Self-Selection and Comparative Advantage in Social Interactions

Online Appendix
(Not for Publication)

November 2016

Contents

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 $Y(r)$ be an indicator variable equal to one if individual r chooses to be a nerd in the old environment (and zero otherwise), and let $Y'(r)$ denote r's choice in the new environment. Then the average treatment effect from manipulating the environment for all r is equal to

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

where r^* and $r^{*\prime}$ 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 is 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 $Y_1(r|\Xi)$ denote r's test score (conditional on environmental variables Ξ) if r chooses to be a nerd, whereas $Y_0(r|\Xi)$ 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

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

where $r^*(\cdot)$ 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).

B Reconciling the Literature on Social Interactions through the Lens of a Roy Model

There is a large empirical literature on peer effects in schools, neighborhoods, and other venues in which individuals interact. Surprisingly, research designs which exploit experimental and quasi-experimental variation often point in conflicting directions, despite comparable samples. In this appendix, we argue that our Roy model of social interactions is flexible enough to reconcile the seemingly disparate evidence. Put differently, we show that a model in which "peer effects" are due to the systematic sorting of individuals *within* social markets can produce the same empirical patterns that have traditionally been attributed to models in which peer effects take the form of direct externalities.

We divide the empirical literature on peer effects into four mutually exclusive categories: analyses that report no significant peer effects, effects which are linear and positive (i.e. smarter peers increase achievement), effects that are non-linear and positive, and analyses that find negative peer effects.

In what follows, we assume that production functions are concave in labor inputs and provide sufficient conditions for our model to reconcile the findings of various studies.[1] We do not attempt to explain every nuance in the empirical literature on peer effects. For sure, there exist several competing explanations all of which can explain some aspect in isolation. For instance, it has been suggested to us that failure to properly account for endogeneity of regressors may also rationalize contradictory results. While endogeneity concerns are definitely important, we do not believe that they are the whole story, especially since even true randomized experiments draw widely different conclusions regarding the sign and importance of peer effects (see, e.g., Cullen et al. (2006), Duflo et al. (2011), Kling et al. (2007), Carrell et al. (2009), Carrell et al. (2013)). Instead, our goal is to develop a tractable model which can reconcile broad but seemingly disparate findings.

B.1 No Peer Effects

One strand of the literature argues that peer effects are negligible (Angrist and Lang (2004), Cullen et al. (2006), Evans et al. (1992), Lefgren (2004), Lyle (2007), Stinebrickner and Stinebrickner (2006)). Angrist and Lang (2004) evaluate Boston's Metco program, which buses minority students from high poverty neighborhoods in Boston to wealthier suburban schools. Their results indicate that, although the new Metco students are on average lower achieving, the change in peer group induced by these students does not affect test scores of elementary and middle school students in the suburban schools. Cullen et al. (2006) analyze roughly fifteen thousand students who applied to nineteen schools through the Chicago Public Schools choice program. Using data from lotteries, their results imply that the academic impact of attending a new school with higher-performing peers is very small, at best.

Using solely the lens and language of our model, this implies that the marginal student, r^*, in these settings remains the same notwithstanding a change in peer group composition. Hence, the choice of sector for all other individuals does not change as well. To see why this might be the case in practice, consider Boston's Metco program (cf. Angrist and Lang (2004)). A potential explanation for the lack of peer effects in this study is that Metco students are few relative to non-Metco ones. Despite the fact that average ability declines and both the "supply" and "demand" curves shift downward, the impact of Metco students on relative wages is likely small. As the relative ability of their non-Metco peers remains the same, our model predicts that almost none of them change sectors, leading to negligible peer effects.

Appendix Figure A.1 illustrates this point. Imagine an increase in the number of Metco students, which shifts the supply curve downward (from σ to σ') and the demand side equilibrium schedule from δ inward to δ'. Significant shifts, however, only occur to the left of the initial equilibrium. If this is indeed what happened, then it is not surprising that the marginal student would remain virtually the same.

[1] It is straightforward to conduct an identical analysis under the assumption of convex production functions, which we leave to the reader.

