Comparative Advantage in Social Interactions*

Steve Cicala — *University of Chicago*
Roland G. Fryer, Jr. — *Harvard University*
Jörg L. Spenkuch — *Northwestern University*

First Draft: March 2011
This Version: January 2016

Abstract

We propose a model of social interactions based on comparative advantage. When comparative advantage is the guiding principle of social interactions, the effect of moving a student into an environment with higher-achieving peers depends on where in the ability distribution she falls and the shadow prices that clear the social market. We show that the model's key prediction—an individual's ordinal rank predicts her behavior and test scores, *ceteris paribus*—is borne out in one randomized controlled trial in Kenya as well as two large observational data sets from the U.S. To test whether comparative advantage can explain the effect of rank on outcomes, we conduct an experiment with nearly 600 public school students in Houston. The experimental results suggest that social interactions are, at least in part, governed by comparative advantage.

*Previous versions of this paper circulated under the title "A Roy Model of Social Interactions." We are grateful to Gary Becker, Edward Glaeser, Bryan Graham, Richard Holden, Lawrence Katz, Steven Levitt, Franziska Michor, Bruce Sacerdote, Chris Shannon, Jesse Shapiro, Andrei Shleifer, Glen Weyl, as well as seminar participants at Harvard and Chicago for many helpful comments and suggestions. Brad Allan, Vilsa Curto, Tanaya Devi, Matt Davis, Ryan Fagan, Natalya Naumenko, and Wonhee Park provided excellent research assistance. Financial support from the Weatherhead Center for International Affairs and Institute for Humane Studies [Cicala], the Education Innovation Lab at Harvard University [Fryer], and the German National Academic Foundation [Spenkuch] is gratefully acknowledged. Correspondence can be addressed to the authors at Harris School of Public Policy, University of Chicago, 1155 E 60th Street, Chicago IL 60637 [Cicala]; Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge MA 02138 [Fryer]; MEDS Department, Kellogg School of Management, 2001 Sheridan Road, Evanston IL 60208 [Spenkuch]; or by e-mail: scicala@uchicago.edu [Cicala], rfryer@fas.harvard.edu [Fryer], or j-spenkuch@kellogg.northwestern.edu [Spenkuch]. The usual caveat applies.

1. Introduction

Comparative advantage is one of the most important ideas in modern economic thought. Developed by Ricardo (1817), the classical theory of comparative advantage was meant to explain why nations may gain from engaging in trade, even if one country's workers are more efficient at producing every single good than their counterparts abroad. Nearly two hundred years after Ricardo's insight, the principle of comparative advantage has permeated well beyond international trade. Relative, rather than absolute, advantage has played an important role in explaining phenomena as wide-ranging as economic growth (Grossman and Helpman 1990; Matsuyama 1992), product differentiation and specialization by firms (Krugman 1981; Rosen 1974), the sorting of workers into occupations and industries (Heckman and Scheinkman 1987; Miller 1984; Roy 1951), educational investment decisions (Willis and Rosen 1979), immigration (Borjas 1987), and the division of labor within households (Becker 1991).

Yet, to what extent individuals' *social* interactions are governed by comparative advantage remains largely unknown. At an abstract level, social interactions involve a decision about what to do with whom. We note that similar decision problems arise in many other economic settings, such as the choice of an occupation or industry, and go on to explore the role of comparative advantage in social interactions—both theoretically and empirically.

Our analysis begins by developing a theory of social interactions that builds on Roy's classic account of self-selection in the labor market (Roy 1951). To model the endogeneity of social contacts, we posit the existence of an implicit price mechanism in the "market for peers." In our simple theoretical framework, agents care only about social status from membership in peer groups. In equilibrium, heterogeneity in ability leads to individuals selecting into groups based on comparative, rather than absolute, advantage. There are no intrinsic externalities built into the model. When comparative advantage is the guiding principle of peer group formation, an individual's behavior is an equilibrium outcome. It depends on where in the ability distribution she falls, and on the shadow prices that clear the social market. Put differently, selection into peer groups is determined by the scarcity of various skills, and peer effects arise due to the endogenous sorting of agents into peer groups *within* a social setting. An important and novel prediction of our model is that a student's academic achievement and problem behaviors depend on her ordinal rank among her peers.[1]

We document the impact of rank in three different data sets.[2] The first one comes from

[1] Throughout the paper, ordinal rank will be used to refer to a student's percentile in a group.

[2] Following our working paper in 2011 (Cicala et al. 2011), others have also documented a relationship between ordinal rank and student outcomes. Murphy and Weinhardt (2014) use administrative data on students in the U.K. to show that rank in primary school correlates with their secondary school achievement. Their empirical approach mirrors the one we take with the NYCPS and NELS data, and their results replicate

a randomized controlled trial in Kenya, and was collected by Duflo et al. (2011). In 2005, 121 Kenyan primary schools with a single first-grade class received additional funds to hire another teacher and create a second section. In 61 of these schools, students were randomized into classrooms. In the remaining 60 schools, students were assigned to sections based on initial achievement. The intervention lasted 18 months. Relying on the experimentally induced variation in the within-classroom rank of students with equal baseline tests scores, we show that increasing a student's rank by fifty percentiles boosts test scores at endline by about .2 standard deviations.

To provide additional evidence on the relationship between ordinal rank and student outcomes, we use administrative data from New York City Public Schools (NYCPS). Our research design for these data exploits transitions from elementary to middle school, i.e., from fifth to sixth grade. We estimate that a fifty percentile decrease in rank among schoolmates is associated with roughly a 2.5 percentage point increase in the probability of a serious behavioral incident—about 32% of its mean. To account for systematic sorting into schools as well as potential issues of reverse causality, we use students' hypothetical change in rank if they had attended the school for which they were zoned as an instrument for the actual change in rank based on the school they chose to attend. Although our IV estimates are less precise, they are qualitatively very similar to their OLS counterparts.

Our third data set is the National Educational Longitudinal Study (NELS). NELS allows us to relate the same student's behavior in different classrooms to a proxy for her course specific rank. We show that a fifty percentile decline in rank across classes is associated with a ten percentage point increase in the probability that the teacher reports behavioral problems in the course for which she has the lower rank (relative to a basis of 40%).

Consistent with our comparative advantage approach to social interactions, the evidence from three different datasets and research designs suggests that students' ordinal rank exerts a significant effect on their achievement and behavior—even conditional on standard measures of peer quality.[3] Yet, these data cannot rule out other mechanisms. For instance, if

ours, at least qualitatively. Elsner and Isphording (2015) use data from the National Longitudinal Study of Adolescent Health (AddHealth) and exploit within-school differences in the ability distribution of cohorts. The results suggest that ordinal rank increases high school completion and college enrollment. Tincani (2014, 2015) explores the effect of rank when students *intrinsically* care about their ordinal ranking, i.e., when rank directly enters the utility function.

[3] Our findings are also related to an emerging literature on the economic effects of *relative* incomes. Luttmer (2005) and Card et al. (2012), for instance, demonstrate that own well-being and satisfaction depend negatively on the earnings of neighbors and coworkers. Bertrand et al. (2015) show that spouses' relative incomes affect marriage formation and the division of household production. Charles et al. (2009) argue that conspicuous consumption serves as a costly signal of economic position. Lastly, Kuziemko et al. (2014) provide experimental evidence to suggest that people are "last-place averse" and that low-income individuals oppose redistribution because it disproportionately benefits those ranking just below them.

teachers always target the top of the class, then higher ranked students would benefit from more appropriate instruction and thus experience an increase in test scores. Lower ranked students may receive less attention and act out instead. While such an explanation would not invalidate our empirical result that ordinal rank matters for student outcomes, it illustrates the multiplicity of plausible channels through which the effect of rank might operate.

Finding direct evidence in support of comparative advantage in social interactions is challenging—in large part because neither the shadow prices that clear the social market nor students' comparative advantage are directly observable in standard data sets. Similarly, given the potential number of alternative theories that predict a relationship between ordinal rank and economic outcomes, and the data requirements associated with testing each, it is unclear how one could rule out all, or even most, of them.[4]

Thus, rather than trying to pin down the share of the relationship between rank and outcomes that is attributable to a particular mechanism, we pursue the more modest goal of providing additional evidence that some portion of the effect of rank on behavior is due to comparative advantage—as this is the precise mechanism explored in our model.

To this end, we conducted an experiment. Between February 2015 and May 2015, we recruited nearly six hundred children from two public middle schools in Houston, TX and incentivized them to "solve" mazes in a custom-made computer game. According to our conversations with principals and teachers, children at these schools have extensive experience playing comparable games on their phones or even on the schools' computers. By embedding our experiment in a natural context, we hope to replicate as many of the myriad situational factors that may affect students' behavior as we possibly can, while maintaining enough experimental control to identify causal effects.

During the initial stage of the game, students were asked to solve a common set of mazes in order to establish a baseline measure of ability. The software then publicly revealed the ordinal rank as well as the cardinal performance of all participants in the same experimental session—similar to the scoreboard in many popular video games. In the next stage, all students were afforded the opportunity to practice solving mazes at a fixed, randomly determined cost per maze. In addition, children in the treatment group could pay to "slime" the screen of any practicing peer. Sliming another student's screen carried no monetary benefit, but it blocked a portion of the maze on which the respective participant was working, thereby negating the benefits of practicing for her. In the third and last stage of the game, students

[4]For instance, to implement a convincing test of the teacher channel, we would need objective information on how much effort and attention teachers place on students at every part of the ability distribution. We are unaware of such data. In Appendix C, we report results from a partial test of the hypothesis that teacher behavior varies with students' rank. Specifically, we test whether teachers' perception of their students' ability depends rank. Conditional on actual test scores, we find no evidence that this is the case.

were asked to solve more difficult mazes and were rewarded with a piece rate for each one they successfully completed. At no point during the experiment did monetary payoffs depend on ordinal rank.

Among children in the control group we observe that, conditional on actual performance in the first stage, ordinal rank is negatively correlated with willingness to practice. Put differently, children who were paired with better peers and, therefore, find themselves closer to the bottom of the distribution are, if anything, more likely to invest in becoming better at solving mazes. This is not the case in the treatment group. Being able to publicly "slime" their peers, lower-ranked students substitute *away* from practicing and pay to disrupt others instead.

Given that children were randomly assigned to either treatment or control, our experimental design permits us to test the hypothesis that the opportunity to engage in a second, disruptive activity exerts a disproportionate effect on lower-ranked children. That is, by giving students the choice between two different activities we allow for comparative advantage to affect social interactions in the treatment group, but not the control. Perversely, the very children who would ordinarily practice more to overcome their relative disadvantage chose to "act out" instead. This suggests that students' behavior depends, in part, on their comparative advantage within narrowly defined social settings.

Although it is difficult to generalize from the findings of any given experiment—and perhaps even more so in our case—we note that our experimental design holds many, if not all, candidate variables for the effect of rank on student behavior fixed. For instance, the experimental results cannot be due to a change in students' cognitive or noncognitive skills, teacher conduct, or environmental influences. This does not necessarily mean that these factors are irrelevant for explaining the patterns that we document in the real-world data. It does, however, suggest that comparative advantage is one of potentially several channels through which ordinal rank affects outcomes.

The remainder of the paper is organized as follows. The next section develops a formal model of comparative advantage in social interactions. Section 3 presents evidence from three data sets that shows how students' test scores and behavior depend on their ordinal rank. Section 4 describes an experiment designed to test the key mechanism highlighted in our model. The final section concludes with a brief discussion of the broader implications of a comparative advantage approach to social interactions.[5]

[5]There are five appendices. Appendix A considers the implications of comparative advantage for predicting the efficacy of social interventions ex ante. Appendix B discusses identification of traditional peer effects in the presence of comparative advantage. Appendix C explores whether teacher behavior can explain the relationship between ordinal rank and outcomes. In Appendices D and E we describe the data used in our analysis and provide further details regarding the implementation of our experiment. All appendices are

2. A Comparative Advantage Theory of Social Interactions

The model we propose in this paper is a simplified version of the well-known multi-sector choice problem, building upon impressive literatures designed to understand the evolution of earnings, the hedonic pricing of skills, and the assignment of workers to firms (e.g., Heckman and Sedlacek 1985; Murphy 1986; Rosen 1974; Roy 1951; Sattinger 1979). The novelty of our approach lies in the application of these classic methods to develop a theory of social interactions where contacts within a social market are endogenous and peer effects arise due to the sorting of agents within narrowly defined social settings.

2.1. *Basic Building Blocks*

Let there be a continuum of agents with unit mass. Every agent is endowed with one unit of non-transferable time. There are two social activities in which agents can engage with their peers: studying or mischief. These activities are exclusive and undertaken by separate social groups: "nerds" and "troublemakers." Agents acquire social status from membership in peer groups. How much status membership in group $j = N, T$ conveys depends on the effective group size, L_j, and other exogenously given factors, which we label capital, K_j. We allow capital to broadly represent any non-human input into groups' activities, such as the availability of textbooks and sharp scissors, the quantity of policing, or school and neighborhood quality more generally.

Agents are heterogeneous along two dimensions. Their varying size and strength yield differences in the ability to cause trouble, whereas heterogeneity in cognitive ability implies differences in their ability to be a true nerd. Let $\sigma_N(i)$ denote the effective units of "nerdiness" that agent i is capable of contributing to the group (e.g., expertise in differential geometry). Analogously, agent i's troublemaking ability is given by $\sigma_T(i)$. For simplicity, we assume that agents are solely interested in maximizing their social status

$$(1) \qquad U(i) = \max\{s_N\sigma_N(i), s_T\sigma_T(i)\},$$

where the shadow prices s_N and s_T denote the (endogenously determined) status per effective unit of nerd and troublemaking ability, respectively. Thus, total status from membership in group j is given by $s_j\sigma_j(i)$. Note, there are no explicit externalities built into agent's utility. Conditional on "prices," the behavior of others has no influence on own decisions—as in analyses of traditional markets. The key assumption in equation (1) is that, all else equal, "nerdier" individuals, i.e. those with higher $\sigma_N(i)$, will derive more utility from joining the nerd sector than agents with less nerd ability.

provided on the authors' websites.

In Cicala et al. (2011), we allow for more general utility functions (e.g., individuals care about more than just social status). Here, however, we present a very simple and parsimonious model in order to demonstrate how comparative advantage in social interactions can produce "peer effects." It is important to note that the main results continue to hold as long as the benefits from joining a particular group are increasing in the respective dimension of ability, so that sorting into peer groups is at least partially determined by comparative advantage.[6]

Individuals maximize their social status by choosing either the nerd or troublemaking group according to a simple cut-off rule. Let $\sigma(i) \equiv \frac{\sigma_N(i)}{\sigma_T(i)}$ denote agent i's skill as a nerd relative to that as a troublemaker, and order agents such that $\sigma'(i) \geq 0$. The agent indifferent between the two sectors, i^*, has a skill ratio of

$$\sigma(i^*) = \frac{s_T}{s_N}. \tag{2}$$

By individual optimization, all agents with index $i \geq i^*$ join forces with the nerds, and individuals with $i < i^*$ become troublemakers. In our Roy model of social interactions, comparative (rather than absolute) advantage determines an individual's choice of peer group.

As a result, the supply of skills to both groups is given by:

$$L_N^* = \int_{i^*}^{1} \sigma_N(q)\, dq \tag{3}$$

$$L_T^* = \int_{0}^{i^*} \sigma_T(q)\, dq. \tag{4}$$

Equilibrium, however, also depends on the endogenously determined shadow prices, and, therefore, on the relationship between social status, s_j, and effective group size, L_j. Since we don't have strong priors as to the functional relationship between s_j and L_j, we explore two different possibilities.

2.2. *Case I: Social Status Decreases with Effective Group Size*

In a traditional Roy model it is usually assumed that labor exhibits diminishing returns to scale. If, for instance, increasing the number of troublemakers does more to increase the probability of getting caught than of winning a fight, then social status may be decreasing in effective group size, L_T. Similarly, intelligence may confer greater status when this

[6] In Cicala et al. (2011), we also extend the basic model to allow for many groups and n-dimensional skill (Heckman and Scheinkman 1987), hierarchies (Rosen 1982), and show that the basic results of our model hold when the sectoral choice problem is cast in a general social multiplier framework (Becker and Murphy 2000; Glaeser et al. 2003).

skill is scarce and not readily available within a social environment. Hence, we first assume $\frac{\partial s_j(L_j,K_j)}{\partial L_j} < 0$.

In equilibrium, market clearing and equation (2) yield the following condition:

$$\delta\left(i^*\right) \equiv \frac{s_T(L_T(i^*), K_T)}{s_N(L_N\left(i^*\right), K_N)} = \sigma\left(i^*\right), \tag{5}$$

where $\delta\left(i\right)$ denotes the ratio of social status in both sectors when the marginal agent is i. Since $\delta'\left(i\right) < 0$ for all i, the relative "price" schedule in the market for peers is strictly downward sloping. To see this, note that equations (3) and (4) respectively imply $\frac{dL_N^*}{di} < 0$ and $\frac{dL_T^*}{di} > 0$, which causes the absolute as well as the relative status of troublemakers to decrease as agents shift from the nerd into the troublemaking group.

We can now describe equilibrium graphically. Figure 1 depicts the situation when social status is decreasing in effective group size. As described above, it features upward sloping "supply" and a downward sloping "price" schedule. There is a unique equilibrium at i^* with market clearing relative status, $\frac{s_T}{s_N}^*$. All individuals with $i < i^*$ select into the troublemaker group and individuals with $i \geq i^*$ choose to become nerds.

Social status may also depend on the specifics of the exogenously given environment. To allow for this possibility we let s_j depend not only on group size, but also on any number of environmental inputs K_j. Suppose $\frac{\partial s_j(L_j,K_j)}{\partial K_j} > 0$ and imagine a shift in the "capital" available to the troublemaking sector—less police surveillance, an increase in the availability of drugs, weapons, or alcohol. Holding everything else constant an increase in troublemakers' productive capital, K_T, is represented by an outward shift of the δ-schedule, which results in higher status for troublemakers and, therefore, fewer nerds. A decrease in the "capital" available to troublemakers has the opposite effects. Thus, with respect to features of the physical environment our Roy model of social interactions features conventional predictions.

Comparative statics with respect to the skill distribution, however, can be quite counterintuitive. Consider, for instance, an increase in nerd skill among the population holding troublemaking ability fixed. First, an increase in agents' nerdiness raises σ_N relative to σ_T and thus shifts the "supply" curve inward. Second, the equilibrium price schedule shifts outwards due to the fact that with more academically able peers there will be more effective units of nerd skill supplied at any i, which lowers s_N in equation (5). While both shifts lead to an unambiguous rise in the relative wage of troublemakers, the effect on quantities is indeterminate.

Figure 2 illustrates this point. In the left panel, social status is largely irresponsive, despite the large shift in the distribution of relative skill. On net the shift in $\delta(i)$ outweighs that in

$\sigma(i)$, which results in an expansion of the nerd group. In the panel on the right, however, nerds' social status drops rapidly with group size, leading to a much larger outward shift of the price schedule, $\delta_1(i)$.[7] In the new equilibrium, *fewer* agents choose to become nerds, despite the fact that everyone has higher nerd ability than before.

2.3. *Case II: Social Status Increases with Effective Group Size*

Much of the literature on social interactions assumes that the marginal utility of a social activity is increasing in overall participation. In models with a social multiplier, peer effects arise because an individual's utility from taking a particular action increases in the number of agents in her reference group who behave in the same way (e.g., Becker and Murphy 2000; Glaeser et al. 2003). By assuming that social status rises with effective group size, i.e. $\frac{\partial s_j(L_j,K_j)}{\partial L_j} > 0$, our comparative advantage theory can replicate these models.

As in the case of decreasing status, the equilibrium price schedule continues to be given by

$$\delta(i^*) = \frac{s_T(L_T(i^*), K_T)}{s_N(L_N(i^*), K_N)}.$$

The key difference when social status is increasing in group size is that $\delta'(i) > 0$ and that there may exist multiple equilibria.[8]

Figure 3 depicts such a scenario. Here, increasing status yields equilibria at the origin, at i^*, and at i^{**}. But only equilibria in which the price schedule, δ, intersects the "supply" curve from above are locally stable.[9] Given the existence of multiple equilibria, the Roy model approach can rationalize starkly different behaviors of individuals in observationally similar environments. Moreover, small changes in the environment may lead to large behavioral responses. Consider, for instance, a decrease in the inputs available to the troublemaking group and assume that $\frac{\partial s_j(L_j,K_j)}{\partial K_j} > 0$. As shown in Figure 4, the initial decrease in K_T lowers the relative status of troublemakers, and, thus, causes troublemakers close to the initial

[7] Note that the effective size of the nerd group is integrated on $[i^*, 1]$, so that, for any i^*, the shift in δ will be larger the more concentrated a given increase in the distribution of σ_N is in the upper end of the distribution.

