Comment:

Theoretical Approaches in Stratification Economics

by Brendan Brundage, Dan J. McGee, and Daniele Tavani

Alex Imas

University of Chicago Booth School of Business

Brundage, McGee, and Tavani have written a very important and comprehensive chapter on the role of formal theory in Stratification Economics. The chapter does a nice job connecting different papers and literatures, providing a clear roadmap for researchers interested in the intersection of SE and formal modeling. It is the kind of piece that will serve as a reference for years—both for those already working in SE, who want to see how game theory and mechanism design can sharpen their intuitions, and for formal theorists outside the tradition who are looking for substantive, important questions to work on. The authors deserve particular credit for the very useful distinction they draw between strategic discrimination and the more atomistic sources of discrimination typically studied in mainstream economics. This distinction is not merely definitional; it carries important conceptual and analytical consequences, and the chapter shows why.

In what follows, I develop some of the comments I provided the authors on the chapter that I hope will be generative for the research agenda. My remarks are organized around three broad themes: the structure of the micro section, the treatment of macro models, and some suggestions for how the chapter might more forcefully demonstrate the value of formal theory for the SE research program. I also sketch some simple formal arguments to illustrate the kinds of results I have in mind.

The case for formalism in SE. Let me start with what I think is the most important contribution of the chapter: the argument for why formal theory matters for SE. The authors lay out four compelling reasons—precision about the objects being studied (at the cost of breadth), the ability to derive non-obvious predictions, the push toward generalizable mechanisms, and the clarification of race as a socio-political tool rather than a culturally deterministic folk category. I think there is room to provide support for each reason.

Consider, for instance, the claim that formal models yield non-obvious predictions. This is true,

and it is one of the great virtues of theory. Research should develop these non-obvious predictions in the case of SE to demonstrate formal theory's value for the field. The divert-and-exploit framework, for example, produces the striking result that White agents would prefer to interact with Black opponents rather than members of their own group—a prediction that runs directly counter to standard taste-based models of discrimination. Developing this prediction further within a richer framework and providing clean empirical evidence for the specific mechanism of the prediction would be a real contribution. Similarly, the discussion of how cultural narratives can ease coordination on asymmetric equilibria could be connected more explicitly to specific historical episodes where the predictions of the formal model align with (or diverge from) what we observe in the data.

Benchmark micro model of SE. It would be useful to flesh out a benchmark micro model of SE using the concepts introduced in the micro section. Begin with a general setup where agents choose effort, face some form of competition or cooperation, and belong to identifiable social groups. In that environment, lay out the conditions under which symmetric and asymmetric equilibria arise. Then show how divert-and-exploit and divide-and-conquer each emerge as special cases of this workhorse model under different assumptions about the structure of interaction (horizontal versus vertical, negative versus positive spillovers). The advantage of this approach is that it makes the connections between the different models transparent. It also makes transparent exactly where the SE assumptions—the centrality of groups, the rivalrous nature of social rewards, the strategic character of discrimination—are doing the analytical work.

To see what I mean, consider a stripped-down version of the chapter's framework that nests both patterns. Two agents $i \in \{1, 2\}$ with group identities $g_i \in \{B, W\}$ simultaneously choose effort $e_i \geq 0$. Agent i's payoff is

$$U_i(e_i, e_j) \;=\; f(e_i, e_j) \;+\; \sigma \cdot h(e_j) \;-\; \kappa\, e_i, \tag{1}$$

where f captures the direct return to own effort (potentially mediated by a contest), $h(e_j)$ captures spillovers from the other agent's effort, $\sigma \in \mathbb{R}$ indexes the sign and magnitude of those spillovers, and $\kappa > 0$ is the marginal cost of effort. When $\sigma < 0$, we are in the competitive world of the Tullock contest: the other agent's effort hurts me. When $\sigma > 0$, we are in the cooperative world of public goods production: the other agent's effort helps me. The chapter's two patterns of strategic discrimination emerge from this single specification as follows:

- *Divert and exploit* ($\sigma < 0$). Because e_j is harmful, the advantaged group discriminates by *increasing* effort, forcing the marginalized group to back off. In the asymmetric equilibrium, $e_W = \bar{e} > e^{sym} > \underline{e} = e_B$, and group W captures a disproportionate share of the contested

prize.

