix the probability that the agent believes her opponent will choose high effort. Then suppose that the expected utility from high effort exceeds the expected utility from low effort. Since costs are iid, the probability that her opponent chooses high effort does not depend on her cost realization. Hence, the expected utility will be strictly higher at any lower cost, so she will still prefer high effort. Symmetrically, if she prefers low effort at some cost, she will strictly prefer low effort at any higher cost. Thus, there exists a threshold c^* as a function of the probability with which she expects her opponent to choose high effort. ■

Proof of Lemma 2:

Proof. Fix the perceived probability that the opponent chooses high effort as $P(H|\tilde{\pi})$. Then by Lemma 1, the threshold c_i^* is the point at which the agent is indifferent between high and low effort. If she chooses high effort, she receives:

$$P(H|\tilde{\pi})(\tilde{\pi}^{G_i}V - c) + (1 - P(H|\tilde{\pi}))(\tilde{\pi}^{G_i}V + b_1 - c)$$

versus $P(H|\tilde{\pi})(-b_2)$ if she does not. Equating the two for c^* and rearranging yields:

$$c^*(\tilde{\pi}) = \tilde{\pi}^{G_i}V + P(H|\tilde{\pi})b_2 + (1 - P(H|\tilde{\pi}))b_1$$

as desired. Then ,since $0 \leq P(H|\tilde{\pi}) \leq 1$, for any beliefs: $\tilde{c}_L^{G_i} \leq c^*(\tilde{\pi}) < \tilde{c}_U^{G_i}$. To show the inequalities are strict, note that since $\underline{c} < b_2$, $P(H|\tilde{\pi}) > 0$ for any beliefs and since $V + b_1 < \bar{c}$, $P(H|\tilde{\pi}) < 1$ for any beliefs. An identical argument verifies the claim for Player 2. ■

Proof of Lemma 3:

Proof. First, to verify that c_0 exists, by Lemma 2, a player's strategy can be described by (3). Under common knowledge of π, players are correct in equilibrium about their opponent's strategy. Thus:

$$c_1 = \pi V + b_2 F(c_2) + b_1(1 - F(c_2))$$

Looking first for symmetric equilibria, I obtain:

$$c = \pi V + b_2 F(c) + b_1(1 - F(c)).$$

Note that the right-hand side is decreasing in c as $b_1 > b_2$. Then since $\underline{c} < b_2$ and $V + b_1 < \bar{c}$ and the right-hand side is continuous, a unique solution exists. Now consider the possibility of asymmetric equilibria, which could occur if all players use group identity as a public coordination device. The recursive best response function for a group W player is:

$$c = (\pi V + b_1) - (b_1 - b_2)F[(\pi V + b_1) - (b_1 - b_2)F(c^W)]$$

where c^W is the strategy chosen by all other group W players. An equilibrium is a fixed point of this function; that is, $c = c^W$. The above argument verifies that c_0 is a fixed point of this function. Next, note that $c > 0$ for $c^W = 0$, but c is bounded above as $c^W \to \infty$. Hence, there must be an odd number of crossing points. Next, since $f(c)$ is single-peaked, there can be at most 3 crossing points. Finally, consider the derivative of the above function:

$$0 < (b_1 - b_2)^2 f(c^W) f((\pi V + b_1) - (b_1 - b_2)F(c^W)) < 1$$

where the second inequality holds since $f(c) < 1/|b_1 - b_2|$. Therefore, the function must cross the 45° line from above at c_0. But then, if there is a crossing less than c_0, there must be at least two. However, since strategies are strategic substitutes and c_0 is the unique symmetric equilibrium, there must be one crossing above c_0 for each one below. But since there cannot be more than 3 total crossings, it follows that c_0 is the unique solution. Hence, the unique

equilibrium under common knowledge of π is the symmetric benchmark equilibrium c_0. In this equilibrium, players ignore group identity. ■

Proof of Proposition 1:

Proof. To show the Proposition, I first must construct explicit expressions for the players' optimal thresholds. To do so, it is useful to switch between the players' probabilities of action and their threshold. Since $F(c)$ has full support on $[\underline{c}, \bar{c}]$, any threshold, c^* maps uniquely to a probability of acting $F(c^*)$. As such, any probability of acting implies a unique optimal threshold. To identify this probability, I use the fact that Nash Equilibrium strategies are those that survive the iterative deletion of dominated strategies, i.e., those that are rationalizable given common knowledge of rationality (Brandenburger and Dekel, 1987; Dekel et al., 2007). To look at this process using probabilities, Player 1 constructs the probability that she believes Player 2 acts according to the following pattern of logic:

> "If Player 2 draws $c < \tilde{\pi}_{1,2}^{G_2} V + b_2$, he will exert high effort for sure. Likewise, if he draws $c > \tilde{\pi}_{1,2}^{G_2} V + b_1$, he will definitely not choose high effort. In between these points, he will follow the same logic as me and act when he expects me not to act."