[8] When social status is independent of group size, i.e. $\frac{\partial s_j(L_j,K_j)}{\partial L_j} = 0$, then the horizontal price schedule intersects the "supply" curve exactly once, leading to a single equilibrium with traditional comparative statics.

[9] To see this, consider the adjustment process following a small shock to "prices." From the initial equilibrium at i^{**}, a small decrease in relative prices (along the δ-schedule) will lead to agents flowing out of the troublemaking and into the nerd sector, which will cause relative status to decline further and lead to even more agents switching sectors. The process continues until the market reaches a new equilibrium at the origin. Conversely, a small increase in relative prices (along the δ-schedule) will lead to agents flowing into the troublemaking sector. This causes relative status to increase even more, thereby inducing more nerds to become troublemakers until the market reaches equilibrium at i^*. Similar reasoning shows that the equilibrium at i^{**} is stable.

equilibrium to switch peer groups. The decrease in the size of the troublemaking group further lowers s_T, which leads to an even larger outflux, and so on. Analogous to traditional models of a social multiplier, agents' behavior may become very elastic when their choices are complements.

2.4. *Empirical Implications*

In contrast to conventional analyses, however, our theory allows for heterogeneity in agents' skill endowments and, therefore, for different behavior in a common environment. This form of heterogeneity is necessary to explore the idea of comparative advantage in social interactions, and it is at the root of the key testable prediction that differentiates our approach from existing models.

Consider, again, Figures 1 and 3. An individual at the bottom of the skill distribution will be more likely to join forces with the troublemakers than an agent in the right tail of the distribution—irrespective of whether social status is increasing or decreasing in effective group size. Put differently, within any social market and for any set of shadow prices, individuals in the upper tail of the relative skill distribution (i.e., those with high i) have more to gain from joining the nerd group than those in the lower tail (i.e., those with low i). As a consequence, agents' behavior correlates with their ordinal rank in a narrowly defined social environment.

For a large class of skill distributions, a ranking based on individuals' academic abilities alone will be a good proxy for one in terms on their (unobserved) relative skills, i.e. $\sigma(i)$.[10] If correct, then our theory predicts that students' scholastic achievement and their proclivity to "act out" are related to their *ordinal* rank in the ability distribution.

Anecdotally, the phenomenon that the behavior of children varies with their relative standing has been observed among some programs for gifted minority youth held at MIT each summer. These programs attract a subset of black children who are among the best and brightest in their schools. At MIT, however, they interact with more academically able peers, leading them to engage in a wide range of problem behaviors (Suskind 1998). Similarly, in his memoir, Canada (1995) speculates that even the most violent youth in Boston would only be mediocre fighters in the South Bronx and, therefore, be forced to change their ways.

Broadly summarizing, the comparative advantage approach to social interactions delivers a novel, testable prediction: students who fall near the bottom of their reference group

[10] Formally, ability-based rank will be positively correlated with rank in terms of relative nerd and troubmaking skill whenver $\sigma_N(i) \geq \sigma_N(j)$ implies that $\mathbb{E}[\sigma(i)] \geq \mathbb{E}[\sigma(j)]$. This condition holds trivially if σ_N and σ_T are independently distributed. It is also satisfied if nerd and troublemaking ability are negatively or not "too positively" correlated. Only if σ_T were to increase proportionately more than σ_N, would ability-based rank fail to be a valid proxy.

should be more prone to behavioral problems than their equally able counterparts who find themselves closer to the top in another environment. Conversely, the latter should display higher academic achievement than the former. In the following section, we demonstrate that this implication of our theory is borne out in three very different data sets.

3. Ordinal Rank and Peer Effects

The perfect data to test our theory would span multiple social markets—say, schools or classrooms—and contain information on shadow prices, individuals' choices of peer groups as well as all of their skills. With such data in hand we could directly test whether comparative advantage determines behavior by comparing social status across groups and relating it to agents' choices. However, in the absence of any information on status, and in lieu of imposing restrictive parametric assumptions, we confine ourselves to providing reduced form evidence which shows that individual behavior does depend on ordinal rank. That is, in the spirit of Friedman (1953) we test a stark prediction of our approach—one that does not follow from standard theories of social interactions.

Specifically, we analyze the impact of rank in three different data sets: a randomized controlled trial involving more than one hundred primary schools in Kenya, administrative data from New York City Public Schools (NYCPS) from the 2003-04 through 2008-09 school years, and survey data from the National Education Longitudinal Study of 1988 (NELS).[11]

3.1. *Evidence from Primary Schools in Kenya*

Our first piece of evidence comes from the Extra Teacher Provision (ETP) intervention by Duflo et al. (2011).[12] Starting in May 2005, ETP provided 121 Kenyan primary schools that had a single first-grade class with additional funds to hire an extra teacher and create a second section. In 61 randomly selected "tracking schools," students were assigned to sections based on scores on exams administered by the schools prior to the intervention. Students above the median were grouped in one section, and those below the median in another one. In the remaining 60 "non-tracking schools," students were randomized into sections. After assignment of students to sections, each of a school's two sections was also randomly assigned to either a civil service teacher or to one hired on a contractual basis.

[11] We also tested for an impact of ordinal rank in data from Project STAR (Krueger 1999; Word et al. 1990) and MTO (Kling et al. 2005, 2007). Although the point estimates were consistent with the model's predictions, large standard errors prevented sharp conclusions.

[12] Since a full description of the experiment is available in the aforementioned paper, we restate only the intervention's most salient features here and refer the interested reader to Duflo et al. (2011) for additional details.

This intervention spanned 18 months.[13]

Table 1 displays summary statistics for the 121 schools in the sample of Duflo et al. (2011). Due to random assignment, tracking and non-tracking schools look very similar on pre-treatment observable characteristics. The same is true for students assigned to either the contract or government teacher section within non-tracking schools. Within tracking schools, students assigned to the top section have on average .81 standard deviation higher test scores and are almost .4 years older than their low-ability counterparts.

Duflo et al. (2011) demonstrate that tracking increased the subsequent test scores of *all* students, regardless of their initial place in the distribution. The authors rationalize this finding with high-ability students benefitting primarily from positive spillover effects due to more able peers, whereas for students in the low-ability section the direct effect of worse peers is more than outweighed by better targeted instruction.

Given random assignment of students to classrooms, the non-tracking schools in the experiment Duflo et al. (2011) provide an ideal testing ground for the impact of ordinal rank on student outcomes. In what follows we exploit the experimentally generated variation in the within-section rank of children with equal ability to estimate the causal effect of rank on academic achievement. As Duflo et al. (2011), we base our results on the initial random assignment of all students who attended first grade in May 2005.[14]

Specifically, we implement the empirical setup of Duflo et al. (2011), but add a student's ordinal rank to the following linear model:

$$(6) \qquad y_i = \varphi r_i + \mathbf{X}_i'\beta + \alpha\overline{y}_{-i} + \mathbf{T}_i'\theta + \epsilon_i,$$

where y_i denotes individual i's standardized total test score at endline, r_i is her section-specific rank (i.e., her percentile in the distribution of pre-treatment test scores), and $\mathbf{X}_i$ is a vector of individual controls including the baseline score and its square, gender, age, etc. $\overline{y}_{-i}$ represents the mean standardized baseline score of i's peers, and $\mathbf{T}_i$ marks a vector of treatment indicators. In alternative specifications, we also include school or section fixed effects, which help to account for unobserved heterogeneity at the school or section level.

Table 2 presents the results from estimating equation (6) using ordinary least squares. To estimate the impact of rank as cleanly as possible, the upper panel restricts attention to nontracking schools, i.e. to students for whom, conditional on test scores at baseline, variation in ordinal rank is purely random. For completeness, in the lower panel of Table

[13] Across five unannounced vistis to each school, both section were found to be combined 14.4% of the time in nontracking schools and 9.7% of the time in tracking schools. When sections were not combined, 92% of students in nontracking schools and 96% of students in tracking schools respected their initial assignment.

[14] About 21% of students in tracking schools and 23% of those in nontracking schools repeated first grade and participated in the program for only the first year.

2 we present results for all students with nonmissing baseline scores, including students in tracking schools. Moving from the left to the right within each panel, the set of controls grows.

As reported in Duflo et al. (2011), contract teachers have a positive impact on test scores. More importantly for our purposes, there is a positive relationship between students' ordinal rank and their academic achievement. With one exception the point estimates are statistically significant, and they are always economically large. Critically, compared to its baseline value in the first column, the estimated impact of rank on test scores actually increases with the inclusion of additional controls, such as peers' mean test score, peers' mean test score interacted with a student's own position in the ability distribution, age, gender, etc. Moreover, the point estimate is also robust to using only within-school or within-section variation as sources of identification. Taking the lowest estimate in the upper panel of Table 2 at face value, a fifty percentile increase in rank increases test scores at endline by about .2 standard deviations.[15]

The coefficients in the lower panel show that the estimated impact of ordinal rank on test scores is qualitatively robust to including all students in the experiment of Duflo et al. (2011), though the point estimates do decline somewhat. Including children in tracking schools has the benefit of introducing more variation in the section-specific ordinal rank of similar students (i.e., children just above and below the cutoff for tracking). As a consequence, the point estimates become more precise.

The downside of including students in tracking schools is that, by design, the expected quality of their peers is not constant. Comparing two students with nearly identical baseline test scores, the one assigned to the "top section" will, on average, have more able peers. But she will also have a lower section-specific rank. Thus, for comparable children in tracking schools ordinal rank will be negatively correlated with peer ability. Unless peers' mean test scores (or higher order polynomials thereof) adequately control for their unobserved quality, there is reason to believe that the estimates in the lower panel of Table 2 are *downward* biased. The fact that they continue to be economically large suggests that ordinal rank exerts nontrivial effects.[16]

It is also important to note that by relying on experimentally induced variation in rank, we sidestep the possibility that measurement error in baseline test scores is responsible for the observed patterns. Since test scores are only a noisy measure of ability, one might worry

[15]The data of Duflo et al. (2011) also contain component test scores for math and literacy. Our results are qualitatively very similar when using these instead of total test scores, but we note that the impact of rank appears to be stronger for math scores.

[16]Additional results (available from the authors upon request) show that the estimated effect of rank remains almost unchanged when we control for higher order polynomials of peers' mean test score.

that a student's rank proxies for unmeasured skills, and that it is for this reason related to future achievement. Yet, in the experimental data of Duflo et al. (2011) this cannot be the case because variation in the rank of students with equal baseline test scores is randomly generated. Put differently, random assignment of students in non-tracking schools guarantees that students who attend the same school and have equal baseline test scores but a different within-classroom rank have, on average, the same unmeasured ability. Taken together, the evidence from this randomized controlled trial suggests that, even conditional on standard measures of peer quality, ordinal rank has an economically meaningful impact on children's academic achievement.[17]

Although the intervention of Duflo et al. (2011) provides us with exogenous variation to test for an impact of rank on achievement, it does not come without drawbacks. As with any experiment, one may wonder about external validity. Even more importantly for our purposes, the data do not contain measures of student behavior, which prevents us from probing the prediction that ordinal rank also affects problem behaviors.

3.2. *Evidence from New York City Public Schools*

To ameliorate these shortcomings, we now turn to administrative data for all students in New York City Public Schools (NYCPS)—the largest school district in the US. The NYCPS data contain student-level information on approximately 1.1 million students per year across the five boroughs of New York City. Our data span the 2003-04 to 2008-09 school years and include student race, gender, free and reduced-price lunch eligibility, behavior, attendance, and matriculation with course grades for all students, as well as state math and English/Language Arts (ELA) test scores for students in grades three through eight. Summary statistics for the variables we use in our core specifications are displayed in Table 3.

Our research design for these data relies on transitions from elementary to middle school, i.e., from fifth to sixth grade. During this transition students typically move from small, local elementary schools to larger middle schools, which disrupts ordinal rank when the feeder schools are heterogeneous. Specifically, to estimate the impact of ordinal rank, we relate *changes* in the behavior of equally able children from the same elementary school to *changes* in rank induced by a switch to different middle schools. Our empirical approach, therefore, accounts for students' inherent tendencies to act out. We first exploit the sheer size of the NYCPS data and estimate semiparametric models, which allow us to explore

[17] Moreover, the results above suggest that students in the lower half of the achievement distribution may have benefitted from tracking at least in part because their ordinal rank increased. Rank, therefore, provides a way to rationalize the results of Duflo et al. (2011) without relying on non-convexities in teachers' payoffs.

potential nonlinearities in the relationship between rank and behavior. Finding little to no evidence of nonlinearities, we then address the potential endogeneity of changes in rank via an instrumental variables strategy based on school zoning regulations.

In order to examine the functional relationship between rank and behavior we estimate semiparametric specifications of the following form

$$(7) \qquad \Delta y_i = f(\Delta r_i) + \mathbf{X}_i'\beta + School_i + Year_i + \epsilon_i,$$

while restricting attention to the set of students who change schools in the transition from fifth to sixth grade. Our behavioral measure, y_i, in each year is an indicator equal to one if a student has at least one reported behavioral incident from that year and zero otherwise. Hence, $\Delta y_i \in \{-1, 0, 1\}$. The three most common behavioral incidents in our data are "engaging in an altercation or physically aggressive behavior with other student(s)," "behaving in a manner that disrupts the educational process (horseplay)," or "engaging in verbally rude or disrespectful behavior / insubordination."

A student's rank in fifth grade is the student's percentile ranking based on her achievement on the New York State exam relative to other students who are in the same school in fifth grade.[18] We also compute each student's position relative to her peers in the sixth-grade school, and denote the difference between these two rankings Δr_i. Results are reported using both math and ELA scores to compute the change in percentile.

We include a standard set of controls, i.e. $\mathbf{X}_i$, consisting of the test score in the same subject from the previous year, an exhaustive set of race dummies, sex, free lunch eligibility, English Language Learner (ELL) status, and special education designation. Using these covariates we attempt to control for factors that plausibly influence changes in behavior and might be correlated with rank. Finally, we include year fixed effects and school fixed effects (for both a student's elementary and middle school). By including school fixed effects we account for average peer quality as well as for the fact that schools might have heterogeneous propensities to classify the same demeanor as a behavioral incident.

Our semiparametric estimates of the link between changes in rank and changes in behavior are displayed in Figure 5. Independent of whether we calculate rank based on ELA or math scores, the behavior of students whose rank decreases in going from elementary to middle school worsens significantly compared to students whose relative standing improves. A student experiencing a 50 percentile decline in rank is approximately five percentage points more likely to have a behavioral incident on record than a student whose rank improves by 50 percentiles—with the estimated effect being slightly larger if we calculate rank based on

[18] The state math and ELA tests are high-stakes exams conducted in the winters of third through eighth grade. For additional information on these tests, see the Data Appendix.

math scores than if we do so based on ELA scores. Given sample means of 8.7% for sixth grade and 4.9% for fifth grade, our estimates are nontrivial in size.[19]

Although the NYCPS data allow us to control for students' natural proclivities to cause trouble by relating *changes* in behavior to *changes* in rank induced by the transition from elementary to middle school, there exists the possibility that estimates of equation (7) are driven by reverse causality. That is, behavioral problems during sixth grade might have caused changes in test scores and, therefore, class rank. Another concern is systematic choice of school. Students who chose an academically less challenging middle school might have experienced less of an increase in behavioral problems, even if their rank had not improved. To address these issues, we also estimate two-stage least squares (2SLS) specifications in which we instrument for a student's change in rank with the predicted change based on the schools they were zoned to attend (given their residential address). Specifically, we estimate the following linear model:

$$\Delta y_i = \varphi \Delta r_i + \mathbf{X}_i' \beta + School_i + Year_i + \epsilon_i \,, \tag{8}$$

where the first stage is given by

$$\Delta r_i = \delta \widehat{\Delta r_i} + \mathbf{X}_i' \gamma + School_i + Year_i + \nu_i,$$

and $\widehat{\Delta r_i}$ denotes student i's counterfactual change in rank at the *beginning* of sixth grade (using *fifth* grade tests scores) had all students attended the schools for which they were zoned. In symbols, let $a_{i,t-1}$ denote student i's test score in fifth grade and let $\text{rank}_I\left(a_{i,t-1}\right)$ be the percentile ranking of a student with score $a_{i,t-1}$ among the set of students I, given their respective test scores at $t-1$. Then,

$$\widehat{\Delta r_i} \equiv \underset{S_{i,t}}{\text{rank}}\left(a_{i,t-1}\right) - \underset{S_{i,t-1}}{\text{rank}}\left(a_{i,t-1}\right),$$

where $S_{i,t-1}$ and $S_{i,t}$ are the sets of students who are zoned for the same elementary and middle school as i, respectively. Intuitively, our IV approach compares observationally identical students from the same elementary school who experience a differential change in rank because school zoning regulations led them to attend different middle schools.

Table 4 presents the resulting 2SLS estimates of the effect of school rank on behavior, as well as the corresponding OLS ones for comparison. In the upper panel we use ELA scores

[19]Interestingly, Figure 5 shows that the relationship between rank and behavior is almost linear, except for in the extremes, where there is less data to deliver precise estimates. This suggests that simple linear models may provide decent approximations to the true functional relationship.

to construct rank, whereas math scores are used in the lower one. Based on the OLS point estimates, one would expect a student experiencing a 50 percentile decline in rank to be three to five percentage points more likely to have a behavioral incident on record than a student whose rank improves by 50 percentiles—consistent with our previous semiparametric results.

Due to the large number of observations, our OLS estimates are very precise. Unfortunately, this is not the case when we estimate equation (8) by 2SLS. Although the first stage F-statistic is well above conventional critical values (Stock and Yogo 2005), our instrument explains little residual variation in the excluded variable, as evidenced by small values of Shea's R^2 (Shea 1997). One potential explanation for this is that only 45.3% (53.9%) of students attend the middle (elementary) school for which they are zoned.

Nevertheless, not including school fixed effects, the 2SLS estimates are at least as large as their OLS counterparts and statistically significant. If we include school fixed effects the standard errors almost double and the point estimates cease to be significant. They do, however, continue to be negative and economically very large.

In order to investigate whether these effects persist beyond sixth grade, we have replicated the analysis in Table 4, focusing on the change in behavior from fifth to eighth grade instead. Table 5 presents the results. Interestingly, all point estimates are negative and economically meaningful. Six out of the eight estimates are even larger than those in the previous table. Although the 2SLS results are, again, fairly imprecise, the sum of the evidence suggests that behavioral effects from changes in ordinal rank do not dissipate over time.

Taken together, the NYCPS data point to an economically significant effect of ordinal rank on problem behaviors.

3.3. *Evidence from the National Educational Longitudinal Study*

Our third data set is the National Educational Longitudinal Study (NELS). NELS was initiated in 1988 with a nationally representative sample of 24,599 eighth graders, who were then resurveyed in 1990, 1992, 1994, and 2000. The available information on these students covers a wide range of topics, including school, work, and home experiences, educational resources and support, neighborhood characteristics, educational and occupational aspirations, as well as the perceptions of other students. For the first three waves, students completed achievement tests in reading, social studies, mathematics and science. The data set also includes survey results from teachers, parents, and school administrators. Table 6 displays summary statistics for all variables we use in our analysis.

We examine NELS data from 1988 and 1990, when students were in eighth and tenth grade. An important limitation of the NELS data is that only 25 students per school were surveyed, yielding a noisy measure of rank. To reduce the impact of measurement error, we limit our sample to students in classrooms with at least five observations.[20] However, NELS allows us take advantage of the fact that the data include teacher reports on behavior and student self-reported grades from exactly two subjects in the same year. By estimating a model that relates differences in a student's behavior *across classrooms* to differences in her rank, we can implicitly account for students' natural tendencies to cause trouble, and we can rule out that systematic sorting into schools drives our results.