B.2 Linear Peer Effects

Another portion of the peer effects literature shows that peer effects operate linearly, typically based on mean group characteristics. Hanushek et al. (2003) show, in a large matched panel data set of third through sixth graders in Texas public schools, that a one standard deviation increase in mean peer test score results in a .20 standard deviation increase in own test scores. Hoxby (2000) uses year to year variation in class-level gender and race composition; finding effects that range from .15 to .40 points for every one point increase in the class mean reading score.[2]

Through the lens of our model, this implies that a constant fraction of individuals shift sectors for every one unit increase in peers' mean test score. Appendix Figure A.2 shows an example in which peer effects would operate linearly. In this example, the supply schedule (i.e. the distribution of relative ability) undergoes a nearly parallel shift close to the initial equilibrium. Furthermore, the demand curve has constant negative slope around the initial equilibrium and is relatively unresponsive to fluctuations in labor supply. Therefore, changes in peer ability lead to constant changes in relative wages, and a constant fraction of individuals switches sectors—resulting in linear peer effects.[3]

An explanation along these lines may explain the results of Hoxby (2000). As Hoxby (2000) identifies peer effects through plausibly random variation in gender and race composition in classrooms, it may be reasonable to assume that, over the relevant range, the ability distribution shifts one-to-one with its mean. Moreover, given the limited variation in cohorts' gender and racial composition, the demand schedule might be approximately linear in a neighborhood around the initial equilibrium. Of course, whether these or equivalent conditions do indeed hold in Hoxby (2000), or in other analyses which also report linear peer effects, is unknown.

B.3 Heterogeneous Positive Effects

A third category of the literature describes positive, but non-linear peer effects. Hoxby and Weingarth (2005), for instance, exploit a desegregation program in Wake County, NC, which produces exogenous changes in classroom peer groups. They find that based on a linear-in-means model a student's test score is expected to increase .25 standard deviations given a 1 standard deviation increase in peers' mean score. However, when they allow their results to differ depending on the decile of peer performance, they show that students benefit more from peers with an achievement level similar to theirs. For example, students in the bottom decile benefit most from the addition of students in the second and third deciles (a 10% increase in peers at the 15th percentile increases their performance by .19 standard deviations more than an additional 10% of students in the 8th decile). Carrell et al. (2009) investigate peer effects among freshmen at the Air Force Academy who are randomly assigned to squadrons. They demonstrate that a one standard deviation increase in peers' average verbal SAT score results in a .565 standard deviation increase in freshman fall GPA for students in the bottom third of the expected achievement distribution, compared to .361 and .312 for those in the

[2]Other contributions in this vein include Boozer and Cacciola (2001), Gaviria and Raphael (2001), Kang (2007), and Goux and Maurin (2007).

[3]We emphasize that the conditions we provide in the text are sufficient, but in no way necessary.

middle and top third. Relying on Census data, Crane (1991) argues that the fraction of high-status workers in a neighborhood is negatively related to the likelihood of teen pregnancy and dropping out of school. These effects become much stronger at the lowest levels of high-status workers.[4]

Under the auspices of our model, Appendix Figure A.3 considers a scenario consistent with the results of Crane (1991). The left panel depicts the situation in a neighborhood with a large number of high achievers (e.g., nerds in the language of the previous section). An increase in the presence of highly skilled individuals (shifting σ to σ') marginally decreases r^*. The right panel features an identical inward shift of the supply curve, but a much larger increase in the equilibrium size of the nerd sector. The larger effect for a given change in supply depicted in the right panel is due to demand being more elastic around the equilibrium featuring fewer nerds.

B.4 Heterogeneous Negative Effects

In stark contrast to the aforementioned studies, a nascent literature provides credible evidence that own achievement might decline in peer quality. Lavy et al. (2009) use a sample of over a million students taking British age-14 tests to examine peer effects in English high schools. Exploiting the fact that students in their sample enter high school, and thus encounter a peer group that is 87% new on average, they demonstrate that peer effects are different for boys and girls. Girls are positively affected by peers in the top 5% (.07 standard deviation for a 10% increase) while boys are negatively affected (-.05 standard deviations). For boys, the negative effects are strongest among those at the top of the achievement distribution.

Based on the non-linear results in Carrell et al. (2009), Carrell et al. (2013) implemented an experiment at the US Air Force Academy aimed at increasing the GPA of incoming freshmen who were predicted to fall in the bottom tercile of the distribution. To achieve this goal squadrons in the treatment group were negatively sorted, while the composition of those in the control group continued to be random. Yet, the experiment did not have the intended effect. Students in the treatment group projected to fall in the bottom tercile, i.e. those students the experiment was designed to help, experienced a small *decline* in GPA compared to their counterparts in the control group.