- *Divert and exploit under positive spillovers* ($\sigma > 0$). Now e_j is beneficial, so the advantaged group discriminates by *decreasing* effort, free-riding on the marginalized group's contributions. Here $e_W = \underline{e} < e^{sym} < \bar{e} = e_B$, and group W enjoys the fruits of B's labor while contributing less.

In both cases, group identity serves as the coordination device that selects the asymmetric equilibrium over the symmetric one. This is the core analytical content of strategic discrimination, and the fact that a single framework generates both patterns—simply by varying the sign of σ—is itself a non-obvious insight. With $\sigma < 0$, racial hierarchy looks like aggression; with $\sigma > 0$, it looks like exploitation of labor. Both are equilibrium phenomena sustained by the same underlying logic.

The divide-and-conquer pattern extends this naturally to a vertical setting. A mechanism designer (the "elite") faces two agents whose joint action is costly to the designer. By offering differential treatment $\delta_A > \delta_B$ to break coordination, the designer achieves the same deterrence at lower cost. What the unified framework makes visible is that the horizontal and vertical patterns are complements: the elite's divide-and-conquer creates the initial asymmetry, and the agents' divert-and-exploit sustains it.

It would be particularly exciting to see how this workhorse model can explicitly generate the phenomenon that Darity (2022) describes as groups "acting as individuals." This is one of the core ideas of SE, and it is currently outside of formal theory on discrimination: how do you get from individual incentives to coordinated group behavior when each person has an incentive to free ride? The chapter discusses several mechanisms—social norms, club goods, identity capital—it would be helpful to use them as building blocks within an integrated framework.

The chapter's general principal-agent setup between individuals and their groups provides the right scaffolding. Each member of group G chooses effort e_i, and the group's collective interest calls for $e^*_G > e^*_{ind}$ because individuals do not fully internalize the returns their effort creates for the group. Community enforcement closes this gap: in a repeated setting, the group sustains $e_i = e^*_G$ by threatening deviators with punishment (reversion to e^*_{ind}, or exclusion from group-produced club goods). The question is: when does this work, and when does it break down?

A simple observation is useful here. The sustainability of the group norm depends on the patience of group members (their effective discount factor δ), the severity of punishment, and the temptation to deviate. For the advantaged group engaged in strategic discrimination, the temptation to deviate is to stop discriminating and pocket the short-run gains from cooperating with the outgroup. But if the group can exclude deviators from the club goods that group membership provides, then the

condition for sustaining strategic discrimination is approximately

$$\frac{\delta}{1-\delta}\Big[\underbrace{U_i(\bar{e},\underline{e})+\phi(G)}_{\text{group norm payoff}}-\underbrace{U_i(e^{sym},e^{sym})}_{\text{punishment payoff}}\Big] \;\geq\; \underbrace{U_i(e^{dev},\underline{e})-U_i(\bar{e},\underline{e})}_{\text{deviation gain}}, \tag{2}$$

where $\phi(G)$ represents the flow value of club goods accessible only to members in good standing. In writing (2) I assume that a deviator enjoys $\phi(G)$ in the deviation period but is excluded starting the following period; if exclusion is immediate, the deviation gain on the right-hand side should instead be $U_i(e^{dev},\underline{e})-U_i(\bar{e},\underline{e})-\phi(G)$. The interesting prediction is that strategic discrimination is *easier* to sustain when the group's club goods are more valuable—that is, when group identity provides more material benefits. This connects the "how discrimination benefits the advantaged" question directly to the "how groups cohere" question in a way that generates testable implications. A real contribution would be generating further non-obvious predictions about when and how groups cohere. It would also, I believe, open the door for other economic theorists to work in this space, because they would have a clear, canonical model to build on and extend.