Specifically, we estimate the following specification:

$$\Delta y_i = f(\Delta r_i) + \mathbf{X}_i'\beta + Grade_i + \epsilon_i, \tag{9}$$

where y_i is an indicator for whether the teacher reported that the student had any behavioral problems, and Δy_i refers to the difference in this indicator across subjects within the same year. Teachers were asked whether the student had a problem in any of six different categories: the student performed below his ability, the student did not complete homework, the student was frequently absent, the student was frequently tardy, the student was inattentive, or the student was disruptive. Our indicator variable is equal to one if the teacher reported that the student had at least one of these behavioral problems.[21]

We use students' self-reported grades to compute subject-specific rank r_i and let Δr_i denote the difference in these ranks across subjects within the same year. Our vector of covariates $\mathbf{X}_i$ includes the mean score across subjects from the same year and its square, a complete set of race indicators, sex, English Language Learner status, indicator variables for parents' marital status, parental education, school type (public, Catholic, or other private), and school location (urban, suburban, or rural). Moreover, we include indicator variables for socioeconomic status quartiles, birth year, and birth month. $Grade_i$ marks a grade-level fixed effect.

Figure 6 displays our semiparametric estimate of the relationship between ordinal rank and behavior. As was the case in the NYCPS data, we find that changes in a student's rank within a narrowly defined social setting are related to changes in her behavior. For instance, students whose rank is fifty percentiles lower in English class than in Math class are estimated to be approximately ten percentage points more likely to act out in the former than the latter—relative to a mean of about 44%. Rank, therefore, appears to exert a significant influence on behavior.[22]

[20] We obtain qualitatively identical results for alternative threshold levels of ten and one.

[21] It is worth noting that the NELS measure of behavioral problems encompasses a far more benign set of offenses than those typically reported in the NYCPS data set.

[22] Instead of using an indicator variable for whether the teacher reports any behavioral incidents, we have also constructed a summary index of children's behavior by factor analyzing different teacher-reported behaviors. Our results are qualitatively identical for both outcomes. We have also re-estimated equation (9)

3.4. *Discussion*

The findings above suggest that, all else equal, students' test scores decrease and problem behaviors worsen as their relative standing declines. These results are noteworthy not only because they point to a hitherto underexplored source of peer effects, but also because they come from three very different, independent settings. The fact that we find an effect of ordinal rank for primary school children in Kenya as well as for middle and high school students in the U.S. suggests a more general phenomenon.

Although the data are consistent with the idea that comparative advantage shapes social interactions, they cannot rule out other mechanisms. For instance, another explanation for our findings is that teacher behavior depends on the entire distribution of student ability. Suppose that a student's perceived ability matters for how much teachers invest in her. If teachers invest more in students who are thought to be smarter, and if ordinal rank serves as a (noisy) signal about ability, then a teacher-focused explanation is consistent with the finding that test scores increase with adolescents' rank. If lack of teacher attention causes students to act out, then such an explanation can also rationalize why problem behaviors worsen as students' rank declines.

The data requirements to test this alternative hypothesis are very demanding. To implement a convincing test, we would need objective information on how much effort and attention teachers places on students at every part of the ability distribution. We are unaware of such data.[23] Moreover, there are likely additional theories that predict a relationship between rank and student outcomes, and we do not have a principled way to narrow down the set of plausible mechanisms. Thus, rather than trying to estimate the precise share of the relationship between rank and outcomes that is attributable to a particular mechanism, we pursue the more modest goal of providing additional evidence that some portion of the effect of rank is due to comparative advantage.

4. An Experiment to Test the Comparative Advantage Mechanism

To explicitly test the comparative advantage channel, we implemented a "framed" field experiment in the Houston Independent School District.[24] Specifically, we recruited nearly six hundred students from two open-enrollment public middle schools and incentivized them to solve mazes in a custom-made computer game. According to our conversations with principals

using grades from previous waves to construct Δr_i. Although the slope of $f(\cdot)$ is estimated to be negative almost everywhere, large standard errors prevent us from drawing sharp conclusions.

[23] In Appendix C, we report results from a partial test of the teacher behavior hypothesis. Specifically, we test whether teacher perception of student ability depends on ordinal rank. Conditional on actual test scores, we find no evidence that this is the case.

[24] Harrison and List (2004) define a "framed" field experiment as laboratory experiment with a nonstandard subject pool and field context in either the commodity, task, or information set that the subjects can use.

and teachers, children at these schools have extensive experience playing simple video games on their phones or even on the schools' computers. Our decision to embed the experiment in the context of a game reflects the desire to replicate as many as possible of the myriad ways in which situational variables may affect students' behavior. That is, we sought to create an environment in which peer effects arise naturally, all the while maintaining enough experimental control to shed light on the mechanism through which rank affects outcomes.

4.1. *Experimental Design*

Each student participated in one experimental session, which was held in her school's computer lab (see Appendix E for details regarding recruitment, parental consent, implementation logistics, etc., and for a copy of the experimental instructions). Sessions lasted about sixty minutes and included, on average, twenty children from the same school.

In the first stage of the game, students were asked to solve either five or twenty mazes, depending on the experimental session.[25] "Solving" a computerized maze entailed using the arrow keys to steer a cursor from the entrance of the maze to its exit (see Figure 7 for screenshots). Children earned \$.25 per maze that they successfully completed in this stage. All students worked on the same set of mazes and were ranked (among participants in the same session) according to the time it took them to complete the task. The ordinal ranking as well as children's cardinal performance was then displayed on everyone's screens in order to make both common knowledge—similar to the scoreboard feature in many of the most popular video games.

In the second stage, children in the control group were given the opportunity to practice on up to 20 additional mazes at a fixed, randomly determined cost per maze. Students were instructed that ten of these mazes would reappear again in the third stage of the game. Before the software determined the cost of practicing, students were asked for their maximal willingness to pay to see and work on a maze. Children whose willingness to pay exceeded the cost per maze were allowed to practice on as many mazes as they wished, paying for each one as they went along.[26] Participants whose willingness to pay did not exceed the session specific cost were not allowed to practice at all. Our procedure for eliciting students' willingness to pay thus resembled the well-known BDM mechanism (Becker et al. 1964), with the important difference that quantity was not fixed at one. Instead, we allowed students to

[25]When examining the data from sessions in which the first stage involved a different number of mazes, we found no differences in student behavior. We, therefore, pool these data in the analysis below.

[26]Students were told that they could go "go into debt" during this stage of the experiment, i.e., that they could spend more money than they had earned during the previous stage. Any extra spending would be substracted from their earnings in the third stage. At the end of the experiment, no child ended up with negative earnings.

choose quantity knowing the realized price per maze.[27]

Children in the treatment group were also allowed to practice, and the software elicited their willingness to pay for practicing in the same way. In addition, they were asked how much they would be willing to spend in order to "slime" the screen of a peer of their choosing. Sliming another child's screen carried no monetary benefit, but it prevented the other student from practicing by blocking a portion of the maze on which she was working (cf. Figure 7). Both activities were nonrival in the sense that children could practice and slime other participants at the same time. However, students were only allowed to engage in a particular activity if their stated willingness to pay exceeded the respective, randomly determined price. Importantly, if allowed to slime, students could do so as often as they were willing to incur the cost. A ticker publicly displayed who slimed whom in real time.

In the final stage of the game, all children were asked to complete ten mazes in twenty minutes, for a payoff of $3.00 per maze they successfully solved. On average, students earned a total of $29.30, including a $2 show-up fee. Neither students' earnings nor their performace during the last stage of the game were made common knowledge. It is also worth noting that at no point during the experiment did monetary payoffs depend on ordinal rank, and that students were made aware of the payoff structure.

Our experimental design exploits two different sources of random and quasi-random variation. All students in the same experimental session were randomly assigned to either the treatment or the control group. Thus, whether the children in a particular session were faced with the opportunity to disrupt their peers was purely random. Unfortunately, administrative constraints imposed by the schools prevented us from randomizing students into experimental sessions. Specifically, we were only allowed to pull students from their elective or social studies classes, which limited the set of potential peers during the experiment. Since all of our results condition on children's own baseline performance, it is the ability of *other* students in the same session that drives variation in the ordinal rank of observationally similar students. As a consequence, variation in within-session rank is only quasi-random.

Nevertheless, by comparing the relationship between ordinal rank and students' willingness to practice across treatment and control, the experimental design permits us to assess how the *opportunity* to engage in a second, disruptive activity causes investment behavior to change among different sets of students, i.e., lower- versus higher-ranked ones. Put differently, the control group allows us to establish a baseline correlation between rank and willingness to practice. If comparative advantage shapes social interactions, then, when given a chance to be disruptive, lower-ranked students—who have a relative disadvantage at solving mazes—should be disproportionately likely to substitute away from practicing. Hence, the experiment

[27] We chose not to elicit an entire demand curve in order to reduce complexity and simplify the instructions.

allows us to explicitly test the hypothesis of comparative advantage in social interactions.

4.2. *Experimental Results*

Table 7 presents descriptive statistics for the children in our experiment, by treatment and control status. With the exception of grade level, students in the treatment and control group are statistically indistinguishable. While the difference in grade level is economically small, it is statistically significant at the 10%-level. Notwithstanding the fact that a Kolmogorov-Smirnov test is unable to reject the null hypothesis that the p-values in the rightmost column are uniformly distributed on the unit interval ($p = .305$)—as one would expect under truly random assignment—we address the issue of imbalance by presenting results that do and do not condition on covariates. If anything, our findings become *stronger* when we control for observables. In addition, we show in Appendix Table A.1 that the results are qualitatively robust to conducting our analysis within each grade level. Although point estimates disaggregated by grade are far less precise, out of the thirty coefficients in Table A.1 only two change sign compared to our main analysis. It, therefore, seems unlikely that imperfect randomization is driving our results.

Pooling over all 573 students who completed the experiment, Table 8 displays the findings from our "framed" field experiment. The numbers therein correspond to the coefficients on rank ($r_{i,s}$) and rank interacted with a treatment indicator (T_s) in the following econometric model:

$$y_{i,s} = \varphi r_{i,s} + \gamma T_s \times r_{i,s} + \alpha b_i + \mathbf{X}_i' \beta + \mu_s + \varepsilon_{i,s}, \tag{10}$$

where $y_{i,s}$ is the outcome of interest for student i in experimental session s, b_i is her baseline performance at solving mazes, and $\mathbf{X}_i$ denotes a vector of controls, which consists of all covariates that are listed in Table 7. Since the cost of practicing and sliming vary at the level of the experimental session, we also include μ_s, a session fixed effect. To allow for arbitrary forms of correlation in the residuals of children within the same experimental session, all standard errors are clustered at the session level. Given the small number of clusters, we follow the bootstrapping procedure recommended by Cameron et al. (2008) whenever we report p-values for hypothesis tests.

The results in columns (1) and (2) show that, among children in the control group, willingness to pay for practicing is negatively correlated with ordinal rank. That is, conditional on baseline performance, a student at the very bottom of her reference group is willing to spend 10 cents more on practicing a maze than an observationally similar one who ranks at the top of the distribution because she happened to be paired with less able peers. Given a sample mean of 20 cents, this disparity is economically large and statistically significant ($p = .009$).

Interestingly, the correlation between ordinal rank and willingness to practice disappears among students in the treatment group. The difference in the slope estimates between both groups is not only statistically significant ($p = .045$), but the coefficient on the interaction term is of opposite sign and almost as large as that on rank itself. Hence, when children are given the choice between practicing and disrupting their peers, it is no longer the case that lower-ranked students invest more than higher-ranked ones.

The next two columns show that this conclusion is qualitatively robust to examining actual money spent on practicing rather than self-declared willingness to pay. Although the point estimates lose much of their precision—in large part because most students practice on only one or two mazes—the sign pattern of the coefficients is identical to that in columns (1) and (2), and the coefficient on the interaction term continues to be economically large and marginally significant ($p = .090$).

The remaining two columns demonstrate that ordinal rank is *negatively* correlated with how much money children in the treatment group spent on disrupting others. Taking the point estimate in column (6) at face value, a student who is paired with more able peers and, therefore, ranks at the bottom of her peer group spends 39 cents more on sliming than an identical child who happens to be at the very top of her reference distribution ($p = .004$). Compared to a sample mean of 14 cents, the effect of rank on sliming is very large.

Given that children were randomly assigned to either treatment or control, we conclude that the *opportunity* to engage in a second, disruptive activity caused lower-ranked students to substitute away from investing. Instead, they paid to engage in socially wasteful behavior. Note, in the control group, where students can only choose between practicing and waiting, there is little to no room for comparative advantage to affect behavior. But in the treatment group, where children are faced with the choice between two very different activities, students with a relative disadvantage at solving mazes opted out of the very activity in which they did poorly compared to their peers. Our experimental results, therefore, suggest that students' behavior is, in part, governed by comparative advantage.[28]

[28] Our findings are also consistent with the idea that students derive utility from ordinal rank itself, say, from being "first" or "not last." The key to reconciling rank-based utility with the comparative advantage approach is to let social status, i.e. s_T and s_N, depend on ordinal rank. In such a world, agents would choose to engage in whichever activity they do better relative to their reference group, i.e., sliming or solving mazes. Thus, even if individuals care directly about their ordinal rank, comparative advantage continues to influence behavior as long as agents can choose among multiple activities.

5. Concluding Remarks

Drawing on traditional models of self-selection in the labor market, we propose a theory of social interactions based on comparative advantage. When comparative advantage is the guiding principle of peer group formation, an individual's behavior is an equilibrium outcome. It depends on where in the ability distribution she falls, and on the shadow prices that clear the social market. That is, in our model selection into peer groups is determined by the scarcity of various skills, and peer effects arise due to the endogenous sorting of agents into peer groups *within* a social setting. An important implication that distinguishes our theory from traditional models of peers effects is that student outcomes should depend on ordinal rank.

Our empirical findings show that this key prediction is borne out in one randomized controlled trial in Kenya as well as two large observational data sets from the United States. To further probe the mechanism through which ordinal rank affects behavior, we implemented a "framed" field experiment with nearly 600 public school students in Houston. The experimental evidence supports the idea of comparative advantage in social interactions.

The results in this paper speak to a large literature on peer and neighborhood effects (see Epple and Romano 2011 and Ioannides 2011 for surveys). To better understand these and related phenomena, economists have developed models of social interactions by putting environmental variables—such as the mean behavior in one's social group or the mean educational attainment in one's neighborhood—into agents' utility functions. In this class of models, peers are a source of positive externalities; unruly peers cause more trouble and smarter peers encourage higher achievement (Akerlof 1997; Becker 1974, 1996; Benabou 1993; Bernheim 1994).

Empirical evidence in support of theories that predict that favorable social interactions lead to better outcomes has been ambivalent. While many authors confirm the hypothesis relying on plausibly exogenous variation in data sets ranging from primary students in Texas to freshmen at Dartmouth College, others find no or even negative peer effects.[29]

Recently, Carrell et al. (2013) present intriguing evidence from a field experiment designed to boost the achievement of low-ability freshmen at the US Air Force Academy. Based on earlier, experimental estimates indicating nonlinear, positive effects of peers' mean test scores, Carrell et al. (2013) "optimally" construct peer groups by pairing low-ability students with a greater share of high-ability ones. However, contrary to the authors' expectations, the intervention had a *negative* and statistically significant effect on the very students it was

[29] For confirmatory evidence, see Bursztyn and Jensen (2015), Carrell et al. (2009), Duflo et al. (2011), Goux and Maurin (2007), Hanushek et al. (2003), Hoxby (2000), Hoxby and Weingarth (2005), Imberman et al. (2012), or Sacerdote (2001). For negative evidence, see Angrist and Lang (2004), Cullen et al. (2006), Kang (2007), or Sanbonmatsu et al. (2006).

designed to help. That is, in the experiment of Carrell et al. (2013), if anything, "better" classmates led to worse outcomes.

One potential explanation for these and other "anomalous" findings is that mean test scores are not a sufficient statistic to gauge the impact of one's peers on own behavior. Our results, for instance, suggest that students' behavior and achievement depend on their ordinal rank within narrowly defined social settings. If correct, then the net effect of being paired with better peers is likely to depend on the entire distribution of ability, not just its first moment.

Interestingly, Carrell et al. (2013) note that within their "optimally" designed peer groups, low-ability students avoided the very peers with whom they were intended to interact and instead formed more homogeneous subgroups. Our comparative advantage theory of social interactions predicts this observation.

In Cicala et al. (2011) we argue that the comparative advantage approach has the potential to rationalize many of the seemingly contradictory findings in the literature on social interactions. In particular, since individuals' ordinal rank may deteriorate as a result of many well-intentioned interventions, our theory provides a simple, economic explanation for why ostensibly better peers do not always lead to more favorable outcomes (see, e.g., Carrell et al. 2013; Kling et al. 2005, 2007; Sanbonmatsu et al. 2006).

References

Akerlof, G. A. (1997). "Social Distance and Social Decisions." *Econometrica,* 65(5): 1005–1028.

Angrist, J. D., and K. Lang (2004). "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program." *American Economic Review,* 94(5): 1613–1634.

Becker, G. S. (1974). "A Theory of Social Interactions." *Journal of Political Economy,* 82(6): 1063–1093.

——— (1991). *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

——— (1996). *Accounting for Tastes*. Cambridge, MA: Harvard University Press.

——— , and K. M. Murphy (2000). *Social Economics: Market Behavior in a Social Environment.* Cambridge, MA: Belknap Press of Harvard University.

Becker, G. M., M. H. Morris, and J. Marschak (1964). "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science*, 9(3): 226–232.

Benabou, R. (1993). "Workings of a City: Location, Education, and Production." *Quarterly Journal of Economics,* 108(3): 619–652.

Bernheim, B. D. (1994). "A Theory of Conformity." *Journal of Political Economy,* 102(5): 841–877.

Bertrand, M., J. Pan, and E. Kamenica (2015). "Gender Identity and Relative Income within Households." *Quarterly Journal of Economics,* 130(2): 571–614.

Blume, L. E., W. A. Brock, S. N. Durlauf, and Y. M. Ioannides (2011). “Identification of Social Interactions,” (pp. 853–964) in J. Benhabib, A. Bisin, and M. O. Jackson (eds.), *Handbook of Social Economics, Vol. 1.* Amsterdam: Elsevier.

Borjas, G. J. (1987). “Self-Selection and the Earnings of Immigrants.” *American Economic Review,* 77(4): 531–553.

Brock, W. A., and S. N. Durlauf (2007). “Identification of Binary Choice Models with Social Interactions.” *Journal of Econometrics,* 140(1): 52–75.

Bursztyn, L., and R. Jensen (2015). “How Does Peer Pressure Pressure Affect Educational Investments?” *Quarterly Journal of Economics*, 130(3): 1329-1367.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). “Bootstrap-Based Improvements for Inference with Clustered Errors.” *Review of Economics and Statistics*, 90(3): 414–427.

Canada, G. (1995). *Fist, Stick, Knife, Gun: A Personal History of Violence in America.* Beacon Press.

Card, D., A. Mas, E. Moretti, and E. Saez (2012) “Inequality at Work: The Effect of Peer Salaries on Job Satisfaction.” *American Economic Review,* 102(6): 2981–3002.

Carrell, S. E., R. L. Fullerton, and J. E. West (2009). “Does Your Cohort Matter? Measuring Peer Effects in College Achievement.” *Journal of Labor Economics,* 27(3): 439–464.

———, B. I. Sacerdote, and J. E. West (2013). “From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation.” *Econometrica,* 81(3): 855–882.

Charles, K. K., E. Hurst, and N. Roussanov (2009) “Conspicuous Consumption and Race.” *Quarterly Journal of Economics,* 124(2): 425–467.

Cicala, S., R. G. Fryer, and J. L. Spenkuch (2011). “A Roy Model of Social Interactions.” NBER Working Paper No. 16880.

Cooley, J. C. (2010). “Desegregation and the Achievement Gap: Do Diverse Peers Help?” mimeographed. University of Wisconsin – Madison.

Cullen, J. B., B. A. Jacob, and S. D. Levitt (2006) “The Effect of School Choice on Participants: Evidence from Randomized Lotteries.” *Econometrica,* 74(5): 1191–1230.

Digman, J. M. (1990). “Personality Structure: Emergence of the Five-Factor Model.” *Annual Review of Psychology,* 41: 417–440.

Duflo, E., P. Dupas, and M. Kremer (2011). “Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya.” *American Economic Review,* 101(5): 1739–1774.