One possible explanation for these perplexing results is illustrated in Figure A.4. In the left panel, we consider a candidate relative supply distribution σ' when, due to random assignment, a squadron has a larger fraction of academically able cadets. Compared to the distribution of the cohort depicted in bold lines, the resulting equilibrium consists of more nerds. That is, some otherwise low achieving students chose become nerds due to a change in market prices. However, when squadrons are arranged according to negative sorting, the supply curve becomes S-shaped (since the middle tercile of students is removed). Thus, as shown in the panel on the right, even a mean-preserving spread of the skill distribution may result in social dynamics with a lower fraction of high achievers, i.e. nerds. Indeed, Carrell et al. (2013) present suggestive evidence that their results are due to students endogenously

[4]Similar non-linear peer effects are found in Cooley-Fruehwirt (2010), Ding and Lehrer (2007), Duflo et al. (2011), Figlio (2007), Imberman et al. (2012), Zimmer and Toma (2000), and Zimmerman (2003).

sorting into groups *within* the squadron—as predicted by our Roy model.

Another striking example of heterogenous peer effects comes from the Moving to Opportunity (MTO) experiment. MTO provided housing vouchers for families in high poverty areas of Baltimore, Chicago, Los Angeles, and New York City, enabling them to relocate to lower poverty neighborhoods (Kling et al. (2005), Kling et al. (2007)). Evaluations of MTO show that female youth benefited from living in an ostensibly better neighborhood. Relative to the control group, female youth are 6.9% less likely to have ever had anxiety symptoms, are 9.1% less likely to have consumed alcohol during the past month, and have .08 fewer lifetime arrests for violent crimes. In stark contrast, male youth were affected negatively. Relative to the control group, male youth are 8.7% more likely to have had serious nonsports accidents, are 10.3% more likely to have smoked during the past month, and have .15 more lifetime arrests for property crimes.

Our model would predict the findings in MTO if *relative* wages increased (compared to the old neighborhood) for boys, but decreased for girls. Figure 8 depicts shifts of the supply and demand curves which could produce such a result. The top two panels refer to boys and the bottom two to girls. The panels on the left illustrate the conditions in the pre-treatment neighborhood and the panels on the right demonstrate an equilibrium in the treatment neighborhood. The set of students who switch sectors is bounded by r^*, the marginal individual in the old neighborhood, and by $\widetilde{r}$, the counterfactual marginal individual given the relative wages in the new environment. In moving to the new neighborhood there are two opposing effects. The nerdier population tends to shift the δ-schedule outwards, while at the same time greater educational resources and supervision in the new neighborhood increase the relative returns to studying, which would shift the δ-schedule inwards. In contrast to models that predict a positive influence from the new environment, here the direction of the treatment effect is *a priori* indeterminate.

In Appendix Figure A.5, the inward shift in supply by going from one environment to the other is the same for boys and girls. That is, boys and girls in the experimental group face the same set of smarter peers after they move. If, however, the relative demand curve for girls shifts sufficiently far inward to overcome the outward shift in relative supply, then relative equilibrium wages move in opposite directions. At least in principle, a smaller shift in the demand curve for boys might be caused by multiple factors. For instance, if the the nerd production function for girls is less concave than that for boys, then girls will be less likely to be crowded out of the nerd group in a new environment with greater supply. Conversely, if the troublemaking production function for boys is more concave than that for girls, then causing trouble will command a higher relative return in the new neighborhood where such skills are scarce.

C 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

$$u\left(T;\mathbf{X}_i,\mathbf{Z}_m\right)-u\left(N;\mathbf{X}_i,\mathbf{Z}_m\right)=\kappa+\mathbf{X}_i'\boldsymbol{\beta}_0+\mathbf{Z}_m'\boldsymbol{\gamma}_0+\alpha_0\overline{y}_m+\nu_{im}\geq 0. \tag{1}$$

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:

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

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).[5] While applied researchers have often found clever strategies to deal with these two problems, group-level unobservables have received much less attention.[6]

[5] 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).

[6] However, there are several notable exceptions. 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.