At the first step of reasoning, Player 1 believes that Player 2 has a dominant strategy to choose high effort with probability: $F(\tilde{\pi}_{1,2}^{G_2} V + b_2)$, which depends on Player 1's second-order beliefs about Player 2's group. Additionally, Player 1 thinks Player 2 will act if his cost falls in the interim region and Player 1's cost gives her a dominant strategy of low effort, which occurs with probability:

$$F(\tilde{\pi}_{1,2}^{G_2} V + b_1) - F(\tilde{\pi}_{1,2}^{G_2} V + b_2)[1 - F(\tilde{\pi}_{1,3}^{G_1} V + b_1]$$

according to Player 1's third-order beliefs. If either of these events occur, Player 1 thinks Player 2 will choose high effort. If instead, both players think the other drew a cost in the

interim region, the pattern of reasoning starts again. Hence, I obtain the following infinite sum:

$$P(H|\tilde{\pi}) = F(\tilde{\pi}_{1,2}^{G_2}V + b_2) + [F(\tilde{\pi}_{1,2}^{G_2}V + b_1) - F(\tilde{\pi}_{1,2}^{G_2}V + b_2)] \cdot \\ \left[1 - F(\tilde{\pi}_{1,3}^{G_1}V + b_1) + [F(\tilde{\pi}_{1,3}^{G_1}V + b_1) - F(\tilde{\pi}_{1,3}^{G_1}V + b_2)]...\right]$$

As this depends on the agent's full hierarchy of higher-order beliefs, I must specify what higher-order belief biases an agent holds to define the perceived probability of high effort. Suppose that Player 1 projects her beliefs and is does not have stereotype awareness. For brevity, let $\tilde{c}_{i,L}^{G_j} = \tilde{\pi}_i^{G_j}V + b_2$ and $\tilde{c}_{i,U}^{G_j} = \tilde{\pi}_i^{G_j}V + b_1$ where $\tilde{\pi}_i^{G_j}$ is Player i's belief about the ability of group G_j $(i, j = 1, 2)$. Applying Definitions 2 and 3, this infinite sum becomes:

$$\begin{aligned} P(H|\tilde{\pi}) &= F(\tilde{c}_{1,L}^{G_2}) + [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})]\left[1 - F(\tilde{c}_{1,U}^{G_1}) + [F(\tilde{c}_{1,U}^{G_1}) - F(\tilde{c}_{1,L}^{G_1})]...\right] \\ &= \sum_{x=0}^{\infty} \left([F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})][F(\tilde{c}_{1,U}^{G_1}) - F(\tilde{c}_{1,L}^{G_1})]\right)^x \left[F(\tilde{c}_{1,L}^{G_2}) + [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})](1 - F(\tilde{c}_{1,U}^{G_1}))\right] \\ &= \frac{F(\tilde{c}_{1,L}^{G_2}) + [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})](1 - F(\tilde{c}_{1,U}^{G_1}))}{1 - [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})][F(\tilde{c}_{1,U}^{G_1}) - F(\tilde{c}_{1,L}^{G_1})]} \end{aligned}$$

where the infinite series converges because $0 < F(\tilde{c}_{i,U}^{G}) - F(\tilde{c}_{i,L}^{G}) < 1$. Hence:

$$P(H|\tilde{\pi}) = \frac{F(\tilde{c}_{1,L}^{G_2}) + [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})](1 - F(\tilde{c}_{1,U}^{G_1}))}{1 - [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})][F(\tilde{c}_{1,U}^{G_1}) - F(\tilde{c}_{1,L}^{G_1})]} \tag{16}$$

Then, using Lemma 2, Player 1's optimal threshold, which defines their probability of choosing high effort is decreasing in $P(H|\tilde{\pi})$ as $b_1 > b_2$. As the construction of $P(H|\tilde{\pi})$ proceeds identically by using Definitions 2 and 3 and taking the infinite sum for other higher-order belief biases, I merely summarize the results.