Elsner, B., and I. Isphording (2015). “A Big Fish in a Small Pond: Ability Rank and Human Capital Investment.” IZA Discussion Paper No. 9121.

Epple, D., and R. E. Romano (2011). “Peer Effects in Education: A Survey of the Theory and Evidence,” (pp. 1053–1163) in J. Benhabib, A. Bisin, and M. O. Jackson (eds.), *Handbook of Social Economics, Vol. 1.* Amsterdam: Elsevier.

Friedman, M. (1953). “The Methodology of Positive Economics,” in *Essays in Positive Economics.*

Chicago: University of Chicago Press.

Glaeser, E. L., B. I. Sacerdote, and J. A. Sheinkman (2003). "The Social Multiplier." *Journal of the European Economic Association,* 1(2–3): 345–353.

Goux, D., and E. Maurin (2007). "Close Neighbours Matter: Neighbourhood Effects on Early Performance at School." *Economic Journal,* 117(523): 1193–1215.

Graham, B. S. (2008). "Identifying Social Interactions through Conditional Variance Restrictions." *Econometrica,* 76(3): 643–660.

Grossman, G. M., and E. Helpman (1990). "Comparative Advantage and Long-Run Growth." *American Economic Review,* 80(14): 796–815.

Hansen, K. T., J. J. Heckman, and K. M. Mullen (2004). "The Effect of Schooling and Ability on Achievement Test Scores." *Journal of Econometrics,* 121(1–2): 39–98.

Hanushek, E. A., J. F. Kain, J. M. Markman, and S. G. Rivki (2003). "Does Peer Ability Affect Student Achievement?" *Journal of Applied Econometrics,* 18(5): 527–544.

Harrison, G. W., and J. A. List (2004). "Field Experiments." *Journal of Economic Literature*, 42(4): 1009-1055.

Heckman, J. J. (1990). "Varieties of Selection Bias." *American Economic Review,* 80(2): 313–318.

——— , and J. A. Scheinkman (1987). "The Importance of Bundling in a Gorman-Lancaster Model of Earnings." *Review of Economic Studies,* 54(2): 243–255.

——— , and G. Sedlacek (1985). "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self-selection in the Labor Market." *Journal of Political Economy,* 93(6): 1077–1125.

Hoxby, C. M. (2000). "Peer Effects in the Classroom: Learning from Gender and Race Variation." NBER Working Paper No. 7867.

——— , and G. Weingarth (2005). "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects." mimeographed. Harvard University.

Imberman, S., A. D. Kugler, and B. Sacerdote (2012). "Katrina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees." *American Economic Review,* 102(5): 2048–2082.

Ioannides, Y. M. (2011). "Neighborhood Effects And Housing," (pp. 1281–1340) in J. Benhabib, A. Bisin, and M. O. Jackson (eds.), *Handbook of Social Economics, Vol. 1.* Amsterdam: Elsevier.

Kang, C. (2007). "Classroom Peer Effects and Academic Achievement: Quasi-Randomization Evidence from South Korea." *Journal of Urban Economics,* 61(3): 458–495.

Kling, J. R., J. Ludwig, and L. F. Katz (2005). "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment." *Quarterly Journal of Economics,* 120(1): 87–130.

——— , J. B. Liebman, and L. F. Katz (2007). "Experimental Analysis of Neighborhood Effects." *Econometrica,* 75(1): 83–119.

KRUEGER, A. B. (1999). "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics,* 114(2): 497–532.

KRUGMAN, P. R. (1981). "Intraindustry Specialization and the Gains from Trade." Journal of Political Economy, 89(5): 959–973.

KUZIEMKO, I., R. W. BUELL, T. REICH, and M. I. NORTON (2014). "Last-Place Aversion: Evidence and Redistributive Implications." *Quarterly Journal of Economics,* 129(1): 105–149.

LUTTMER, E. F. P. (2005). "Group Loyalty and the Taste for Redistribution." *Journal of Political Economy,* 109(3): 500–528.

MANSKI, C. F. (1993). "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies,* 60(3): 531–542.

MATSUYAMA, K. (1992). "Agricultural Productivity, Comparative Advantage, and Economic Growth." *Journal of Economic Theory*, 58(2): 317–334.

MILLER, R. A. (1984). "Job Matching and Occupational Choice." *Journal of Political Economy*, 92(6): 1086–1120.

MURPHY, K. M. (1986). "Specialization and Human Capital." Doctoral Dissertation. University of Chicago.

MURPHY, R., AND F. WEINHARDT (2014). "Top of the Class: The Importance of Rank Position." mimeographed. University College London.

RICARDO, D. (1817). *On the Principles of Political Economy and Taxation*. London: John Murray.

ROSEN, S. (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy,* 82(1): 34–55.

——— (1982). "Authority, Control, and the Distribution of Earnings." *Bell Journal of Economics*, 13(2): 311–323.

ROY, A. D. (1951). "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers,* 3(2): 135–146.

SACERDOTE, B. I. (2001). "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics,* 116(2): 681–704.

SANBONMATSU, L., J. R. KLING, G. DUNCAN, and J. BROOKS-GUNN (2006). "Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment." *Journal of Human Resources,* 41(4): 649–691.

SATTINGER, M. (1979). "Differential Rents and the Distribution of Earnings." *Oxford Economic Papers,* 31(1): 60–71.

SHEA, J. (1997). "Instrument Relevance in Multivariate Linear Models: A Simple Measure." *Review of Economics and Statistics,* 79(2): 348–352.

TINCANI, M. (2014). "On the Nature of Social Interactions in Education: An Explanation for Recent Puzzling Evidence." mimeographed. University College London.

——— (2015). "Heterogeneous Peer Effects and Rank Concerns: Theory and Evidence." mimeographed. University College London.

Stock, J. H., and M. Yogo (2005). "Testing for Weak Instruments in Linear IV Regression." (pp. 80–108) in D. W. K. Andrews and J. H. Stock (eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge, UK: Cambridge University Press.

Suskind, R. (1998). *A Hope in the Unseen: An American Odyssey from the Inner City to the Ivy League*. New York: Broadway Books.

Willis, R. J., and S. Rosen (1979). "Education and Self-Selection." *Journal of Political Economy,* 87(5): S7–S36.

Word, E., J. Johnston, H. Bain, B. Fulton, J. Zaharias, N. Lintz, C. Achilles, J. Folger, and C. Breda (1990). *Final Report. Student/Teacher Achievement Ratio (STAR): Tennessee's K–3 Class Size Study*. Nashville, TN: Tennessee State Department of Education.

Yatchew, A. (1998). "Nonparametric Regression Techniques in Economics." *Journal of Economic Literature,* 36(2): 669–721.

Appendix Materials

[NOT FOR PUBLICATION]

Appendix A: Predicting the Efficacy of Social Experiments Ex Ante

The fundamental issue that our comparative advantage approach to social interactions highlights is that contacts within narrowly defined social markets are endogenously determined *after* any potentially random assignment to a neighborhood, classroom, etc. Thus, an intervention's effect will at least partially depend on skill distributions, group-specific capital, and the ensuing relative social status, all of which are generally unobserved. While this may seem to imply that "anything goes" in the market for peers, it does not mean that it is impossible to predict the efficacy of social interventions ex ante.

On the contrary, if correct, our theory suggests a simple heuristic that can help policy makers predict outcomes of small-scale interventions, i.e. those which are unlikely to have large effects on equilibrium "prices." If a policy maker is interested in predicting the behavior of a child after moving to a new neighborhood, a new school, or new classroom, then the relevant statistic is the behavior of children with the same characteristics in the new environment. The reason is simple: children with similar characteristics who face the same social wages will likely have a common comparative advantage and can, therefore, be expected to behave similarly.

The challenge is to find a way to compare agents across markets. Let Θ_j denote the set of individual characteristics which determine sorting in environment j. This may include, for example, test scores or innate ability in a school intervention, or height, weight, and motivation in a neighborhood intervention. If one can identify Θ_j before an intervention commences, then students can be matched across social markets and the heuristic is straightforward.

Consider a few thought experiments. If Θ_j is test scores, then one can compare individuals across cities by their scores. If Θ_j is innate ability, methods developed in Hansen et al. (2004) to extract measures of ability can be used to match individuals with the same ability across markets. If Θ_j involves noncognitive skills such as those psychologists often refer to as "The Big Five"—Openness, Conscientiousness, Extraversion, Agreeableness, and Emotional Stability (e.g., Digman 1990)—one can develop pre-intervention surveys along these dimensions and match students on these five measures. Difficulties arise, however, when we have no theory or empirical evidence to inform Θ_j. In this case, one might use administrative or survey data to match on as many variables as possible, recognizing that the prediction will have more noise.

More generally, the predictions from our model are directly related to traditional program evaluations. Let $Y(i)$ be an indicator variable equal to one if individual i chooses to be a nerd in the old environment (and zero otherwise), and let $Y'(i)$ denote i's choice in the new environment. Then the average treatment effect from manipulating the environment for all i is equal to

$$ATE = \mathbb{E}\left[Y'(i) - Y(i)\right] = i^* - i^{*\prime},$$

where i^* and $i^{*\prime}$ denote the marginal individual in the old and new environments, respectively. In words, the average treatment effect is simply the fraction of individuals who switch sectors.

Interpreting our model more loosely, it perhaps more useful to think of outcomes (Y_0, Y_1), which are different from an agent's actual choice of sector, but nevertheless depend on it. For instance, let $Y_1(i|\Xi)$ denote i's test score (conditional on environmental variables Ξ) if i chooses to be a nerd, whereas $Y_0(i|\Xi)$ is her potential outcome as a troublemaker. In this case, the average treatment effect from transplanting a population of unit mass into a new environment characterized by Ξ' is

$$
\begin{aligned}
ATE = & \int_0^{\min\{i^*(\Xi), i^*(\Xi')\}} \left(Y_0(s|\Xi') - Y_0(s|\Xi)\right) ds \\
& + \mathbf{1}\left[i^*(\Xi) < i^*(\Xi')\right] \times \int_{i^*(\Xi)}^{i^*(\Xi')} \left(Y_0(s|\Xi') - Y_1(s|\Xi)\right) ds \\
& + \mathbf{1}\left[i^*(\Xi) > i^*(\Xi')\right] \times \int_{i^*(\Xi')}^{i^*(\Xi)} \left(Y_1(s|\Xi') - Y_0(s|\Xi)\right) ds \\
& + \int_{\max\{i^*(\Xi), i^*(\Xi')\}}^{1} \left(Y_1(i|\Xi') - Y_1(i|\Xi)\right) ds,
\end{aligned}
$$

where $i^*(\cdot)$ denotes the marginal individual in a given environment, and $\mathbf{1}[\cdot]$ is an indicator function equal to one if the condition in braces is satisfied. The first and last row in the equation above give the change in test scores for those individuals who do not switch sectors, whereas the middle rows denote the change in outcome for those agents who do switch (e.g., for nerds who become troublemakers or vice versa).

The formula highlights that although changing the environment from Ξ to Ξ' might be beneficial in the sense that it raises both Y_0 and Y_1 for *every* individual, if the difference between Y_0 and Y_1 (conditional on the environment) is sufficiently large, then the average treatment effect could still be negative—as observed, for instance, among male youth in the Moving to Opportunity Experiment of Kling et al. (2005, 2007), or in the experiment of Carrell et al. (2013).

Appendix B: Comparative Advantage and Identification of Peer Effects

Above, we have outlined a new way of thinking about social interactions using sorting and comparative advantage as the guiding principles of peer group formation. Here, we consider its implications for the identification of "traditional" peer effects. In doing so, we follow the literature and assume that there exist other factors besides social status that determine the utility of being a troublemaker or a nerd—e.g., personal and neighborhood characteristics, peers' test scores, or their behavior itself.

To fix ideas, consider a student's choice of becoming a troublemaker, T, or a nerd, N. Let $\mathbf{X}_i$ be a set of individual-level covariates, and let $\mathbf{Z}_m$ denote factors varying only at the level of the social market, i.e. schools or neighborhoods. Mean teast scores, or mean behavior in market m, is given by $\overline{y}_m$, and ν_{im} represents an error term known to the individual but not the econometrician. Intuitively, ν_{im} captures all unobserved factors influencing the difference in utility between T and

N. Maximizing utility, student i chooses to become a troublemaker if and only if

$$(11) \qquad u(T;\mathbf{X}_i,\mathbf{Z}_m) - u(N;\mathbf{X}_i,\mathbf{Z}_m) = \kappa + \mathbf{X}_i'\boldsymbol{\beta}_0 + \mathbf{Z}_m'\boldsymbol{\gamma}_0 + \alpha_0\overline{y}_m + \nu_{im} \geq 0.$$

Social status and individual ability are typically not directly observable. Thus, the comparative advantage approach should be viewed as providing a more explicit theory of the error term. Following our theoretical model, we can decompose ν_{im} into the net social payoff from being a troublemaker and some other random variable:

$$\nu_{im} = (s_{Tm}\sigma_{Ti} - s_{Nm}\sigma_{Ni}) + \epsilon_i.$$

Note that only ϵ_i and σ_{ji}, $j \in \{N, T\}$, are possibly independent and identically distributed across individuals, whereas s_{Tm} and s_{Nm} (both of which are measured in utility units) vary only at the market or group level. Therefore, our theory stipulates the existence of group-level unobservables.

It is well known that in the presence of group-level unobservables, not all parameters in the binary choice model are identified from cross-sectional data (Blume et al. 2011; Brock and Durlauf 2007). While $\boldsymbol{\beta}_0$ can be consistently estimated without imposing parametric assumptions (using methods outlined in Heckman 1990), $\boldsymbol{\gamma}_0$ and α_0—the coefficients of interest in the majority of applied work—cannot. Nonidentification is due to the fact that ν_{im} depends on $\mathbf{Z}_m$ and $\overline{y}_m$ in an unknown way. Therefore, only the linear combination of market-level observables and unobservables is identified (see Brock and Durlauf 2007 for a formal proof).

It is important to note that non-identification as a result of group-level unobservables is quite distinct from endogeneity due to systematic sorting of individuals into social markets (such as neighborhoods and classrooms), or the reflection problem, which poses that α_0 cannot be identified if $\mathbf{Z}_m$ and $\overline{y}_m$ are linearly dependent (Manski 1993).[30] While applied researchers have often found clever strategies to deal with these two problems, group-level unobservables have received much less attention.[31]

To appreciate the consequences, denote i's observed behavior by

$$y_i = \begin{cases} 1 & \text{if } u(T;\mathbf{X}_i,\mathbf{Z}_m) - u(N;\mathbf{X}_i,\mathbf{Z}_m) \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

and consider the case in which there are no endogenous peer effects, i.e. $\alpha_0 = 0$. By the Frisch–Waugh Theorem and assuming that $\mathrm{Cov}(\mathbf{X}_i^*, \nu_{im}) = 0$, the probability limit of the ordinary least

[30] If $\mathbf{Z}_m$ exhibits sufficient variation and $\boldsymbol{\gamma} \neq 0$, then the binary choice model of social interactions does not suffer from the reflection problem, as the limited range of the outcome rules out perfect linear dependence (Brock and Durlauf 2007).

[31] However, there are several notable exceptions. Cooley (2010) motivates her instrument in the presence of unobserved differences in teacher quality. Hoxby (2000) uses panel data to remove the effect of group-level unobservables that do not vary over time; and Graham (2008) shows how conditional variance restrictions can be used to identify endogenous peer effects when individual- and group-level unobservables are uncorrelated.

squares estimate of α_0 equals

$$\text{plim}\,\hat{\alpha}_{OLS} = \alpha_0 + \frac{\text{Cov}\left(\overline{y}^*_m, \nu_{im}\right)}{\text{Var}\left(\overline{y}^*_m\right)} = \frac{\text{Cov}\left(\overline{y}^*_m, \nu_{im}\right)}{\text{Var}\left(\overline{y}^*_m\right)},$$

where $\overline{y}^*_m$ denotes the residual from projecting $\overline{y}_m$ onto $\mathbf{X}_i$ and $\mathbf{Z}_m$. Only if $\overline{y}^*_m$ and ν_{im} are uncorrelated, will $\hat{\alpha}_{OLS}$ be consistent.

However, the comparative advantage approach to social interactions predicts that $\text{Cov}\left(\overline{y}^*_m, \nu_{im}\right) > 0$. To see this, condition on $\mathbf{X}_i$ and $\mathbf{Z}_m$ and note that, according to (11), individual i in social market m chooses to become a troublemaker if and only if

$$\nu_{im} \geq \xi\left(\mathbf{X}_i, \mathbf{Z}_m\right), \tag{12}$$

where $\xi\left(\mathbf{X}_i, \mathbf{Z}_m\right) \equiv -\left(\kappa + \mathbf{X}_i'\boldsymbol{\beta}_0 + \mathbf{Z}_m'\boldsymbol{\gamma}_0\right)$. Now, decompose ν_{im} into the market specific mean social payoff, $\overline{v}_m$, and deviations around the mean, $\widetilde{v}_{im}$, which are distributed according to some cumulative distribution function $\Phi_m(\cdot)$. That is, let $v_{im} = \overline{v}_m + \widetilde{v}_{im}$. With this notation in hand, $y_i = 1$ if and only if

$$\widetilde{v}_{im} \geq \xi\left(\mathbf{X}_i, \mathbf{Z}_m\right) - \overline{v}_m,$$

and the fraction of individuals who are troublemakers in market m is equal to

$$\overline{y}_m = \mathbb{E}_{\mathbf{X}_i}\left[1 - \Phi_m\left(\xi\left(\mathbf{X}_i, \mathbf{Z}_m\right) - \overline{v}_m\right)\right].$$

Unless $\mathbf{X}_i$ and $\mathbf{Z}_m$ fully determine individuals' behavior, it will generally be the case that $\frac{d\overline{y}^*_m}{d\overline{v}_m} > 0$, as $\frac{d\overline{y}_m}{d\overline{v}_m} > 0$. Hence, it follows that $\text{Cov}\left(\overline{y}^*_m, \nu_{im}\right) > 0$.

The intuition for non-identification is straightforward. Under the assumptions of our model, a particular behavior will be more prevalent in markets in which the (unobserved) social net payoff to it is higher—say, because of more group-specific capital K_j. It follows that although endogenous social interactions might not be a driver of behavior (i.e. $\alpha_0 = 0$), linear-in-means estimates will be biased toward finding this form of peer effect—even under random assignment to social markets and if one resolves the reflection problem.

Appendix C: Can Teacher Behavior Explain the Relationship between Rank and Outcomes?

As mentioned in Section 3.4, there are several alternative mechanisms that might explain the relationship between ordinal rank and student outcomes. In this appendix, we use data from the Early Childhood Longitudinal Study (ECLS) to provide a partial test of the 'teacher behavior' explanation.

The ECLS is a longitudinal, nationally representative dataset that followed students from the kindergarten cohort of 1998-99 all the way through eighth grade. The ECLS data contain information on demographics for students, parents, and schools, early-childhood education programs, students' behavior and experiences, academic achievement and, importantly for our purposes, teachers' subjective evaluations of students. Table A.2 presents summary statistics for the variables we use in our analysis.

Since the data contain subjective evaluations of students' performance as well as students' actual test scores from two subjects in the same year, we can mimic the empirical approach that we took with the NELS data and relate different teachers' subjective evaluations of the same student to differences in her ordinal rank across classrooms. That is, we estimate

$$(13) \qquad \Delta y_i = \varphi \Delta r_i + \mathbf{X}_i' \beta + Grade_i + \epsilon_i,$$

where y_i denotes student i's standardized score on the Academic Rating Scale (ARS) for a particular subject. ARS is a composite index that is computed from subjective teacher assessments on a variety of skills, graded on a scale from "not yet" to "proficient."[32] As controls we include actual test scores, gender, an exhaustive set of race indicators, whether the student's language at home was English, socio-economic quartile, parents' educational achievement and biological relationship with the student, and school characteristics and location. We also account for birth year, birth month, and grade fixed effects.

For each wave of the ECLS, the upper panel of Table A.3 presents estimates of φ. Moving from the left to the right of the table, we add higher order polynomials of students' test scores to better proxy for true ability. Overall, conditional on actual achievement, there appears to be no systematic relationship between the class rank of students and teachers' subjective assessments. Seventeen of the point estimates are positive, while the remaining thirteen are negative. Individual coefficients range from -.133 to .208 and are, with a few exceptions, statistically indistinguishable from zero There is, therefore, no evidence to conclude that ordinal rank affects how teachers view equally able students.