The macro section and the role of groups. I found the macro section of the chapter well structured, precisely because it begins with a basic workhorse model and then explores implications. But I believe that there is more of SE to add to the macroeconomic framework. The baseline model focuses on individuals—a marginalized individual M who faces lower returns to effort with increasing discrimination, and a dominant individual D who chooses effort and a level of discrimination to maximize his position relative to M. Discrimination emerges in equilibrium and persists across generations, which is an important result. But these agents could be interpreted as cohesive groups rather than individuals, and the chapter does not push hard on what changes (or what new predictions emerge) when you take the group interpretation seriously.

Here are some potential directions. It would be useful to outline the implications for the stability of equilibria when groups—rather than individuals—struggle over positional resources. If the dominant group is itself composed of heterogeneous members with varying commitments to the discriminatory equilibrium (as the micro section suggests), does this affect the persistence of discrimination at the macro level? Conversely, if the marginalized group can partially overcome its collective action problem through the mechanisms discussed in the micro section, what are the macroeconomic consequences? These questions sit at the intersection of the micro and macro, and deriving answers to them formally would further help illustrate the importance of theory for SE. Other questions include the predictions of SE for the dynamics of wealth accumulation, as well as conditions under which discriminatory equilibria collapse versus those under which they become more entrenched.

Incorporating a theory of power. Much of SE deals, explicitly or implicitly, with power: the power of dominant groups to set the rules of competition, to define what counts as merit, to control access to resources. It would be useful to incorporate the formal concept of power into a model of SE. Here is a potential roadmap. The framework developed by Bowles and Gintis (1992) offers a natural starting point. In their model, employment relationships are characterized by a power asymmetry: employers have the ability to extract surplus from workers because the threat of dismissal is costly to workers in a world of involuntary unemployment. The efficiency wage models discussed in the chapter's treatment of divide-and-conquer already gesture in this direction, but the connection could be made more explicit.

Consider the following. In a labor market with efficiency wages, group status can advantage the dominant group by reducing the dismissal threat they face relative to the marginalized group. Following the notation in the chapter, if effort is non-contractible and workers' job-finding rate $f(u)$ is decreasing in unemployment, then the no-shirking wage for a worker from group $G \in \{B, W\}$ is

$$w_G \;\geq\; b + c(e)\left(1 + \frac{r + f(u)}{d_G}\right), \tag{3}$$

where b is the unemployment benefit, $c(e)$ the cost of effort, r the discount rate, and d_G the group-specific dismissal hazard. The chapter already notes that if $d_B > d_W$—Black workers face harsher dismissal threats—then $w_B < w_W$ in equilibrium. But the Bowles-Gintis logic reveals something further. The employer's per-worker surplus from group G is $y - w_G$, where y is output. The *difference* in surplus extracted across groups is

$$\begin{aligned}(y - w_B) - (y - w_W) \;&=\; w_W - w_B \\ &=\; c(e)\left(\frac{1}{d_W} - \frac{1}{d_B}\right)(r + f(u)) \\ &=\; c(e)(r + f(u))\frac{d_B - d_W}{d_W d_B}.\end{aligned} \tag{4}$$

This expression makes two things visible. First, the exploitative surplus wedge is increasing in the dismissal-rate gap $d_B - d_W$, giving employers (or advantaged groups that shape workplace discipline) an incentive to sustain—or widen—unequal firing risks across groups. Second, holding d_B and d_W fixed, the wedge is increasing in $f(u)$: when jobs are easier to find, workers' outside options improve, so the no-shirking wage for the protected group must rise more than for the group under threat, implying a larger equilibrium wage gap $w_W - w_B$ and hence a larger surplus differential. Since $f(u)$ is decreasing in unemployment, this delivers a clean comparative-static prediction that the discriminatory wage gap is larger in tight labor markets (i.e., procyclical). At the same time, the countercyclical patterns in racial labor-market gaps emphasized in the chapter

(e.g., Cajner et al. (2017); Boulware and Kuttner 2019) point to a natural extension in which the discipline wedge itself, $d_B - d_W$, becomes more severe in downturns—amplifying group-specific job-loss risk precisely when aggregate conditions deteriorate.

To sum up: Brundage, McGee, and Tavani have written a chapter that fills an important gap in the literature. For SE researchers, it demonstrates that formal theory is useful for making their insights more precise, testable, and generalizable. For formal theorists, it identifies a set of deep, important questions where their tools can make real contributions.