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

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

and consider the case in which there are no endogenous peer effects, i.e. $\alpha_0 = 0$. Assuming that $\text{Cov}\left(\mathbf{X}_i^*, \nu_{im}\right) = 0$ and applying the Frisch–Waugh Theorem (Frisch and Waugh 1933) to equation (1) with y_i as the left hand-side variable, the probability limit of the ordinary least 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 (1), individual i in social market m chooses to become a troublemaker if and only if

$$\nu_{im} \geq \xi\left(\mathbf{X}_i, \mathbf{Z}_m\right), \tag{2}$$

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 $\Phi_m\left(\cdot\right)$. 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.

D Additional Evidence on Rank Effects

D.1 Results for the Full Sample of Duflo et al. (2011)

In Appendix Table A.1 we replicate and extend the analysis in Section 3.1 by controlling for additional moments the peer skill distribution. In Table A.2 we consider all students in the experiment of Duflo et al. (2011), i.e., students in tracking as well as non-tracking schools. The coefficients in this table show that the estimated impact of ordinal rank on test scores is qualitatively robust to using the full sample of students, though the point estimates do decline somewhat. Including children in tracking schools has the benefit of increasing the number of observations. It also introduces 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 construction, 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 children near the assignment threshold 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.[7]

D.2 Evidence from the National Educational Longitudinal Study

Our third observational 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. Appendix Table A.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.[8] 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. Since classmates change across subjects, one would expect students' choices to vary accordingly if their relative standing in narrowly defined social settings matters for outcomes.

[7] 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.

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

This prediction is supported by some of the social psychology literature on "big-fish-small-pond" effects, which suggests that "self-concept" and, therefore, behavior are context specific (see., e.g., Marsh et al. (1984), Strang et al. (1978), Marsh et al. (2008), Parker et al. (2013)). Rogers (1978), for instance, finds that *within*-classroom rankings are more predictive of students' academic self-concept than other, plausible alternatives.

By estimating a model that relates differences in a student's behavior *across classrooms* to differences in her rank, we can also 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:

$$\Delta y_i = f(\Delta r_i) + \mathbf{X}_i'\beta + Grade_i + \epsilon_i, \tag{3}$$

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.[9]

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. Finally, we further include the variance of peers' subject-specific test scores, and a second order polynomial on peers' subject-specific mean test scores. These covariates attempt to control for moments of the peer ability distribution that might also be correlated with ranks.

Appendix Figure A.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.[10]

[9] 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.

[10] 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 (3) 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.

D.3 Can Teacher Behavior Explain the Relationship between Rank and Outcomes?

As mentioned in Section 3.3, 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. Appendix Table A.7 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

$$\Delta y_i = \varphi \Delta r_i + \mathbf{X}_i' \beta + Grade_i + \epsilon_i, \tag{4}$$

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."[11] 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 Appendix Table A.8 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 Appendix Table A.8 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 (4) 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

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

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.

E Data Appendix

E.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.

E.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.

E.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.

E.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.

E.5 Experimental Data from Houston Public Schools

Willingness to Practice Appendix F 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 F, 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 F, 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 F).

F Experimental Setup

F.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.

F.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.[12]

F.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

[12] 579 students showed up for the experiment and where thus paid a participation fee, as reported below. Due to computer issues only 573 students completed the experiment. Since we do not have data for the six students whose computers froze during the assessment stage, we cannot include them in our analyses.

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.

F.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.

F.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**.

Stage 1: First task

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.

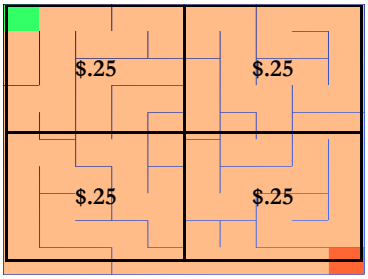
After the task is completed, you will be asked again about your ability at solving mazes and how sure you are of this ability. The second stage of the experiment will begin only after everyone else has finished this task. If you have completed this task quickly and are waiting on others, you will be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

Stage 2: Practice

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.

References

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.

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.

Boozer, M., and S. E. Cacciola (2001). "Inside the 'Black Box' of Project Star: Estimation of Peer Effects Using Experimental Data. "Yale Economic Growth Center Discussion Paper No. 832.

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

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.

Carrell, S. E., 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.

Cooley-Fruehwirt, J. C. (2010). "Desegregation and the Achievement Gap: Do Diverse Peers Help?"Unpublished Manuscript. Department of Economics. University of Wisconsin at Madison.