If Player 1 projects her beliefs, has stereotype awareness, and believes Player 2 projects

his beliefs, I obtain:

$$P(H|\tilde{\pi}) = \frac{F(\tilde{c}_{1,L}^{G_2}) + [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})](1 - F(\tilde{c}_{2,U}^{G_1}))}{1 - [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})][F(\tilde{c}_{2,U}^{G_1}) - F(\tilde{c}_{2,L}^{G_1})]}. \tag{17}$$

If Player 1 projects her beliefs, has stereotype awareness, and believes Player 2 does not project his beliefs, I obtain:

$$P(H|\tilde{\pi}) = \frac{F(\tilde{c}_{1,L}^{G_2}) + [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})](1 - F(\tilde{c}_{1,U}^{G_1}))}{1 - [F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})][F(\tilde{c}_{1,U}^{G_1}) - F(\tilde{c}_{1,L}^{G_1})]}, \tag{18}$$

which is the same as when Player 1 is unaware of stereotypes. Next, if Player 1 does not project and also is not aware of stereotypes, I obtain:

$$P(H|\tilde{\pi}) = \frac{F(\tilde{c}_{2,L}^{G_2}) + [F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2})](1 - F(\tilde{c}_{1,U}^{G_1}))}{1 - [F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2})][F(\tilde{c}_{1,U}^{G_1}) - F(\tilde{c}_{1,L}^{G_1})]}. \tag{19}$$

If Player 1 does not project, has stereotype awareness, and thinks Player 2 projects his beliefs, I obtain:

$$P(H|\tilde{\pi}) = \frac{F(\tilde{c}_{2,L}^{G_2}) + [F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2})](1 - F(\tilde{c}_{2,U}^{G_1}))}{1 - [F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2})][F(\tilde{c}_{2,U}^{G_1}) - F(\tilde{c}_{2,L}^{G_1})]}. \tag{20}$$

Finally, if Player 1 does not project, has stereotype awareness, and thinks Player 2 does not project, I obtain:

$$P(H|\tilde{\pi}) = \frac{F(\tilde{c}_{2,L}^{G_2}) + [F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2})](1 - F(\tilde{c}_{1,U}^{G_1}))}{1 - [F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2})][F(\tilde{c}_{1,U}^{G_1}) - F(\tilde{c}_{1,L}^{G_1})]}, \tag{21}$$

which is the same as when Player 1 is unaware of stereotypes. From the pairs of equations (16) and (18) and (19) and (21), stereotype awareness affects a player's strategy only if she believes her opponent projects his beliefs. Now consider the interim claim that the results in the lemma hold provided:

$$\tilde{\pi}_2^{G_2} > \frac{\tilde{\pi}_1^{G_2}}{D_2}$$

where D_2 satisfies:

$$D_2 \begin{cases} = 1 & \text{if } F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2}) = F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2}) \\ > 1 & \text{if } F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2}) > F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2}) \\ < 1 & \text{if } F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2}) < F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2}) \end{cases}$$

To prove this claim, I compare (16) to (19) and (17) to (20), which both differ only in which player's beliefs about group G_2 matter. There are three cases:

Case 1 - $F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2}) = F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2})$: In this case, the expressions only differ in the first term of the numerator. Hence, the probability is higher when Player 1 projects if and only if $1 - F(\tilde{c}_{1,U}^{G_2}) > 1 - F(\tilde{c}_{2,U}^{G_2})$, which is true exactly when $\tilde{\pi}_1^{G_2} < \tilde{\pi}_2^{G_2}$, so $D_2 = 1$.

Case 2 - $F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2}) > F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2})$: In this case, the denominator of the expressions when Player 1 projects (16) and (17) are smaller than the denominators when she does not (19) and (20). Thus, even if $F(\tilde{c}_{1,L}^{G_2}) > F(\tilde{c}_{2,L}^{G_2})$, it is possible that (19) > (16), which implies Player 1's threshold is higher. In particular, if $\tilde{\pi}_1^{G_2} = \tilde{\pi}_2^{G_2}$, it is still the case that (16) < (19), which implies $D_2 > 1$.