For completeness, exploiting the fact that the last wave of the data also contains teacher reports about problem behaviors, we show in the lower panel of Table A.3 that ordinal rank appears to affect problem behaviors among students in the ECLS—just as it did in the NYCPS and NELS data. Specifically, we estimate the model in equation (13) with the difference in behavioral incidents across classrooms as the outcome.

Although the evidence from the ECLS suggests that teachers do not rely on ordinal rank to assess students, it is insufficient to rule out all teacher-based explanations for our main findings. Suppose, for instance, that teachers always target the top of the class. If this is the case, then, conditional on own ability, higher ranked students would benefit from more appropriate instruction and thus experience an increase in test scores. Lower ranked students may act out due to a lack of attention.

[32] For additional detail regarding the ARS as well as descriptions of all other variables in our empirical specifications, see Appendix D.

Appendix D: Data Appendix

D.1. *ETP Experiment of Duflo et al. (2011)*

Tracking and Nontracking Schools Tracking School and Nontracking School are dichotomous variables equal to one if a student initially attened a school that was randomly chosen for the tracking treatment or the nontracking one, respectively.

Top vs Bottom Section Top and Bottom Section are dichotomous variables equal to one if a student initially attended a school that was randomly chosen for the tracking treatment and if she was assigned (based on initial achievement) to the top or bottom, respectively.

Civil Service and Contract Teachers Contract Teacher is an indicator variable equal to one if a given section was randomly assigned a teacher hired on a contractual basis, and zero if it was taught by a civil service teacher instead.

Additional Controls We use the entire set of control variables contained in the data of Duflo et al. (2011): age at the beginning of the intervention, gender, and indicator variables for whether the school is located in the Bungoma district and whether it was also sampled for another experiment on school-based management.

Test Scores As our dependent variable we use standardized total tests scores at endline, i.e. 18 months after the intervention began. According to Duflo et al. (2011) the underlying test was partially designed by a cognitive psychologist in order to measure a range of age appropriate skills. While part of the test was written, the remainder was orally administered one-to-one by trained enumerators. Students were asked math and literacy questions, such as identifying letters, counting, subtracting three-digit numbers, reading, and understanding sentences. As Duflo et al. (2011) we control for achievement prior to the intervention by using test scores at baseline. Schools administered baseline tests individually, which is why we standardize scores on the school level. Peers' mean test score is defined as the standardized leave-one-out mean baseline test scores of an individual's classmates. In some specifications we allow the impact of peers' mean test score to vary by quartile of the school specific initial test score distribution in which the individual finds herself.

Rank Rank is defined as a student's percentile in the school and grade specific distribution of baseline test scores. Hence, rank ranges from 100 to 0 with smaller values indicating a position closer to the bottom of the distribution.

D.2. *New York City Public Schools*

Demographic Variables Demographic variables that should not vary from year to year (such as race and gender) were pulled from New York City enrollment files from 2003/04 through 2008/09, with precedence given to the most recent files. Race consisted of the following categories: black, Hispanic, white, Asian, and other race. These categories were considered mutually exclusive. The "other race" category consisted of students who were coded as "American Indian." Gender was coded as male, female, or missing.

Demographic variables that may vary from year to year (free lunch status, English Language Learner status, and special education designation) were only pulled from the enrollment file corresponding to the same year as the observation. A student was considered eligible for free lunch if he was coded as "A" or "1" in the raw data, which corresponds to free lunch, or "2", which corresponds to reduced-price lunch. A student was considered non-free lunch if the student was coded as a "3" in the NYC enrollment file, which corresponds to full price lunch. All other values, including blanks, were coded as missing. For English Language Learner status, a student was given a value of one if he was coded as "Y" for the limited English proficiency variable. All other students in the NYC data were coded as zero for English Language Learner status. Special education was coded similarly.

New York State Test Scores NYC state test scores were constructed from the NYC test score files for 2003/04 through 2008/09 for English/Language Arts (ELA) and math. School-wide rankings were constructed based on test scores in 5th grade.

The state math and ELA tests are high-stakes exams conducted in the winters of third through eighth grade. Students in third, fifth, and seventh grades must score level 2 or above (out of 4) on both tests to advance to the next grade without attending summer school. The math test includes questions on number sense and operations, algebra, geometry, measurement, and statistics. Tests in the earlier grades emphasize more basic content such as number sense and operations, while later tests focus on advanced topics such as algebra and geometry. The ELA test is designed to assess students on three learning standards—information and understanding, literary response and expression, and critical analysis and evaluation—and includes multiple-choice and short-response sections based on a reading and listening section, along with a brief editing task. Content breakdown by grade and additional exam information is currently available at <http://www.emsc.nysed.gov/osa/pub/reports.shtml>.

Behavior The number of behavioral incidents for each student was determined from NYC files listing all recorded behavioral incidents from 2004/05 through 2008/09. Students not listed in this file but with a valid test score from the same year were assumed to have zero behavioral incidents. We constructed a behavioral incident indicator with a value of one if the student was listed for a behavioral incident in the file from the relevant year, a value of zero if the student had a valid test score from the same year, and missing otherwise.

D.3. *National Educational Longitudinal Study*

Demographic Variables Demographic variables were taken from the baseline year of the survey. These included: race, sex, English Language Learner status, parents' marital status, parents' education, school type (public, Catholic, or other private), school location (urban, suburban, rural), socioeconomic status, birth month, and birth year.

Behavior Behavior variables were constructed using data from teacher reports on individual students. Teachers were asked to indicate whether the student had problems in each of the following areas: the student performs below his ability, the student does not complete homework, the student is frequently absent, the student is frequently tardy, the student is inattentive, or the student is disruptive. In the baseline year (eighth grade), each student had one teacher report from either Math or Science and another from either English or History, for a total of two teacher reports. Similarly, each student had two reports from the first follow-up year (tenth grade). Teacher reports were also administered in the second follow-up year (twelfth grade) but only in one subject, so these reports are excluded from our analysis, which takes advantage of within-year across-subject variation. For each student, we constructed an indicator that is equal to one if the student's teacher reports that the student has a problem in at least one of the six categories and zero otherwise. The outcomes used for our analysis are the within-year differences across subjects in the behavioral indicator.

Grades The dataset contains self-reported grades for the baseline, first follow-up, and second follow-up years. In the baseline year, students were asked to report for each subject (Math, Science, English, and History) whether their grades since sixth grade had been "mostly A's (90-100)," "mostly B's (80-89)," "mostly C's (70-79)," "mostly D's (60-69)," or "mostly below D (below 60)." Similarly, in the first follow-up year, students were asked to report for each subject whether their grades from ninth grade until now were "mostly A's," "about half A's and half B's," "mostly B's," "about half B's and half C's," "mostly C's," "about half C's and half D's," "mostly D's," or "mostly below D." These responses were converted to the average of the corresponding grade point values on a 4.0 scale, where 1.0 corresponds to D, 2.0 corresponds to C, 3.0 corresponds to B, and 4.0 corresponds to A. These grade values were used to compute a student's percentile rank within each class.

Test Scores The dataset contains test scores for each student in Math, Science, English, and History for each year. We construct a test score control that is the mean of the test scores from the two subjects for which there are teacher reports in the baseline year and first follow-up year. We also calculate its square and use both as controls in our estimates.

D.4. *Early Childhood Longitudinal Survey*

Demographic Variables Demographic variables were taken from the baseline year of the survey. These included: race, sex, language spoken at home, parents' marital status, parents' education, school type (public, Catholic, other religious, or other private), school location (urban, suburban, town, rural), socioeconomic status, birth month and birth year, and grade.

Behavior Behavior variables were constructed using data from the 8th grade teacher questionnaire. These variables did not exist for earlier waves. The questions used were: (i) Has this student fallen behind in school work in this class? (ii) When you assign homework for this class, how often does this student complete it? (iii) How often is this student attentive in your class? (iv) How often is this student disruptive in your class? (v) How often is this student absent from your class? (vi) How often is this student tardy to your class? A indicator variable was set to one if a teacher answered any of these questions negatively, and zero otherwise. To examine the difference in the behavioral incident indicator, we looked at the difference of the behavioral incident variable between math and reading.

Academic Rating Scale Score The ARS score was calculated from teachers' responses to the Academic Rating Scale. Teachers were asked to assess their student in math and reading and grade their performance on a number of skills. The overall score is a continuous score between 1 and 5. The difference in ARS is calculated as the difference in the ARS score in reading and math.

Test Scores This variable was constructed from the Item Response Theory –based overall scaled score provided by ECLS. Percentiles were calculated for all students based on their ranking in their reading or math class. Difference in percentile ranking was calculated as the difference in percentiles in reading and math.

D.5. *Experimental Data from Houston Public Schools*

Willingness to Practice Appendix E describes how we elicited students' willingness to pay to practice on additional mazes. In our analysis we use students' stated willingness to pay (in dollars) and the total dollar amount that students actually spent on practicing.

Willingness to Slime As explained in Appendix E, students in the treatment group could pay to "slime" the screens of other participants. The same appendix describes how we elicited students' willingness to pay for sliming. In our analysis we examine total dollar amount that students spent on "sliming."

Baseline Performance & Ordinal Rank As explained in Appendix E, during the first stage of the experiment we established a baseline measure of students' ability to solve mazes. Specifically, we measure performance by the average time it took a student to complete the set of mazes in that stage. We then rank students accordingly. The variable rank is defined as the percentile ranking among participants in the same experimental session (with higher values assigned to faster students).

Control Variables Our set of control variables is listed in Table 9. With the exception of Self-Assessed Ability and Baseline Performace, these data were furnished to us by the schools. Self-Assessed Ability corresponds to students' answer to the question "how good are you at solving mazes (scale of 1-10)?" We rely on the first answer given, i.e. before students started to work on any mazes (see also Appendix E).

Appendix E: Experimental Setup

E.1. *Schools & Program Launch*

We partnered with the Houston Independent School District (HISD) to implement an artefactual field experiment using a specially-designed computer game. The experiment was implemented at two traditional public middle schools. Enrollment at School A was 623 students. As of the beginning of the 2014-15 school year, 91.97 percent of students are economically disadvantaged, 18.14 percent are special education students, and 43.66 percent are limited English proficiency. The racial breakdown of the school is: 13.80 percent black, 85.23 percent Hispanic, 0.64 percent white, and 0.48 percent American Indian. Females are 44.3 percent and males are 55.7 percent of the student population.

Enrollment at School B was 1,217 students. As of the beginning of the 2014-15 school year, 29.66 percent of students are economically disadvantaged, 6.7 percent are special education students, and 18.3 percent are limited English proficiency. The racial breakdown of the school is: 9.4 percent black, 33.1 percent Hispanic, 44.7 percent white, and 0.58 percent American Indian. Females are 48.6 percent and males are 51.4 percent of the student population.

Before implementation began, we visited School A to see the computer lab facilities, speak with the principal about running the experiment, and distribute parental consent forms. The only important constraint was that students at both schools had to be pulled from specific classes in order to participate (see below). We also visited classrooms to answer any questions from students. Due to time constraints, a pre-implementation visit was not possible at School B. Both schools were offered up to $5,000 to participate in the study.

E.2. *Recruitment and Randomization*

All students at the middle schools were eligible to participate. At School A, parental consent forms (in English and Spanish) were sent home with students on February 16, 2015 and were due back on February 25, seven school days later. Forms were collected and shipped to us on February 26. We received 483 valid consent forms; students were randomized into control ($n = 232$), and treatment ($n = 251$). At School B, parental consent forms (in English and Spanish) were sent home with students on May 4, 2015 and were due back on May 11, one week later. Forms were collected and shipped to us on May 13. We received 306 valid consent forms; students were randomized into control ($n = 151$) and treatment ($n = 155$).

In order to participate, students had to be pulled from specific classes. At School A, it was their elective courses, while at School B it was social studies. At School A, electives take place during each period of the school day, blocks one (1) through five (5). Students indicated their elective block on the consent forms. After randomization, students were sorted into groups with a cap of 30 (because the school's computer lab had 30 computers that could be used at once). Because of this scheduling constraint, sessions at the school were held on three separate occasions: March 2-3, March 9-10, and April 6-7, 2015. At School B, social studies classes take place all periods of the day except for period 4. Due to block scheduling, on one day periods 1, 3, 5, and 7 take place; on the next day, periods 2, 4, 6, and 8 take place. After randomization, students were sorted into groups based on their social studies teacher, with a cap of 32 (the size of the school's computer lab). Due to the end of the school year, we were only able to hold sessions at the school from May 18-20, 2015.

For both schools, once groups were created, a schedule with student rosters for each session was emailed to the principal. The principal and school administrative team distributed passes to the students who appeared on the rosters so that they could leave their elective or social studies course and check in at the library. As students arrived they were given the experimental instructions to read. After all students—or as many as the administrative team could locate—in a group arrived, they were ushered into the computer lab, where they signed and submitted assent forms. Out of the 789 students who returned valid consent forms, 573 ended up completing the experiment.

E.3. *Experiment*

The experiment was run using an online software we developed. After signing the assent form, students would log into the software using their unique student ID and session code. Once they were logged in, we would read the instructions aloud to ensure understanding of the experiment. The stages of the experiment, reflected by the software, are described below.

Stage 1 Students were first asked two questions: (1) how good are you at solving mazes (scale of 1-10)? (2) how sure are you of this ability (scale of 1-10)? Depending on the experimental session, they were then given ten minutes to solve either five or twenty mazes. Students received \$0.25 for each maze completed in this stage. After completing all the mazes or at the end of the ten minutes (whichever came first), students were shown a list of all students in their session, sorted and ranked by average maze-completion time.

Stage 2 For students in the control group, stage 2 began with a prompt from the software: students were asked to enter the amount of money (between \$0.01 and \$0.50) they were willing to pay (per maze) to be able to practice up to 20 mazes. We used a random number generator between \$0.01 and \$0.50 to set the cost of practicing per maze (x). If the amount that the student was willing to pay was higher than x, then she was permitted to practice on many as 20 mazes, at a per-maze cost of x, for up to 10 minutes. If, however, the amount that the student was willing to pay was lower than x, she was not permitted to practice at all.

Students in the treatment group were prompted to enter two numbers: (1) the amount of money (between \$0.01 and \$0.50) they were willing to pay (per maze) to be able to practice on up to 20 mazes; and (2) the amount of money (between \$0.01 and \$0.50) they were willing to pay to "slime" a portion of another student's screen to distract them from practicing. We used a random number generator between \$0.01 and \$0.50 to independently set the costs of practicing per maze (x) and sliming (y). If the amount that the student was willing to pay to practice was higher than x, then she was permitted to practice at the cost of x per maze for 10 minutes. If, however, the amount that the student was willing to pay to practice was lower than x, she was not permitted to practice at all. Similarly, if the amount that the student was willing to pay to slime was higher than y, then she was permitted to slime at the cost of y. If, however, the amount that the student was willing to pay to slime was lower than y, she was not allowed to slime at all.

Students in the treatment whose willingness to pay bids were accepted for both practicing and sliming were permitted to do both: practicing mazes and sliming each had their own tab. In order to slime, students would first select a student from the list of stage 1 rankings, then select the quarter of the maze they want to block, then hit the "Slime!" button. An image depicting green slime would appear on the selected student's practice maze for 30 seconds, then disappear. A ticker on the side of the sliming screen showed who slimed whom at all times.

Stage 3 In the final stage, all students were given five minutes to complete ten mazes. Students received $3.00 for each maze completed in this stage.

E.4. *Payment & Program Figures*

After the end of the experiment, the student's earnings would appear on the screen. Students could earn a maximum of $37.00—$2.00 for showing up to the study, $0.25 per maze in stage 1, and $3.00 per maze in stage 3. Students filled out a subject payment form and received cash immediately following the session.

At School A, 433 students participated in the study: 217 in the control, and 216 in the treatment group. We ran a total of 23 sessions over the course of six school days. We distributed a total of $12,213.70 to students at school A during the course of the experiment. Average earnings were $28.21.

At School B, 146 students participated in the study: 57 in the control, and 89 in the treatment group. We ran a total of 10 sessions over the course of three school days. We distributed a total of $4,751.93 to students during the course of the experiment. Average earnings were $32.55.

E.5. *Experimental Instructions*

The following page shows the experimental instructions for the treatment group. Instructions for the control group mirror those for the treatment group, but did not contain any mention of "sliming."

Instructions

The session that you will participate in has three stages, as well as questions at the beginning and end of the session. The three stages are described below. You may choose to stop participating at any time during any stage.

First you will be asked a brief questionnaire to gauge your beliefs about your ability at **solving mazes**.

<u>Stage 1: First task</u>

You will then be asked to solve 20 mazes. You will receive 25 cents for each maze that you solve. After all participants complete the set of 20 mazes, participants will be ranked by the average time it took to complete the mazes. Each participant's rank and average time will be revealed to all other participants.

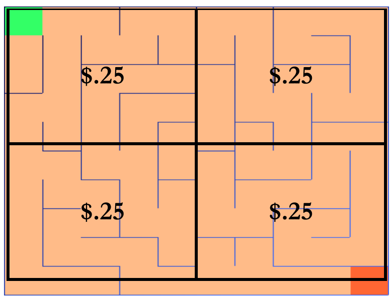
After the task is completed, you will be asked again about your ability at solving mazes and how sure you are of this ability. The second stage of the experiment will begin only after everyone else has finished this task. If you have completed this task quickly and are waiting on others, you will be given access to Internet that you may use as you wish.

<u>Stage 2: Practice</u>

You will be asked to write down two numbers: (1) the maximum amount (between \$0.01 and \$1.00) you are willing to pay per maze to practice them; and (2) the maximum amount (between \$0.01 and \$1.00) you're willing to pay to "slime" a quarter of a maze of another participant while they are practicing. The software will pick two random numbers $\{x, y\}$ between \$0.01 and \$1.00. If the amount you indicate to practice is higher than x, you will be allowed to practice on mazes at the cost of x per maze. If the amount you indicate to practice is lower than x, you will not be allowed to practice at all.

Similarly, if the amount you indicate to slime is higher than y, you will be allowed to slime others' mazes at the cost of y per maze. If the amount you indicate to slime is lower than y, you will not be allowed to slime others at all. If you decide to practice, you will be given 20 mazes. The second task will include 10 mazes from among this pool of 20. These mazes will be more difficult than the mazes from the first task. Each practice maze costs \$$x$ so if you decide to practice all 20 mazes, \$$(20*x)$ will be deducted from your earnings.

If you decide to "slime" other participants' computers, you can log on to the sliming tab and **block** a portion of the maze that they are working on. You'll be able to block one of the four squares in the picture below:



Every participant's rank (based on their performance on the first series of mazes) will be displayed on the screen to help you figure who you wish to slime. You can log out of the sliming tab whenever you wish to stop sliming and start practicing on mazes (if you're allowed to practice). If you wish to slime again, you will have to log in to the sliming tab again.

There is no constraint on who you wish to slime. When a participant has been slimed, the affected region will be blocked for 30 seconds. After the 30 seconds, their original maze screen will appear again. A ticker shows who slimed whom at all times to all participants.

Stage 3: Second Task
You will be asked to complete 10 mazes in 5 minutes. You will receive $3 for each maze completed.