Crane, J. "The Epidemic Theory of Ghettos and Neighborhood Effects on Dropping Out and Teenage Childbearing. " *The American Journal of Sociology*, 96(5): 1226–1259.

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.

Ding, W., and S. F. Lehrer (2007). "Do Peers Affect Student Achievement in China's Secondary Schools? "*Review of Economics and Statistics*, 89(2): 300–312.

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.

Evans, W. N., W. E. Oates, and R. M. Schwab. (1992). "Measuring Peer Group Effects: A Study of Teenage Behavior. "*Journal of Political Economy*, 100(5): 966–991

Figlio, D. N. (2007). "Boys Named Sue: Disruptive Children and their Peers. "*Education Finance and Policy*, 2(4): 376-394.

Frisch, R., and F. V. Waugh (1933). "Partial Time Regressions as Compared with Individual Trends." *Econometrica*, 1(4): 387–401.

Gaviria, A., and S. Raphael (2001). "School-Based Peer Effects and Juvenile Behavior. " *The Review of Economics and Statistics*, 83(2): 257–268.

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.

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.

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

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

Hoxby, C. M., 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.

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.

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

Lavy, V., O. Silva, and F. Weinhardt (2009). "The Good, the Bad and the Average: Evidence on the Scale and Nature of Ability Peer Effects in Schools. "NBER Working Paper Series No. 15600.

Lefgren, L. (2004). "Educational Peer Effects and the Chicago Public Schools " *Journal of Urban Economics,* 56: 169-191.

Lyle, D. S. (2007). "Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point." *Review of Economics and Statistics*, 89(2): 531–542.

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

Marsh, H., and J. W. Parker (1984). "Determinants of Student Self-Concept: Is It Better to Be a Relatively Large Fish in a Small Pond Even if You Don't Learn to Swim as Well? " *Journal of Personality and Social Psychology*, 47(1): 213–231.

Marsh, H., M. Seaton, U. Trautwein, O. Ludtke, K. T. Hau, A. J. O'Mara (1987). "The Big-Fish-Little-Pond-Effect Stands Up to Critical Scrutiny: Implications for theory, methodology, and future research. " *Educational Psychology Review*, 20(3): 319–350.

Parker, P. D., H. W. Marsh, O. Ludtke, and U. Trautwein (2013). "Differential School Contextual Effects for Math and English: Integrating the Big-Fish-Little-Pond Effect and the Internal/External Frame of Reference. " *Learning and Instruction*, 23: 78–89.

Rogers, C.M., M. D. Smith, and J. M. Coleman (1978). "Social Comparison in the Classroom: the Relationship Between Academic Achievement and Self-Concept. " *Journal of Educational Psychology*, 70: 50–57.

Stinebrickner, R., and T. Stinebrickner (2006). "What Can be Learned About Peer Effects Using College Roommates? Evidence from New Survey Data and Students from Disadvantaged Backgrounds. " *Journal of Public Economics*, 90(8-9): 1435–1454.

Strang, L., M. D. Smith, and C. M. Rogers (1978). "Social Comparison, Multiple Reference Groups and the Self-Concept of Academically Handicapped Children Before and After Mainstreaming. " *Journal of Educational Psychology*, 70: 479–487

Zimmer, R. W.. and E. F. Toma (2000). "Peer Effects in Private and Public Schools Across Countries. " *Journal of Policy Analysis and Management*, 19(1): 75–92.

Zimmerman, D. J.. (2003). "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment " *Review of Economics and Statistics*, 85(1): 9–23.