Case 3 - $F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2}) < F(\tilde{c}_{2,U}^{G_2}) - F(\tilde{c}_{2,L}^{G_2})$: Finally, the denominators of (16) and (17) are larger than the denominators of (19) and (20), so the numerators must also be sufficiently larger for Player 1's threshold to be higher under projection. Hence, this implies $D_2 < 1$.

Finally, examine (16) and (17). In both these equations, $P(H|\tilde{\pi})$ is increasing in $\tilde{\pi}_1^{G_2}$ as the direct effect through $F(\tilde{c}_{1,L}^{G_2})$ dominates the indirect effect through $F(\tilde{c}_{1,U}^{G_2}) - F(\tilde{c}_{1,L}^{G_2})$. By Lemma 2, this proves that Player 1's optimal threshold is decreasing in $\tilde{\pi}_1^{G_2}$ when they

project their beliefs.

Next, examine (17) and (20). In both these equations, $P(H|\tilde{\pi})$ is decreasing in $\tilde{c}_2^{G_1}$ by the same argument as before, so Player 1's optimal threshold is decreasing in $\tilde{\pi}_2^{G_1}$ when she has stereotype awareness and thinks Player 2 projects his beliefs. Since the game is symmetric, I can reverse the roles of Player 1 and 2, which shows that Player 2's optimal threshold decreases in $\tilde{\pi}_1^{G_2}$ when he has stereotype awareness and thinks Player 1 projects.

Finally, to prove the proposition, observe that when $\tilde{\pi}_2^{G_2} = \tilde{\pi}_1^{G_2}$, $D_2 = 1$. Since the interim claim holds, Player 1's threshold is decreasing and Player 2's threshold is increasing in $\tilde{\pi}_1^{G_2}$. Therefore, by continuity, it continues to hold in the neighbourhood of $\tilde{\pi}_2^{G_2} = \tilde{\pi}_1^{G_2}$, when $D_2 \neq 1$. ■

Proof of Lemma 4:

Proof. (5) and (6) jointly define a strategy recursion, $c_1 = G(c_{G_1})$, where:

$$G(c_{G_1}) = (1+\theta)(\pi V + b_1) - (1+\theta)(b_1 - b_2)F[(1+\theta)(\pi V + b_1) - (1+\theta)(b_1 - b_2)F(c_{G_1})].$$

This recursion represents the following logic: if the rest of Player 1's group follows threshold strategy c_{G_1} and the opposing parent best responds to that strategy, $G(c_{G_1})$ is Player 1's parent's best response to the opposing parent's strategy. Thus, Nash equilibrium among parents is a fixed point of this strategy function; that is, points such that $G(c_{G_1}) = c_{G_1}$. To show that at least one point exists, note that $G(c_{G_1})$ is bounded below by:

$$(1+\theta)(\pi V + b_1) - (1+\theta)(b_1 - b_2)F[(1+\theta)(\pi V + b_1)] > 0$$

and bounded above by:

$$(1+\theta)(\pi V + b_1) - (1+\theta)(b_1 - b_2)F[(1+\theta)(\pi V + b_2)]$$

Next, the derivative of $G(c_{G_1})$ is:

$$G'(c) = (1+\theta)^2(b_1 - b_2)^2 f(c) f((1+\theta)(\pi V + b_1) - (1+\theta)(b_1 - b_2)F(c)) > 0$$

Then, since $G(c)$ is continuous by the assumption that $F(c)$ has full support, there is at least one point where $G(c_{G_1}) = c_{G_1}$.

Next, observe that there are always an odd number of fixed points as the function $G(0) > c$, but $G(c) < c$ as $c \to \infty$, which implies that at least one fixed point is best response to itself. To show that there can be at most one such point, suppose instead that two points $c_1 < c_2$ were both best response to themselves. By Lemma 2, the strategies that these are best response to, $\hat{c}_1$ and $\hat{c}_2$, respectively must satisfy $\hat{c}_1 > \hat{c}_2$. But this contradicts the assumption that the points are both best responses to themselves. Hence, there is a unique point, c_S which is best response to itself. Then, by Lemma 2, the symmetric equilibrium is given by (7).

Now suppose that Assumption 1 holds, when this is true, $G'(c_S) > 1$, which implies that $G(c)$ must cross the 45° line from below. However, as noted above, $G(0) > c$, but $G(c) < c$ as $c \to \infty$. Hence, if it crosses from below, there must be at least two additional crossings. Finally, the assumption that the distribution $F(c)$ has at most one peak ensures that there are at most 3 fixed points. Thus, when Assumption 1 holds, there are exactly 3 fixed points, two of which are best response to each other, given by (10) and (11) and one is the unique point c_S.