Figure 1: Equilibrium in the Market for Peers when Status Decreases with Effective Group Size

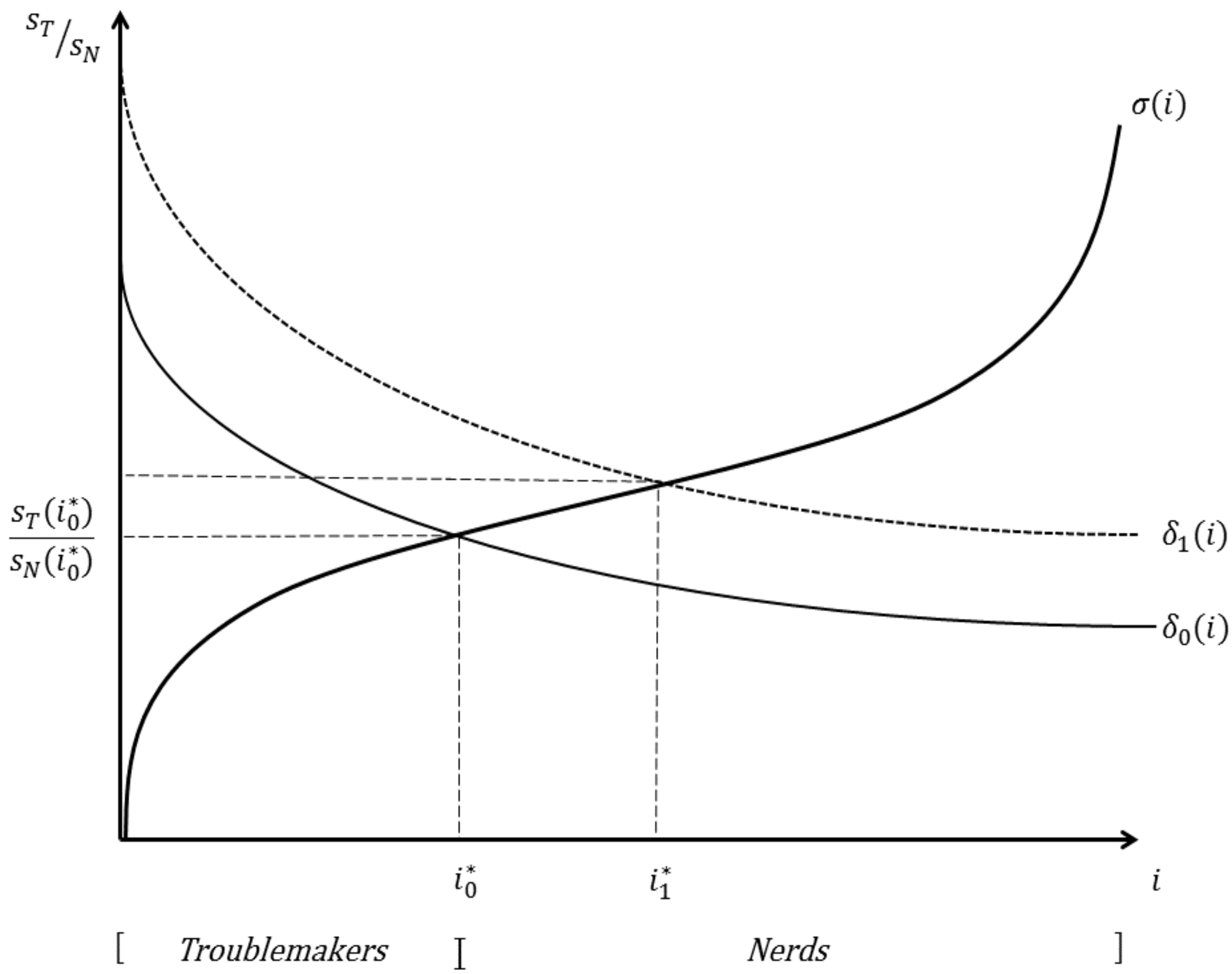


Figure 2: Effect of Higher-Quality Peers when Social Status is Decreasing in Effective Group Size

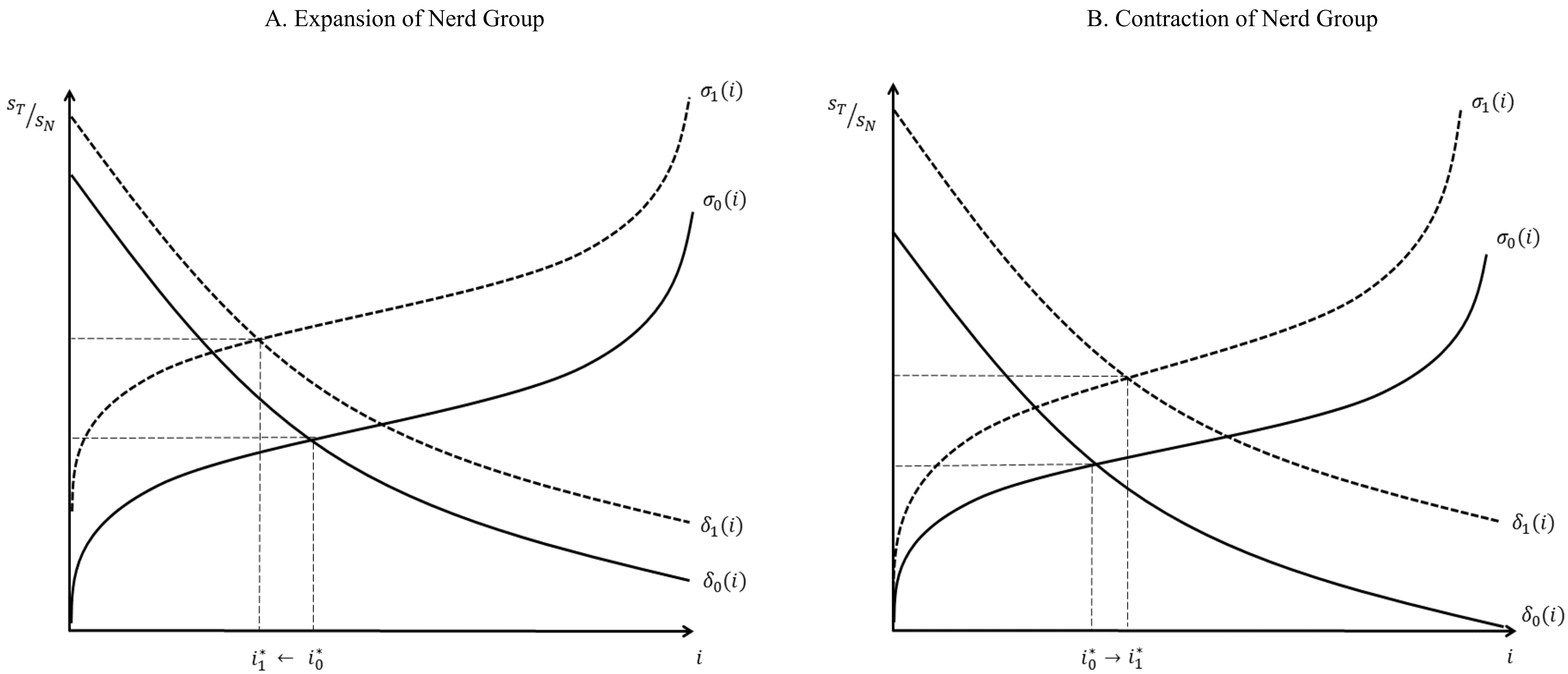


Figure 3: Equilibria in the Market for Peers when Status Increases with Effective Group Size

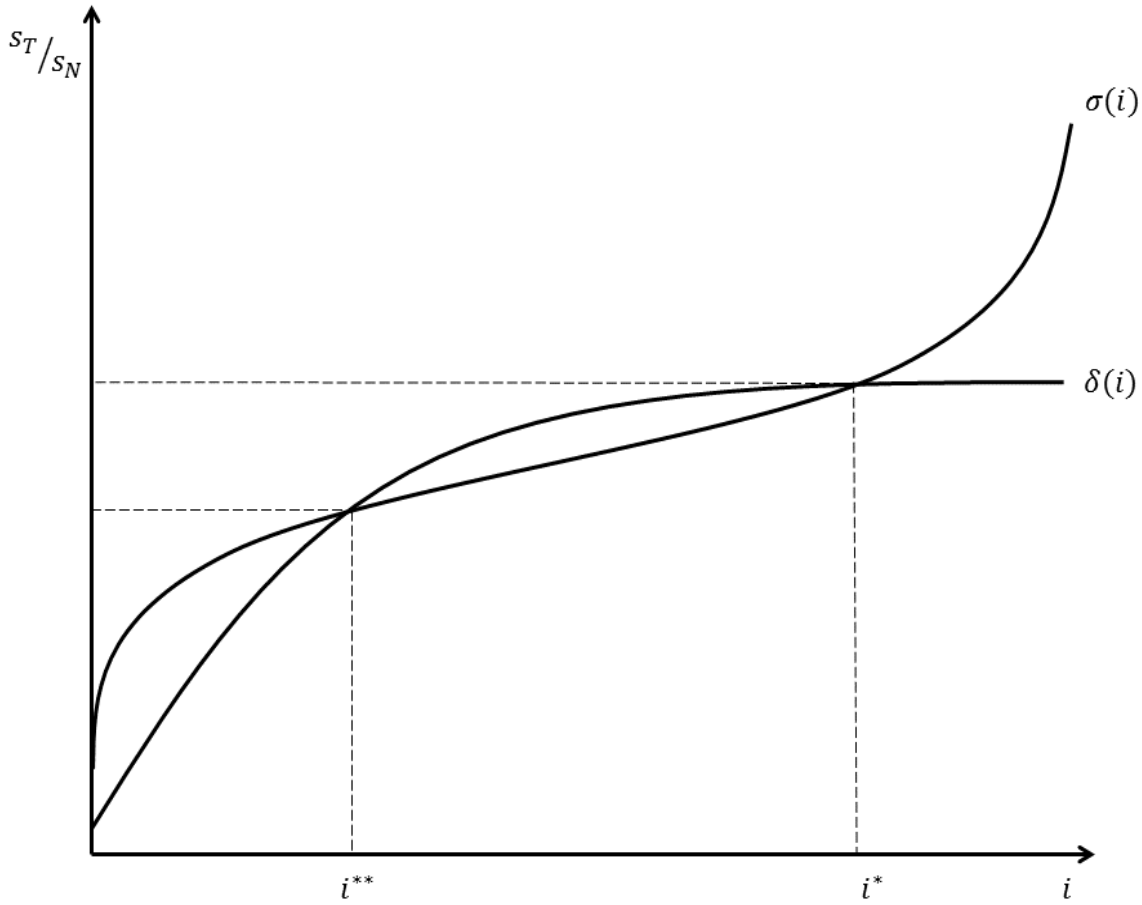


Figure 4: Comparative Statics when Status Increases with Effective Group Size

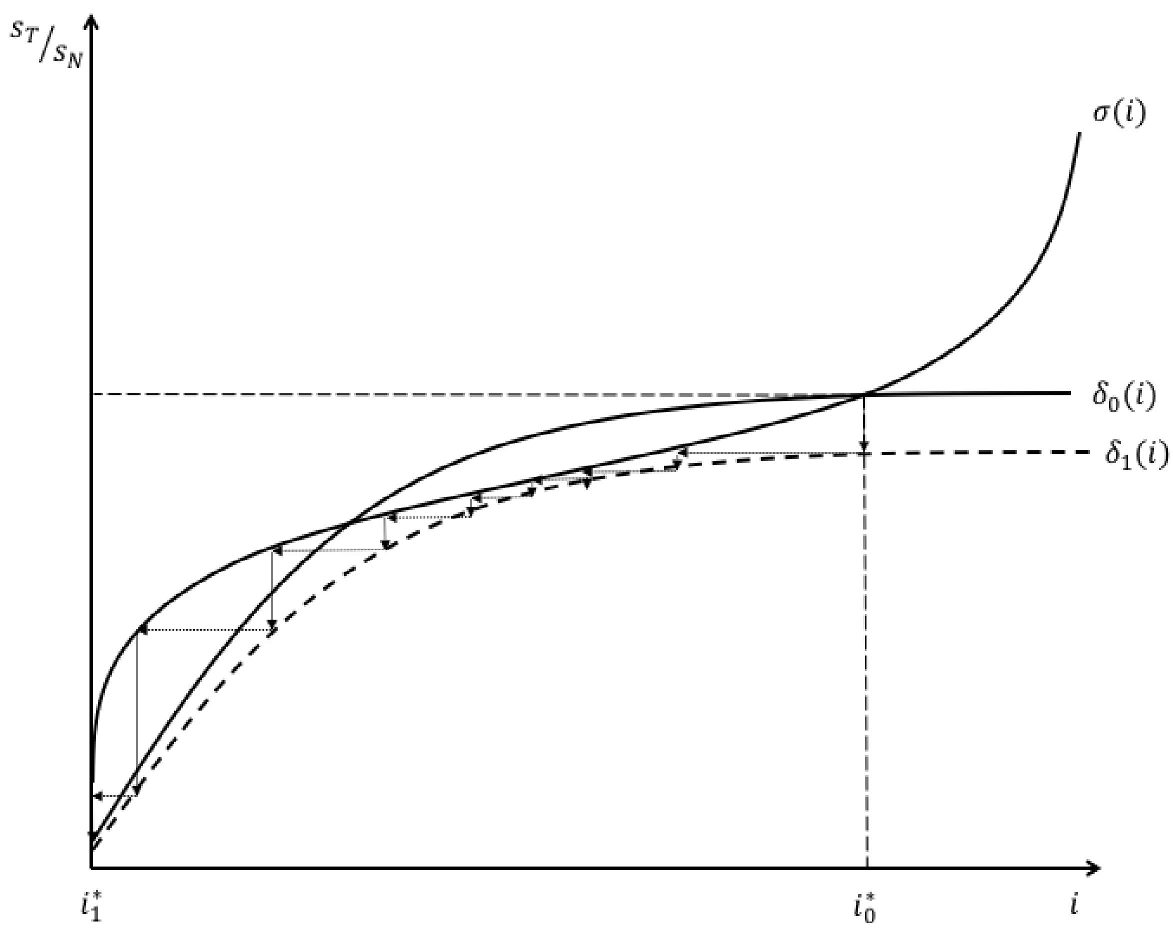
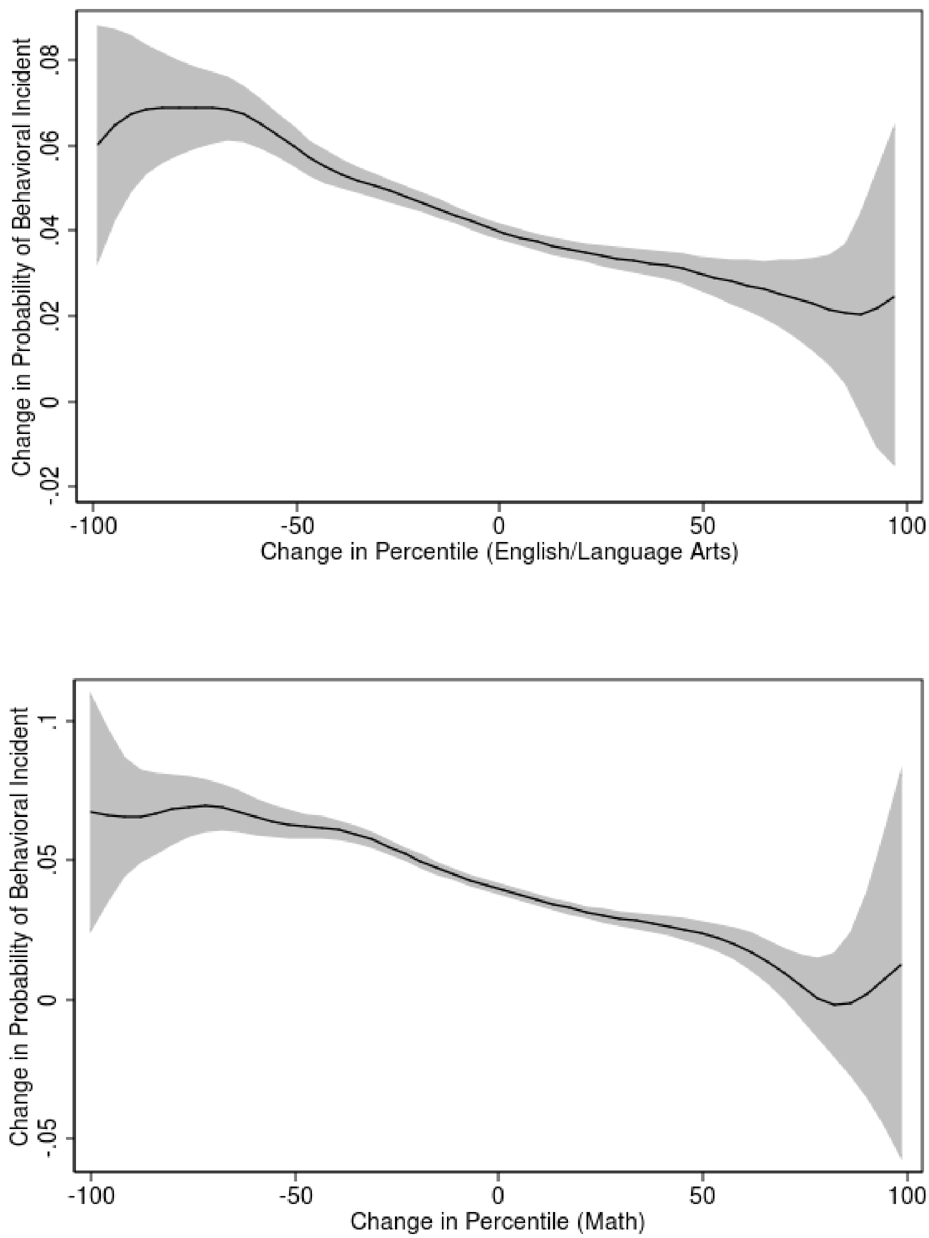
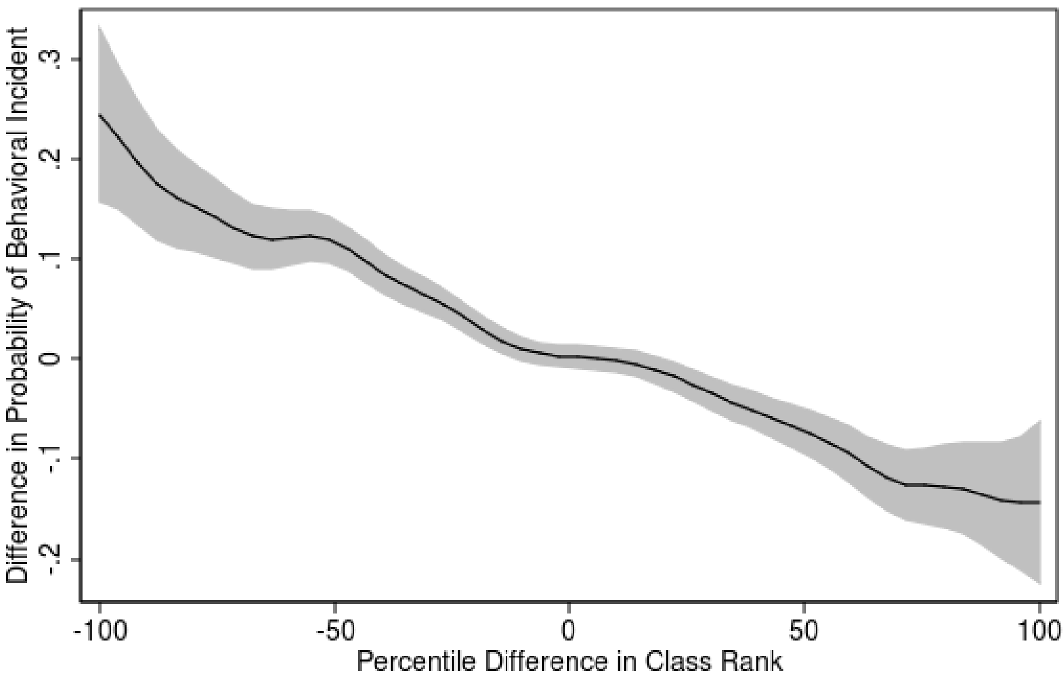


Figure 5: Evidence from New York City Public Schools



Notes: Panels show semiparametric estimates and the associated 95%-confidence intervals of the effect of a change in a student's class percentile rank (in going from elementary to middle school) on the change in an indicator variable for whether she was involved in a behavioral incident, cf. equation (7). The top panel constructs percentiles based on English/Language Arts (ELA) test scores, whereas the lower one uses math test scores. Estimates are obtained using the differencing procedure in Yatchew (1998) and local-mean smoothing with a Gaussian kernel and a bandwidth of 10. Section 3.2 and the Data Appendix provide additional information on the exact econometric specification as well as the sample.