G Appendix Figures and Tables

Figure A.1: Reconciling No Peer Effects with the Comparative Advantage Approach

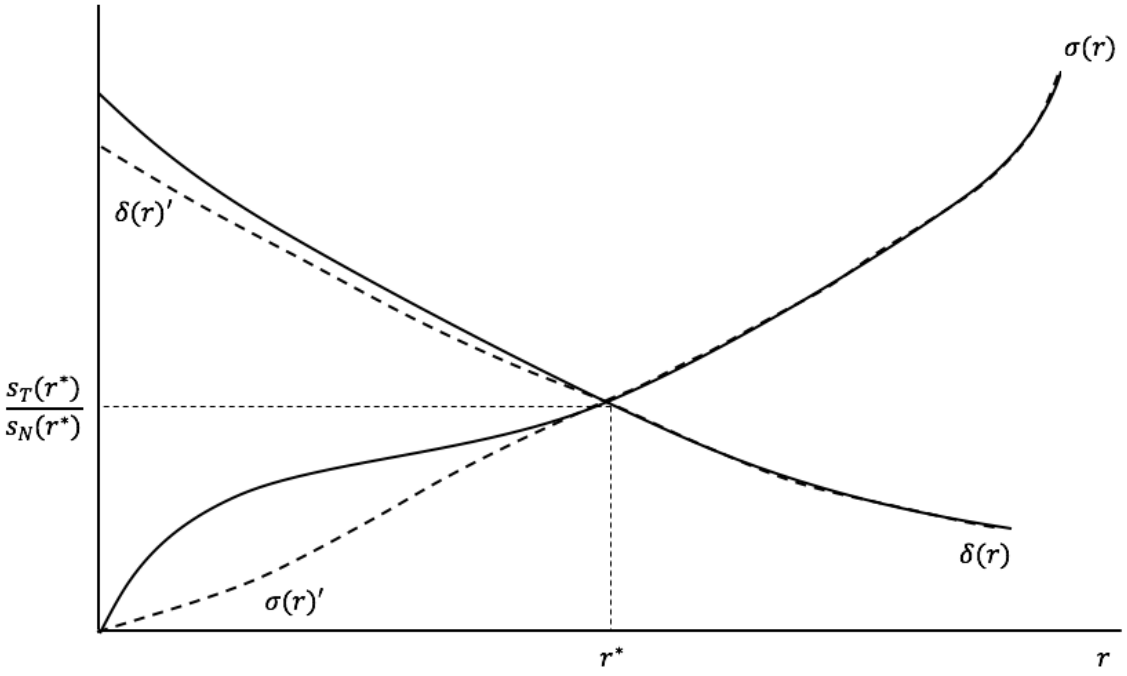


Figure A.2: Reconciling Linear in Means with the Comparative Advantage Approach

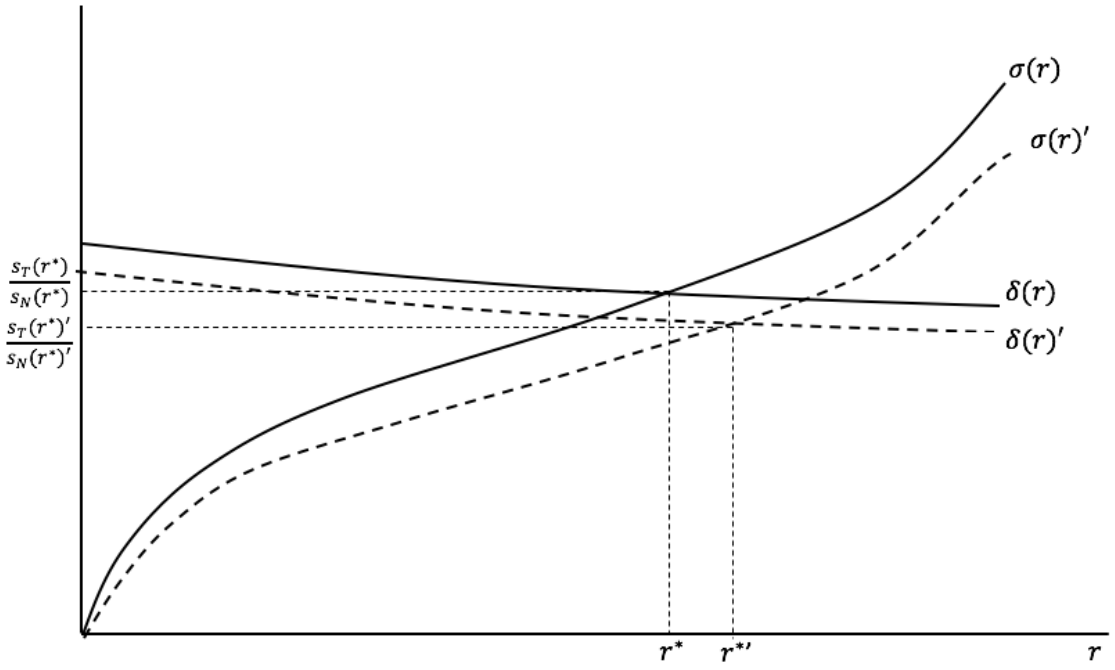


Figure A.3: Reconciling Heterogenous Positive Effects with the Comparative Advantage Approach