Finally, suppose that Assumption 1 is reversed. Then, $G(c)$ must cross the 45° line from

above at c_S. Since this point is best response to itself, by Lemma 2, for any additional pairs of fixed points, there must be one above and one below c_S. But to cross from above, there must be an even number of intersections less than c_S. There cannot be 2 or more by the assumption of single-peakedness, so there must be zero. Then, there cannot be any intersections greater than c_S. Thus, if Assumption 1 is reversed, c_S is the unique fixed point and best response to itself. ■

Proof of Theorem 1:

Proof. A within-group equilibrium requires that both players hold the same beliefs by the symmetry requirement. By Lemma 2, this implies that they follow the same threshold strategy in equilibrium. Hence, the only viable equilibrium is the fixed point c_S, which is best response to itself, and is defined implicitly by (7). Then, by Lemma 2, (8) defines the child's optimal strategy as a function of her parent's choice of beliefs. Setting (7) equal to (8) and rearranging, yields (9). Since at point c_S, with beliefs given by (9), children's strategies are optimal given their beliefs, parents' choices are optimal given the strategies of other parents, and all children of the same group hold the same beliefs, this is an equilibrium. To show that this equilibrium is stable, consider the best response function:

$$G(c_{G_1}|c_{G_1} = c_{G_2}) = (1+\theta)[\pi V + b_2 F(c_{G_1}) + b_1(1 - F(c_{G_1}))]$$

Since this function is decreasing in c_{G_1}, if the group deviated to $c_S + \epsilon$, the best response would be a strategy less than c_S, but greater than $c_S - \epsilon$. Thus, iterating on best responses converges to the unique equilibrium. ■

Proof of Theorem 2:

Proof. First, suppose that Assumption 1 holds. To verify that the strategy pairs (c_U, c_L), (c_S, c_S), and (c_L, c_U) are equilibria, I check that the conditions for equilibria are satisfied.

Consider first the symmetric equilibrium. By Theorem 1, if children believe that the opposing group has the same ability as them, they will believe they follow the same strategy. The same argument as above verifies that this is an equilibrium. Now consider the equilibrium where group W children play c_U and group B children play c_L. Given (10) and (11), these strategies are best response from the parent's perspective. Next, since all children from a group pursue the same strategy, Lemma 2 implies that there exist beliefs such that they all hold the same beliefs. Finally, by Lemmas 2 and 4, since $c_L, c_U \in [G(0), G(\underline{c})]$, there exist beliefs such that each strategy is optimal for children of that group. Hence, this pair is an equilibrium and the same argument verifies the other asymmetric equilibrium. To observe stability of equilibria, suppose that a player's strategy was not at equilibrium and updated according to the recursion $G(c)$, which is one step of best response iteration. When Assumption 1 holds, $G(c) < c$ for $c \in (c_L, c_S)$ and $G(c) > c$ for $c \in (c_S, c_U)$. Hence, the asymmetric equilibria are stable, but the symmetric equilibrium is unstable. Furthermore, this demonstrates that when Assumption 1 is reversed, the unique equilibrium is a stable symmetric equilibrium at c_S.

Next, to identify beliefs in the symmetric equilibrium, (8) shows that if children correctly anticipate their opponent's strategy as c_S, they will play it in response, as their self-image is pinned down by (9). This requires that each group's belief about the opposing group's ability matches what that group thinks about themselves: $\tilde{\pi}_B^W = \tilde{\pi}_W^B = \tilde{\pi}^G$.

Now, consider beliefs in the W-favoured equilibrium. If the group W child had correct beliefs, other than her optimistic self-image, she would choose:

$$c(\tilde{\pi}) = \tilde{\pi}^G V + F(c_L)b_2 + (1 - F(c_L))b_1 < c_U$$

To see that the inequality holds, observe that replacing c_S with c_L in (9) would lead to

a larger $\tilde{\pi}_i^{G_i}$, which implies that inducing c_U solely through overconfidence would require a higher level of confidence than the one pinned down by within-group competition. Hence, the child must hold additional biased beliefs to distort her threshold upward. By Proposition 1, if Player 1 holds negative stereotypes about Player 2 and projects her beliefs, she will choose a higher threshold than if she does not. Hence, $\tilde{\pi}_W^B < \tilde{\pi}^B$ and Player 1 projects her beliefs. Player 2's parent must distort his strategy downward to c_L, compared to c_S in within-group matches. By Proposition 1, this occurs if Player 2 has stereotype awareness. However, since:

$$\tilde{\pi}^B V + F(c_U)b_2 + (1 - F(c_U))b_1 < c_L,$$

by the same argument as above, holding correct beliefs about Player 1's strategy will distort Player 2's strategy downward too far. Thus, again using Proposition 1, Player 2 will hold a negative stereotype of Player 1 and project. Then, Proposition 1 shows that $\tilde{\pi}_W^B < \tilde{\pi}_B^W < \tilde{\pi}^G$.

Finally, to show that there does not exist a mixed strategy, observe that Lemma 2 implies that conditional on beliefs, children do not randomize. Now consider the parents' decision. Suppose that parents in group B randomized between multiple hierarchies of beliefs. Then, by Lemma 2, there exists a unique threshold that the parents in group W want to induce their children to follow. This follows because given beliefs, children do not randomize over thresholds, so the probability of high effort is just the compound probability based on the randomization probability and the threshold. But then, if there is a unique threshold, group W parents do not want to randomize over beliefs. Finally, if group W parents are not randomizing, then group B parents face a single threshold and thus also do not wish to randomize. Hence, there cannot be mixed-strategy equilibria. ■

Proof of Corollary 3:

Proof. Consider the W-favoured equilibrium. Player 1's strategy is c_U, which is the fixed point of $G(c)$ defined by (10). Fix b_1 and b_2 and consider an increase in θ. From $G(c)$, this shifts all fixed points upwards. Furthermore, c_U shifts upward relative to c_S. By Proposition 1, to increase Player 1's threshold, her parent must distort her belief about Player 2 downward. Hence, $\tilde{\pi}^B_W$ is decreasing in θ, which proves the first claim.

Similarly, fix θ and consider an increase in $b_1 - b_2$. Again using $G(c)$, an increase in $b_1 - b_2$ increases the fixed points and causes c_U to shift upwards relative to the others. As before, this implies that $\tilde{\pi}^B_W$ is decreasing in $b_1 - b_2$, which proves the second and third claims. An identical argument holds for the B-favoured equilibrium. ■

Proof of Proposition 2:

Proof. First, define $\tilde{c}^G_{i,L} = \tilde{\pi}^G_i V + b_1$ and $\tilde{c}^G_{i,U} = \tilde{\pi}^G_i V + b_2$. Then, by the same geometric series construction as Proposition 1, the probability Player 1 (group W) acts when she projects her beliefs and Player 2 (group B) has stereotype awareness is:

$$P_1(H) = F(\tilde{c}^W_{1,L}) + [F(\tilde{c}^W_{1,U}) - F(\tilde{c}^W_{1,L})] \left[\frac{F(\tilde{c}^B_{1,L}) + [F(\tilde{c}^B_{1,U}) - F(\tilde{c}^B_{1,L})]F(\tilde{c}^W_{1,L})}{1 - [F(\tilde{c}^W_{1,U}) - F(\tilde{c}^W_{1,L})][F(\tilde{c}^B_{1,U}) - F(\tilde{c}^B_{1,L})]} \right]$$

which is increasing in $\tilde{\pi}^B_W$. Then, note that a player's optimal threshold is increasing in the other player's probability of acting by Lemma 2. Hence, Player 2's threshold is also increasing in $\tilde{\pi}^B_W$. The same argument verifies the claim with the roles reversed. ■

Proof of Lemma 5:

Proof. Using the implicit function theorem on the first-order condition (14), I obtain:

$$\frac{\partial \tilde{\pi}^*}{\partial \alpha} = -\frac{\frac{\partial V^W_{cm}}{\partial \tilde{\pi}} - \frac{\partial V^W_{cp}}{\partial \tilde{\pi}}}{\alpha \frac{\partial^2 V_{cm}}{\partial \tilde{\pi}^2} + (1-\alpha)\frac{\partial^2 V_{cp}}{\partial \tilde{\pi}^2}}$$

which is negative by Assumption 3. Also:

$$\frac{\partial \tilde{\pi}^*}{\partial \gamma} = -\frac{\alpha \frac{\partial^2 V_{cm}}{\partial \tilde{\pi} \partial \gamma} + (1-\alpha)\frac{\partial^2 V_{cp}}{\partial \tilde{\pi} \partial \gamma}}{\alpha \frac{\partial^2 V_{cm}}{\partial \tilde{\pi}^2} + (1-\alpha)\frac{\partial^2 V_{cp}}{\partial \tilde{\pi}^2}}$$