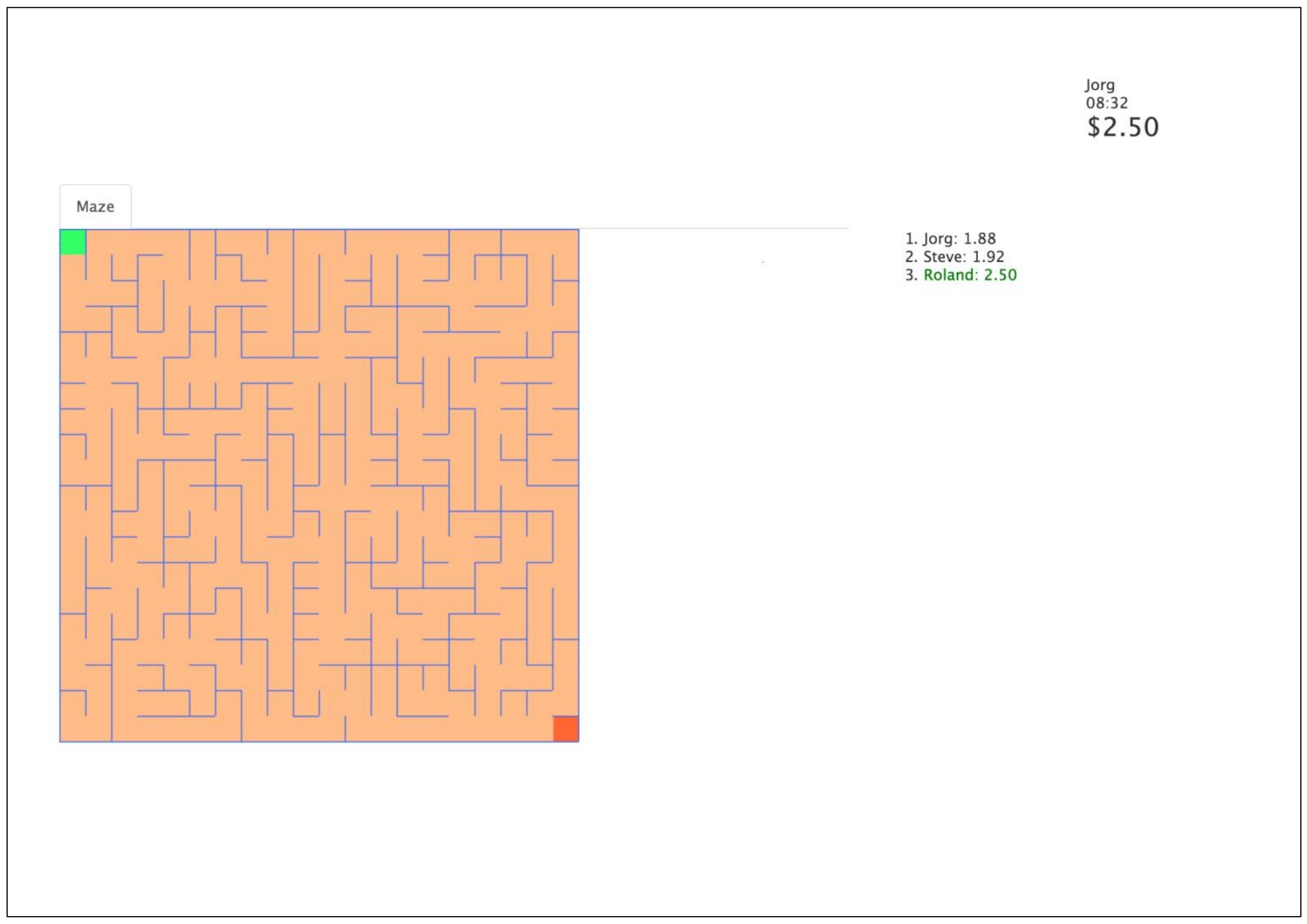
Figure 6: Evidence from the National Educational Longitudinal Study



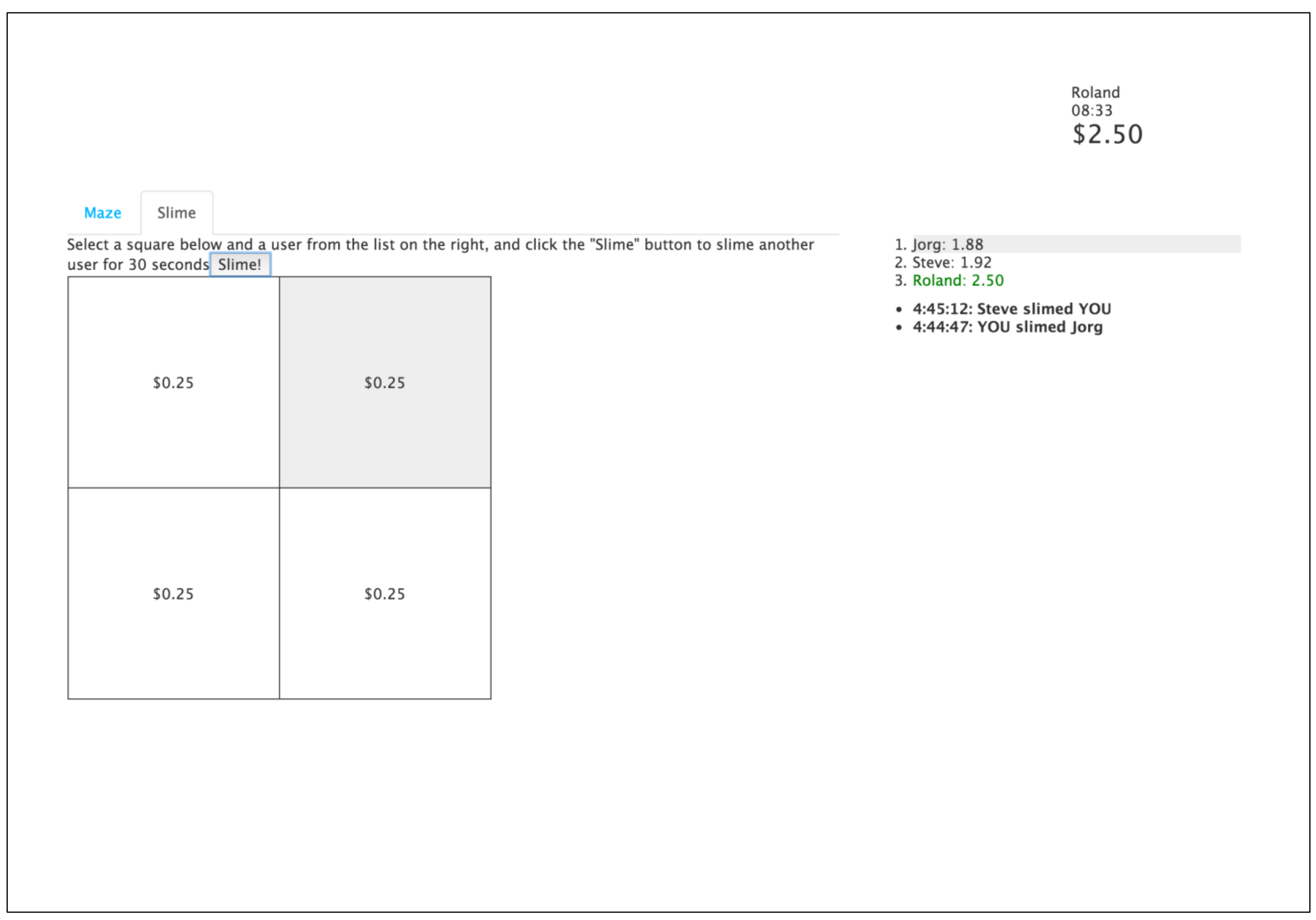
Notes: The figure shows semiparametric estimates and the associated 95%-confidence intervals of the effect of differences in a student's course-specific percentile rank on the difference in two course-specific behavioral outcomes, cf. equation (9). Estimates are obtained using the differencing procedure in Yatchew (1998) and local-mean smoothing with a Gaussian kernel and a bandwidth of 7.5. Section 3.3 and the Data Appendix provide additional information on the exact econometric specification as well as the sample.

Figure 7: Sample Screenshots from the Maze-Solving Experiment

A. Screen of a Participant in the Control Group



B. Sliming Tab in the Treatment Group



C. Screen of a Participant Who Got "Slimed"

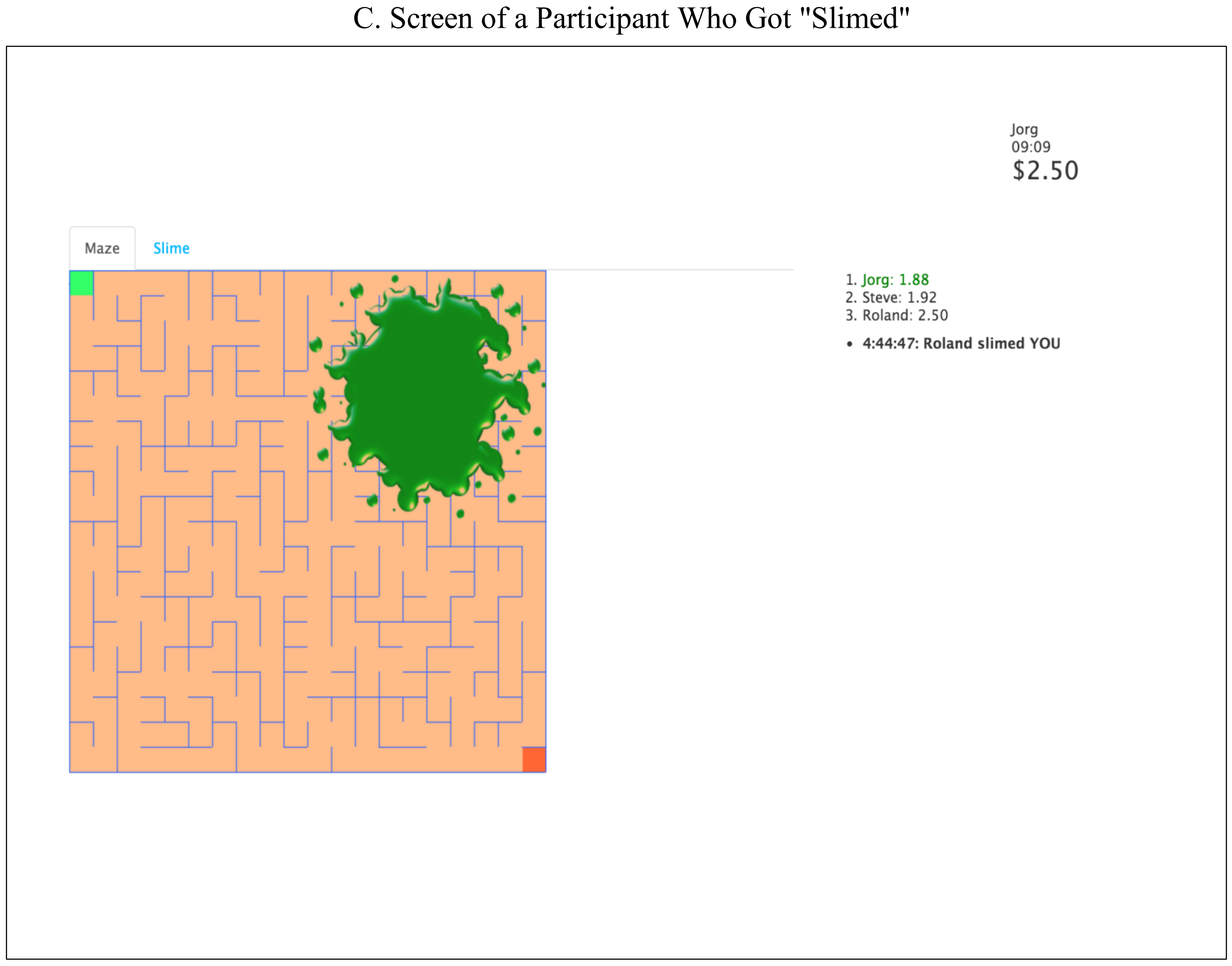


Table 1: Baseline School and Class Characteristics in the Experiment of Duflo et al. (2011), by Treatment Group

	All Schools				
	Nontracking Schools		Tracking Schools		*p* -value
	Mean	SD	Mean	SD	Tracking = Nontracking
School Characteristics at Baseline:					
Total Enrollment	589	(232)	549	(198)	.32
Number of Government Teachers	11.6	(3.3)	11.9	(2.8)	.62
Student/Teacher ratio	37.1	(12.2)	35.9	(10.1)	.56
Performance on National Exam (out of 400)	255.6	(23.6)	258.1	(23.4)	.57
Class Size at Baseline:					
Average Class Size	91	(37)	89	(33)	.76
Proportion of Female Students	.49	(.06)	.49	(.05)	.54

	Within Nontracking Schools				
	Section A: Assigned to Civil Service Teacher		Section B: Assigned to Contract Teacher		*p* -value
	Mean	SD	Mean	SD	Section A = Section B
Proportion Female	.49	(.06)	.49	(.06)	.89
Average Age at Endline	9.07	(.53)	9.00	(.45)	.45
Average Standardized Test Score at Baseline	.003	(.10)	.002	(.11)	.94
Average SD (within Section) of Test Scores at Baseline	1.005	(.08)	.993	(.08)	.43

	Within Tracking Schools				
	Bottom Section		Top Section		*p* -value
	Mean	SD	Mean	SD	Top = Bottom
Proportion Female	.49	(.09)	.50	(.08)	.38
Average Age at Endline	9.04	(.59)	9.41	(.60)	.00
Assigned to Contract Teacher	.53	(.49)	.46	(.47)	.44
Respected Assignment	.99	(.02)	.99	(.02)	.67
Average Standardized Test Score at Baseline	-.81	(.04)	.81	(.04)	.00
Average SD (within Section) of Test Scores at Baseline	.49	(.13)	.65	(.13)	.00

Notes: Table shows averages and standard deviations of selected characteristics of the 121 schools in the ETP experiment of Duflo et al. (2011). Of the 121 schools in the experiment 60 were randomly assigned to the "tracking" treatment, whereas the remaining 61 schools are classified as "nontracking." The rightmost column displays *p* -values for tests of equality across groups.

Source: Duflo et al. (2011)

Table 2: Estimates of the Impact of Percentile on Test Scores in the Experiment of Duflo et al. (2011)

A. Nontracking Schools Only

	Endline Test Score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Percentile (÷100)	.418	.641	.649	.472	.462	.639	.449
	(.269)	(.273)	(.265)	(.225)	(.230)	(.265)	(.225)
Test Score at Baseline	.374	.314	.320	.371	.365	.324	.379
	(.084)	(.086)	(.083)	(.073)	(.075)	(.083)	(.072)
Squared Test Score at Baseline	.015	.020	.020	.022	.025	.020	.022
	(.016)	(.016)	(.016)	(.015)	(.015)	(.016)	(.015)
Contract Teacher	.142	.139	.144	.134		.144	.134
	(.051)	(.048)	(.049)	(.045)		(.048)	(.045)
Peers' Mean Test Score		.550	.546	.438			
		(.206)	(.215)	(.205)			
Peers' Mean Test Score × Bottom Quarter						.518	.457
						(.322)	(.285)
Peers' Mean Test Score × Second Quarter						.599	.426
						(.364)	(.359)
Peers' Mean Test Score × Third Quarter						.263	.022
						(.353)	(.340)
Peers' Mean Test Score × Top Quarter						.793	.820
						(.408)	(.382)
Constant	-.309	-.424	-.230			-.226	
	(.168)	(.164)	(.250)			(.249)	
Additional Controls	No	No	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	No	No	No	Yes	No	No	Yes
Section Fixed Effects	No	No	No	No	Yes	No	No
R-Squared	.240	.243	.254	.390	.413	.255	.391
Number of Observations	2,190	2,190	2,188	2,188	2,188	2,188	2,188

B. All Schools

	Endline Test Score						
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Percentile (÷100)	.321	.388	.364	.251	.294		
	(.122)	(.129)	(.129)	(.120)	(.124)		
Percentile (÷100) × Nontracking School						.370	.445
						(.200)	(.195)
Percentile (÷100) × Tracking School × Top Section						.245	.296
						(.146)	(.147)
Percentile (÷100) × Tracking School × Bottom Section						.313	.372
						(.135)	(.136)
Test Score at Baseline	.390	.367	.385	.424	.408	.396	.372
	(.052)	(.054)	(.055)	(.052)	(.053)	(.060)	(.059)
Squared Test Score at Baseline	.016	.018	.019	.019	.022	.024	.028
	(.012)	(.012)	(.012)	(.011)	(.011)	(.013)	(.013)
Tracking School × Top Section	.240	-.012	.025				
	(.107)	(.133)	(.131)				
Tracking School × Bottom Section	.045	.391	.375	-.289		-.355	
	(.075)	(.134)	(.128)	(.225)		(.238)	
Contract Teacher	.178	.162	.167	.167		.167	
	(.038)	(.041)	(.041)	(.037)		(.037)	
Peers' Mean Test Score		.208	.186	-.046		-.042	
		(.074)	(.068)	(.073)		(.074)	
Constant	-.279	-.305	.005				
	(.097)	(.099)	(.168)				
Additional Controls	No	No	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	No	No	No	Yes	No	Yes	No
Section Fixed Effects	No	No	No	No	Yes	No	Yes
R-Squared	.251	.257	.269	.419	.450	.419	.451
Number of Observations	5,170	5,170	5,147	5,147	5,147	5,147	5,147

Notes: Entries are coefficients and standard errors from estimating equation (6) using ordinary least squares. Heteroskedasticity robust standard errors are clustered at the school level and presented in parentheses. The upper panel shows results for students who attend nontracking schools, whereas the lower panel restricts the sample to all students with nonmissing baseline test scores in the data of Duflo et al. (2011). Going from column (2) to (3) the number of observations decreases because some students are missing information on age and gender. "Additional Controls" include age, gender, whether the school is located in the Bungoma district, and whether it was sampled for school based management. "Bottom Quarter", "Second Quarter", etc. are indicator variables for students' own position in the test score distribution at baseline. See the Data Appendix or Duflo et al. (2011) for a definition of each variable.

Table 3: Summary Statistics for NYCPS Data

	Mean	SD
Behavioral Indicators:		
Behavioral Incident in 5th Grade	.083	(.531)
Behavioral Incident in 6th Grade	.089	(.285)
Behavioral Incident in 8th Grade	.132	(.339)
Test Scores:		
5th Grade Test Score (English / Language Arts)	661	(36.9)
5th Grade Test Score (Math)	668	(41.6)
Demographics:		
White	.149	(.356)
Black	.314	(.464)
Hispanic	.394	(.489)
Asian	.139	(.346)
Other race	.004	(.063)
Male	.507	(.500)
Female	.493	(.500)
Free lunch	.830	(.376)
English Language Learner	.093	(.290)
Special education	.087	(.282)
School Year:		
2004/05	.206	(.405)
2005/06	.194	(.395)
2006/07	.194	(.396)
2007/08	.201	(.401)
2008/09	.205	(.403)

Notes: Entries are means and standard deviations for each variable we use in the NYCPS data. For further details about the NYCPS data see the description in the Data Appendix.

Table 4: Estimates of the Short-Run Impact of Percentile on Behavior in the NYCPS Data

A. Percentile Based on ELA Test Scores

	Δ Behavioral Incident (Grade 5 → Grade 6)			
Independent Variable	OLS	2SLS	OLS	2SLS
Δ Percentile (÷100)	-.025	-.071	-.030	-.054
	(.005)	(.033)	(.005)	(.053)
Test Score at Previous School (÷100)	-.011	-.016	-.011	-.014
	(.001)	(.004)	(.001)	(.005)
Male	.026	.024	.026	.024
	(.003)	(.003)	(.003)	(.003)
Black	.014	.009	.033	.030
	(.006)	(.007)	(.005)	(.005)
Hispanic	-.005	-.007	.008	.007
	(.004)	(.005)	(.003)	(.004)
Asian	-.022	-.021	-.010	-.010
	(.004)	(.004)	(.004)	(.004)
Other Race	-.019	-.023	.006	.004
	(.015)	(.015)	(.014)	(.014)
Free Lunch	.011	.010	.011	.011
	(.003)	(.003)	(.003)	(.003)
English Language Learner	-.010	-.012	-.011	-.012
	(.006)	(.007)	(.006)	(.007)
Special Education	.015	.008	.009	.007
	(.006)	(.008)	(.005)	(.009)
Year Fixed Effects	Yes	Yes	Yes	Yes
School Fixed Effects	No	No	Yes	Yes
First Stage F-Statistic	---	488.3	---	330.2
Shea's Partial R-Squared	---	.031	---	.006
R-Squared	.009	--	.067	--
Number of Observations	122,812	118,750	122,812	118,750

B. Percentile Based on Math Test Scores

	Δ Behavioral Incident (Grade 5 → Grade 6)			
Independent Variable	OLS	2SLS	OLS	2SLS
Δ Percentile (÷100)	-.039	-.065	-.051	-.075
	(.006)	(.028)	(.005)	(.047)
Test Score at Previous School (÷100)	-.010	-.013	-.011	-.013
	(.001)	(.003)	(.001)	(.004)
Male	.028	.027	.027	.026
	(.003)	(.003)	(.003)	(.003)
Black	.012	.009	.031	.029
	(.006)	(.006)	(.004)	(.005)
Hispanic	-.004	-.005	.008	.007
	(.004)	(.005)	(.003)	(.003)
Asian	-.017	-.016	-.007	-.007
	(.004)	(.004)	(.004)	(.004)
Other Race	-.021	-.025	.003	.000
	(.014)	(.014)	(.013)	(.013)
Free Lunch	.012	.012	.012	.011
	(.003)	(.003)	(.002)	(.003)
English Language Learner	-.007	-.007	-.008	-.008
	(.005)	(.005)	(.004)	(.004)
Special Education	.013	.010	.008	.005
	(.006)	(.007)	(.005)	(.007)
Year Fixed Effects	Yes	Yes	Yes	Yes
School Fixed Effects	No	No	Yes	Yes
First Stage F-Statistic	---	526.6	---	389.0
Shea's Partial R-Squared	---	.038	---	.008
R-Squared	.008	--	.065	--
Number of Observations	131,312	126,976	131,312	126,796

Notes: Entries are coefficients and standard errors from estimating the linear model in equation (8) by ordinary least squares as well as two-stage least squares. The dependent variable is listed at the top of each column. The instrument for Δ Percentile is the predicted change in percentile based on school zoning regulations, as explained in the text. The IV specifications contain fewer observations because we do not observe addresses of all students in our data. In the upper panel a student's percentile in his school is calculated based on ELA test scores, whereas the lower panel uses math test scores. Heteroskedasticity robust standard errors are clustered on the school level and reported in parentheses. In addition to the variables shown in the table, indicator variables for missing values of each covariate are also included in the regressions. To facilitate comparisons between Tables 4 and 5 the set of students included in the analysis has been restricted to those observed from fifth through eight grade. See the Data Appendix for the precise definition and source of each variable.

Table 5: Estimates of the Medium-Run Impact of Percentile on Behavior in the NYCPS Data

A. Percentile Based on ELA Test Scores

	Δ Behavioral Incident (Grade 5 → Grade 8)			
Independent Variable	OLS	2SLS	OLS	2SLS
Δ Percentile (÷100)	-.049	-.095	-.057	-.144
	(.008)	(.038)	(.006)	(.070)
Test Score at Previous School (÷100)	-.030	-.035	-.029	-.037
	(.002)	(.005)	(.002)	(.007)
Male	.037	.035	.036	.033
	(.003)	(.003)	(.003)	(.004)
Black	.030	.026	.048	.046
	(.008)	(.009)	(.006)	(.007)
Hispanic	.002	-.001	.012	.011
	(.006)	(.007)	(.005)	(.005)
Asian	-.048	-.047	-.036	-.034
	(.005)	(.005)	(.005)	(.005)
Other Race	-.009	-.010	.011	.011
	(.020)	(.020)	(.018)	(.019)
Free Lunch	.025	.024	.028	.026
	(.004)	(.004)	(.004)	(.004)
English Language Learner	-.016	-.016	-.017	-.020
	(.008)	(.009)	(.007)	(.009)
Special Education	.004	-.001	.000	-.011
	(.007)	(.009)	(.006)	(.011)
Year Fixed Effects	Yes	Yes	Yes	Yes
School Fixed Effects	No	No	Yes	Yes
First Stage F-Statistic	---	488.3	---	330.2
Shea's Partial R-Squared	---	.031	---	.006
R-Squared	.021	--	.070	--
Number of Observations	122,812	118,750	122,812	118,750

B. Percentile Based on Math Test Scores

	Δ Behavioral Incident (Grade 5 → Grade 8)			
Independent Variable	OLS	2SLS	OLS	2SLS
Δ Percentile (÷100)	-.070	-.050	-.086	-.014
	(.009)	(.032)	(.006)	(.059)
Test Score at Previous School (÷100)	-.031	-.029	-.030	-.025
	(.002)	(.003)	(.002)	(.004)
Male	.041	.041	.040	.041
	(.003)	(.003)	(.003)	(.003)
Black	.027	.028	.047	.048
	(.008)	(.008)	(.006)	(.006)
Hispanic	.002	.003	.012	.014
	(.006)	(.006)	(.005)	(.005)
Asian	-.037	-.038	-.029	-.032
	(.005)	(.005)	(.005)	(.005)
Other Race	-.015	-.014	.007	.010
	(.019)	(.020)	(.018)	(.018)
Free Lunch	.028	.028	0.029	.030
	(.004)	(.004)	(.004)	(.004)
English Language Learner	-.017	-.016	-.017	-.015
	(.005)	(.005)	(.004)	(.004)
Special Education	.005	.006	.001	.007
	(.007)	(.008)	(.006)	(.008)
Year Fixed Effects	Yes	Yes	Yes	Yes
School Fixed Effects	No	No	Yes	Yes
First Stage F-Statistic	---	526.6	---	389.0
Shea's Partial R-Squared	---	.038	---	.008
R-Squared	.021	--	.069	--
Number of Observations	131,312	126,976	131,312	126,796

Notes: Entries are coefficients and standard errors from estimating the linear model in equation (8) by ordinary least squares as well as two-stage least squares. The dependent variable is listed at the top of each column. The instrument for Δ Percentile is the predicted change in percentile based on school zoning regulations, as explained in the text. The IV specifications contain fewer observations because we do not observe addresses of all students in our data. In the upper panel a student's percentile in his school is calculated based on ELA test scores, whereas the lower panel uses math test scores. Heteroskedasticity robust standard errors are clustered on the school level and reported in parentheses. In addition to the variables included in the table, indicator variables for missing values of each covariate are also included in the regressions. To facilitate comparisons between Tables 4 and 5 the set of students included in the analysis has been restricted to those observed from fifth through eight grade. See the Data Appendix for the precise definition and source of each variable.

Table 6: Summary Statistics for NELS Data

	Mean	SD
Behavioral Indicators:		
Behavioral Incident - 8th Grade, English	.373	(.484)
Behavioral Incident - 8th Grade, History	.368	(.482)
Behavioral Incident - 8th Grade, Math	.379	(.485)
Behavioral Incident - 8th Grade, Science	.380	(.486)
Behavioral Incident - 10th Grade, English	.539	(.499)
Behavioral Incident - 10th Grade, History	.532	(.499)
Behavioral Incident - 10th Grade, Math	.499	(.499)
Behavioral Incident - 10th Grade, Science	.554	(.497)
Test Scores:		
Mean Test Score, 8th Grade	.088	(.867)
Mean Test Score, 10th Grade	.116	(.870)
Student Demographics:		
Male	.498	(.500)
Female	.502	(.500)
White	.705	(.456)
Black	.103	(.304)
Hispanic	.114	(.318)
Asian	.058	(.234)
Other race	.019	(.137)
English Language Learner	.025	(.156)
Bottom Socioeconomic Quartile	.219	(.414)
Second Socioeconomic Quartile	.240	(.427)
Third Socioeconomic Quartile	.237	(.425)
Highest Socioeconomic Quartile	.304	(.460)
Parent Characteristics:		
Married	.731	(.444)
Divorced	.103	(.304)
Separated	.031	(.174)
Never Married	.020	(.140)
Other Marital Status	.115	(.320)
High School Dropout	.090	(.286)
High School Graduate	.189	(.391)
Some College	.399	(.490)
College Graduate	.159	(.365)
Postgraduate Degree	.100	(.300)
Doctoral Degree	.064	(.245)
School Characteristics:		
Public School	.769	(.421)
Catholic School	.114	(.318)
Independent / Other Private School	.117	(.321)
Urban Area	.285	(.452)
Suburban Area	.422	(.494)
Rural Area	.293	(.455)

Notes: Entries are weighted means and standard deviations for each variable we use in the NELS data. For further details about the NELS data see the Data Appendix.

Table 7: Observable Characteristics of Students in our Framed Field Experiment, by Treatment Status

	All		Treatment		Control		*p*-value
	Mean	SD	Mean	SD	Mean	SD	Treatment = Control
Male	.538	(.499)	.537	(.499)	.538	(.499)	.971
Minority	.873	(.333)	.866	(.341)	.882	(.324)	.843
Grade	6.89	(.762)	6.97	(.715)	6.80	(.803)	.077
Special Education	.134	(.341)	.134	(.341)	.134	(.341)	.988
Limited English Proficiency	.334	(.472)	.339	(.474)	.328	(.470)	.864
Missing Demographic Information	.023	(.149)	.033	(.180)	.013	(.115)	.103
Self-Assessed Ability (pre-period; scale 1–10)	6.08	(1.83)	6.19	(1.79)	5.96	(1.87)	.246
Baseline Performance (seconds per compl. maze)	27.1	(15.8)	26.2	(18.5)	28.0	(12.1)	.342
Number of Students	573		302		271		

Notes: Table shows basic descriptive statistics for students who participated in our framed field experiment, by treatment status. The rightmost column displays *p*-values for tests of equality in means across the treatment and control groups. A Kolmogorov-Smirnov test is unable to reject the null hypothesis that the *p*-values in the rightmost column are uniformly distributed on the unit interval (*p=.305*) For additional information on the experiment see the main text. The Data Appendix provides precise definitions of all variables.

Table 8: Experimental Results

	Willingness to Pay for Practicing		Total Money Spent on Practicing		Total Money Spent on Sliming	
	(1)	(2)	(3)	(4)	(5)	(6)
Percentile (÷100)	-.089	-.103	-.061	-.081		
	(.035)	(.037)	(.074)	(.073)		
Percentile (÷100) × Treatment	.075	.089	.143	.165	-.371	-.390
	(.041)	(.042)	(.092)	(.092)	(.234)	(.165)
H_0: Coefficient on Percentile = 0	.022	.009	.412	.289		
H_0: Coefficient on Percentile × Treatment = 0	.078	.045	.142	.090	.035	.004
Controls	No	Yes	No	Yes	No	Yes
Experimental Session Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Sample	Treatment & Control	Treatment & Control	Treatment & Control	Treatment & Control	Treatment	Treatment
Mean of Dependent Variable	.201	.201	.204	.204	.137	.137
R-Squared	.138	.155	.160	.172	.110	.142
Number of Observations	573	573	573	573	302	302

Notes: Entries are coefficients and standard errors from estimating the linear model in equation (10) by ordinary least squares. The dependent variables are listed at the top of each column. All specifications control for baseline performance and experimental session fixed effects. Additional controls include gender, grade, a minority indicator, special education status, limited English proficiency, self-assessed ability, and indicator variables for missing demographic information. Heteroskedasticity robust standard errors are clustered by experimental session and reported in parentheses. To account for the small number of clusters, reported *p* -values are based on the wild bootstrap procedure suggested by Cameron et al. (2008) with 10,000 iterations. See the Data Appendix for the precise definition and source of each variable.

Table A.1: Experimental Results, by Grade

A. 6th Graders

	Willingness to Pay for Practicing		Total Money Spent on Practicing		Total Money Spent on Sliming	
	(1)	(2)	(3)	(4)	(5)	(6)
Percentile (÷100)	-.128	-.121	-.032	-.047		
	(.039)	(.054)	(.093)	(.109)		
Percentile (÷100) × Treatment	.060	.055	.438	.461	-.391	.075
	(.064)	(.066)	(.240)	(.238)	(.397)	(.153)
H_0: Coefficient on Percentile = 0	.003	.042	.704	.629		
H_0: Coefficient on Percentile × Treatment = 0	.375	.443	.077	.074	.488	.491
Controls	No	Yes	No	Yes	No	Yes
Experimental Session Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Sample	Treatment & Control	Treatment & Control	Treatment & Control	Treatment & Control	Treatment	Treatment
Mean of Dependent Variable	.180	.180	.196	.196	.163	.163
R-Squared	.159	.201	.195	.212	.173	.235
Number of Observations	196	196	196	196	80	80

B. 7th Graders

	Willingness to Pay for Practicing		Total Money Spent on Practicing		Total Money Spent on Sliming	
	(7)	(8)	(9)	(10)	(11)	(12)
Percentile (÷100)	-.199	-.190	-.062	-.075		
	(.071)	(.062)	(.198)	(.177)		
Percentile (÷100) × Treatment	.089	.085	.039	.040	.252	-.118
	(.052)	(.049)	(.156)	(.151)	(.514)	(.389)
H_0: Coefficient on Percentile = 0	.015	.005	.761	.694		
H_0: Coefficient on Percentile × Treatment = 0	.117	.102	.802	.794	.790	.782
Controls	No	Yes	No	Yes	No	Yes
Experimental Session Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Sample	Treatment & Control	Treatment & Control	Treatment & Control	Treatment & Control	Treatment	Treatment
Mean of Dependent Variable	.209	.209	.200	.200	.120	.120
R-Squared	.360	.371	.354	.360	.131	.268
Number of Observations	229	229	229	229	146	146

C. 8th Graders

	Willingness to Pay for Practicing		Total Money Spent on Practicing		Total Money Spent on Sliming	
	(13)	(14)	(15)	(16)	(17)	(18)
Percentile (÷100)	-.060	-.070	-.023	-.065		
	(.089)	(.092)	(.118)	(.149)		
Percentile (÷100) × Treatment	.082	.104	.139	.209	-.305	-.255
	(.105)	(.097)	(.150)	(.233)	(.341)	(.274)
H_0: Coefficient on Percentile = 0	.542	.499	.847	.707		
H_0: Coefficient on Percentile × Treatment = 0	.447	.321	.379	.503	.479	.442
Controls	No	Yes	No	Yes	No	Yes
Experimental Session Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Sample	Treatment & Control	Treatment & Control	Treatment & Control	Treatment & Control	Treatment	Treatment
Mean of Dependent Variable	.219	.219	.232	.232	.140	.140
R-Squared	.230	.254	.191	.212	.254	.398
Number of Observations	135	135	135	135	72	72

Notes: Entries are coefficients and standard errors from estimating the linear model in equation (10) separately by grade. The dependent variables are listed at the top of each column. All specifications control for baseline performance and experimental session fixed effects. Additional controls include gender, a minority indicator, special education status, limited english proficiency, self-assessed ability, and indicator variables for missing demographic information. Heteroskedasticity robust standard errors are clustered by experimental session and reported in parentheses. To account for the small number of clusters, reported p-values are based on the wild bootstrap procedure suggested by Cameron et al. (2008) with 10,000 iterations. See the Data Appendix for the precise definition and source of each variable.

Table A.2: Summary Statistics for ECLS Data

	Mean	SD
Behavioral Indicators:		
Any behavioral incident, English	0.277	(0.447)
Any behavioral incident, Math	0.284	(0.451)
Test Scores:		
Standardized ARS Reading	0.000	(1.000)
Standardized ARS Math	0.000	(1.000)
Item Response Score: Math	65.57	(43.002)
Item Response Score: Reading	82.02	(51.537)
Student Demographics:		
Male	0.504	(0.500)
Female	0.496	(0.500)
White	0.600	(0.490)
Black	0.135	(0.342)
Hispanic	0.149	(0.356)
Asian	0.057	(0.233)
Other race	0.058	(0.233)
Speaks English at Home	0.893	(0.309)
Bottom Socioeconomic Quintile	0.153	(0.360)
Second Socioeconomic Quintile	0.189	(0.392)
Third Socioeconomic Quintile	0.204	(0.403)
Fourth Socioeconomic Quintile	0.218	(0.413)
Fifth Socioeconomic Quintile	0.236	(0.424)
Parent Characteristics:		
Biological Mother and Biological Father	0.675	(0.468)
Biological Mother and Other Father	0.080	(0.272)
Other Mother and Biological Father	0.009	(0.092)
Biological Mother only	0.181	(0.385)
Biological Father only	0.018	(0.133)
Adoptive Parents	0.014	(0.118)
Guardians	0.023	(0.150)
Mother's Education: none - very low	0.113	(0.317)
Mother's Education: High School	0.301	(0.459)
Mother's Education: Voc/tech Program	0.055	(0.228)
Mother's Education: Some College	0.276	(0.447)
Mother's Education: Bachelor's Degree	0.168	(0.374)
Mother's Education: Professional Degree	0.087	(0.281)
Father's Education: none - very low	0.110	(0.313)
Father's Education: High School	0.310	(0.462)
Father's Education: Voc/tech Program	0.052	(0.222)
Father's Education: Some College	0.214	(0.410)
Father's Education: Bachelor's Degree	0.184	(0.387)
Father's Education: Professional Degree	0.130	(0.336)
School Type:		
Catholic	0.123	(0.329)
Other Religious	0.062	(0.241)
Other Private	0.028	(0.165)
Public	0.787	(0.410)
Location Characteristics:		
Suburb	0.233	(0.423)
City	0.767	(0.423)

Notes: Entries are means and standard deviations for each variable we use in the ECLS data. For further details about the ECLS data see the description in the Data Appendix.

Table A.3: Exploring the Relationship between Percentile and Subjective Teacher Assessments in the ECLS, all Waves

A. Percentile and Subjective Teacher Assessments

	Δ Teacher Assessment				
Coefficient on Δ Percentile (÷100):					
Fall, Kindergarten Year	0.052	-0.011	0.024	0.011	0.019
	(0.033)	(0.048)	(0.054)	(0.054)	(0.055)
	10,571	10,571	10,571	10,571	10,571
Spring, Kindergarten Year	0.098	-0.041	-0.025	-0.045	-0.054
	(0.025)	(0.034)	(0.035)	(0.035)	(0.037)
	15,209	15,209	15,209	15,209	15,209
Spring, 1st Grade Year	0.208	0.031	0.049	0.017	0.007
	(0.033)	(0.039)	(0.039)	(0.040)	(0.040)
	13,376	13,376	13,376	13,376	13,376
Spring, 3rd Grade Year	0.035	0.047	0.061	0.067	0.066
	(0.044)	(0.044)	(0.045)	(0.046)	(0.046)
	9,063	9,063	9,063	9,063	9,063
Spring, 5th Grade Year	0.104	-0.009	-0.021	-0.007	-0.011
	(0.087)	(0.091)	(0.092)	(0.092)	(0.092)
	4,003	4,003	4,003	4,003	4,003
Spring, 8th Grade Year	0.074	-0.123	-0.129	-0.128	-0.133
	(0.104)	(0.120)	(0.123)	(0.123)	(0.124)
	2,659	2,659	2,659	2,659	2,659
Baseline Controls	Yes	Yes	Yes	Yes	Yes
Subject Score Polynomial:					
First Order	Yes	No	No	No	No
Second Order	No	Yes	No	No	No
Third Order	No	No	Yes	No	No
Fourth Order	No	No	No	Yes	No
Fifth Order	No	No	No	No	Yes

B. Percentile and Reports of Behavioral Incidents

	Δ Behavioral Incident				
Δ Percentile (÷100)	-0.049	-0.065	-0.100	-0.101	-0.100
	(0.056)	(0.067)	(0.066)	(0.066)	(0.066)
Baseline Controls	Yes	Yes	Yes	Yes	Yes
Subject Score Polynomial:					
First Order	Yes	No	No	No	No
Second Order	No	Yes	No	No	No
Third Order	No	No	Yes	No	No
Fourth Order	No	No	No	Yes	No
Fifth Order	No	No	No	No	Yes
R-Squared	0.293	0.294	0.296	0.296	0.296
Number of Observations	2566	2566	2566	2566	2566

Notes: Entries are coefficients and standard errors from estimating the linear model in equation (13) by ordinary least squares. The dependent variable in the upper panel is the difference in standardized ARS score between reading and math subjects. Subject specific ARS scores are standardized by wave and grade, restricted to students in the sample on which regressions are run for that particular wave. The dependent variable in the lower panel is the difference in teacher reported behavioral incidents across subjects. Heteroskedasticity robust standard errors are clustered on the school level and reported in parentheses. In addition to the usual baseline covariates, each column includes a higher order of subject score poynomial. See the Data Appendix for the precise definition and source of each variable.