The denominator is clearly negative by Assumption 3, so the expression is positive if and only if the numerator is positive. By Assumption 3, the first term is negative and the second is positive. Hence, there exists $\hat{\alpha}$ such that the expression is positive if and only if $\alpha < \hat{\alpha}$, which verifies the claim. ■

Proof of Lemma 6:

Proof. Equation (15) holds with equality only if the term in the square brackets is negative as $\lambda'(a) < 0$ by Assumption 2. Next, $V_{cp}^B(\tilde{\pi}, \gamma) \geq V_{cm}^B(\tilde{\pi}, \gamma)$. By Lemma 5, $\tilde{\pi}$ is decreasing in α, so given the assumptions, (15) is decreasing in α, which proves the existence of $\bar{\alpha}$ such that (15) holds with equality for $\alpha > \bar{\alpha}$.

Suppose that (15) holds with equality. Then, applying the implicit function theorem. I obtain:

$$\frac{\partial a^*}{\partial \alpha} = -\frac{\lambda'(a^*)(V_{cm}^B - V_{cp}^B)}{\lambda''(a^*)[\alpha V_{cm}^B(\tilde{\pi}, \gamma) + (1-\alpha)V_{cp}^B(\tilde{\pi}, \gamma)] - c''(a^*)}.$$

Since a^* is a maximizer of group B agents' avoidance choice, the second-order condition must be satisfied, implying that the denominator is negative. Then, since $\lambda'(a)$ is negative and $V_{cm}^B < V_{cp}^B$ when (15) holds with equality, the numerator is positive. Hence:

$$\frac{\partial a^*}{\partial \alpha} > 0$$

Next, again using the implicit function theorem:

$$\frac{\partial a^*}{\partial \tilde{\pi}} = -\frac{\lambda'(a^*)[\alpha \frac{\partial V_{cm}^B}{\partial \tilde{\pi}} + (1-\alpha)\frac{\partial V_{cp}^B}{\partial \tilde{\pi}}] - c'(a)}{\lambda''(a^*)[\alpha V_{cm}^B(\tilde{\pi}, \gamma) + (1-\alpha)V_{cp}^B(\tilde{\pi}, \gamma)] - c''(a^*)} < 0$$

where the numerator is negative by Assumptions 2 and 3 and the denominator is the second-order condition. Finally:

$$\frac{\partial a^*}{\partial \gamma} = -\frac{\lambda'(a^*)[\alpha \frac{\partial V_{cm}^B}{\partial \gamma} + (1-\alpha)\frac{\partial V_{cp}^B}{\partial \gamma}] - c'(a)}{\lambda''(a^*)[\alpha V_{cm}^B(\tilde{\pi}, \gamma) + (1-\alpha)V_{cp}^B(\tilde{\pi}, \gamma)] - c''(a^*)} < 0$$

■

Proof of Proposition 3:

Proof. Consider group W agents' first-order condition when avoidance is possible:

$$\lambda'(a)\frac{\partial a}{\partial \tilde{\pi}}[\alpha V_{cm}^W + (1-\alpha)V_{cp}^W] + \lambda(a)[\alpha \frac{\partial V_{cm}^W}{\partial \tilde{\pi}} + (1-\alpha)\frac{\partial V_{cp}^W}{\partial \tilde{\pi}}]$$

Since $\lambda'(a)$ and $\frac{\partial a}{\partial \tilde{\pi}}$ are negative and $[\alpha V_{cm}^W + (1-\alpha)V_{cp}^W]$ is positive, the first term is positive, which implies that the second term must be negative for the condition to hold. However, the second term is the first-order condition when no avoidance is possible. Since the second derivatives of V_{cm} and V_{cp} are negative, this implies that when (15) holds with equality, $\tilde{\pi}_0^* < \tilde{\pi}_a^*$. Obviously, if group B agents would not choose to avoid intergroup matches when given the option, group W's choice is identical to where no avoidance is possible. Hence if $a = 0$, $\tilde{\pi}_0^* = \tilde{\pi}_a^*$, which proves the claim. ■