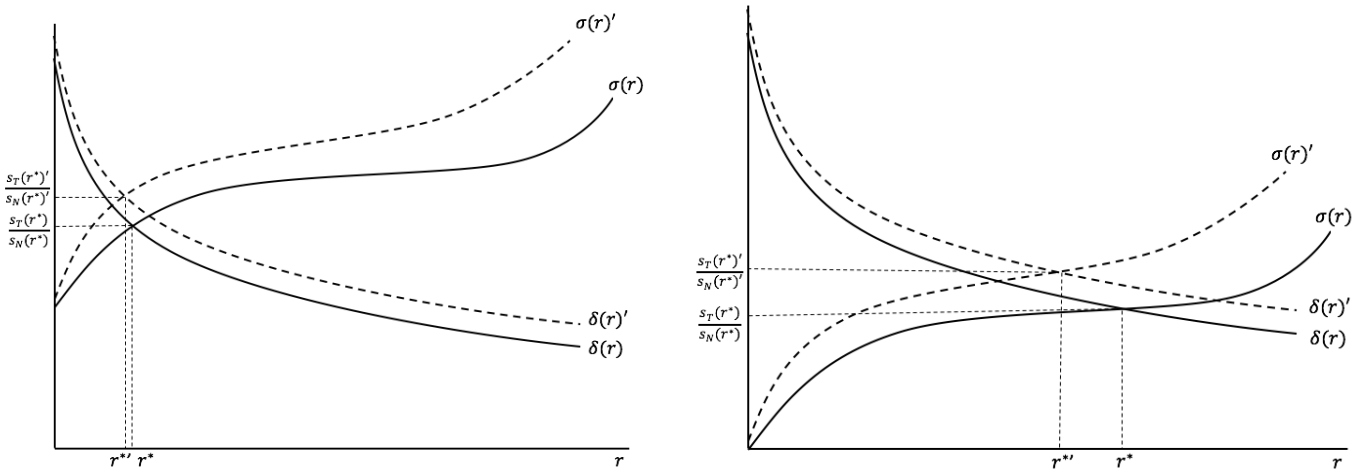
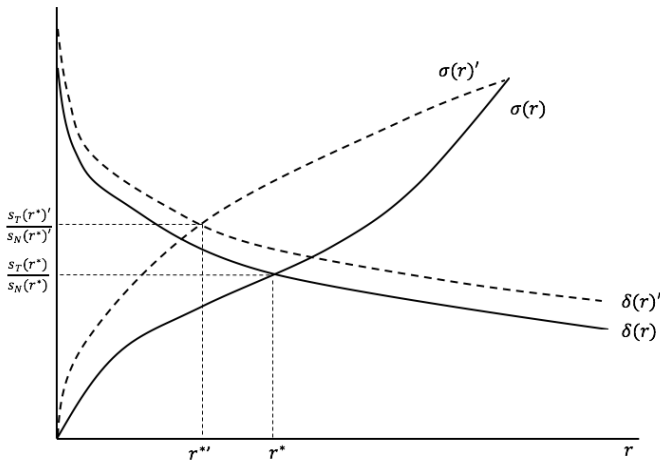


Figure A.4: Reconciling Heterogenous Negative Effects in the Air Force Example with the Comparative Advantage Approach

A) Random Assigment



B) Experimental Assignment

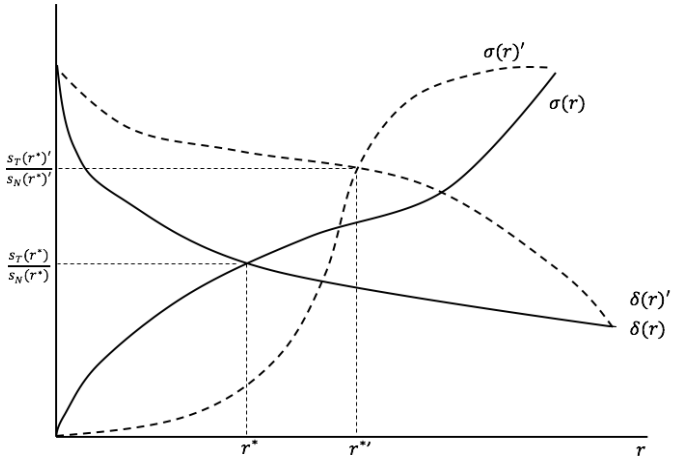
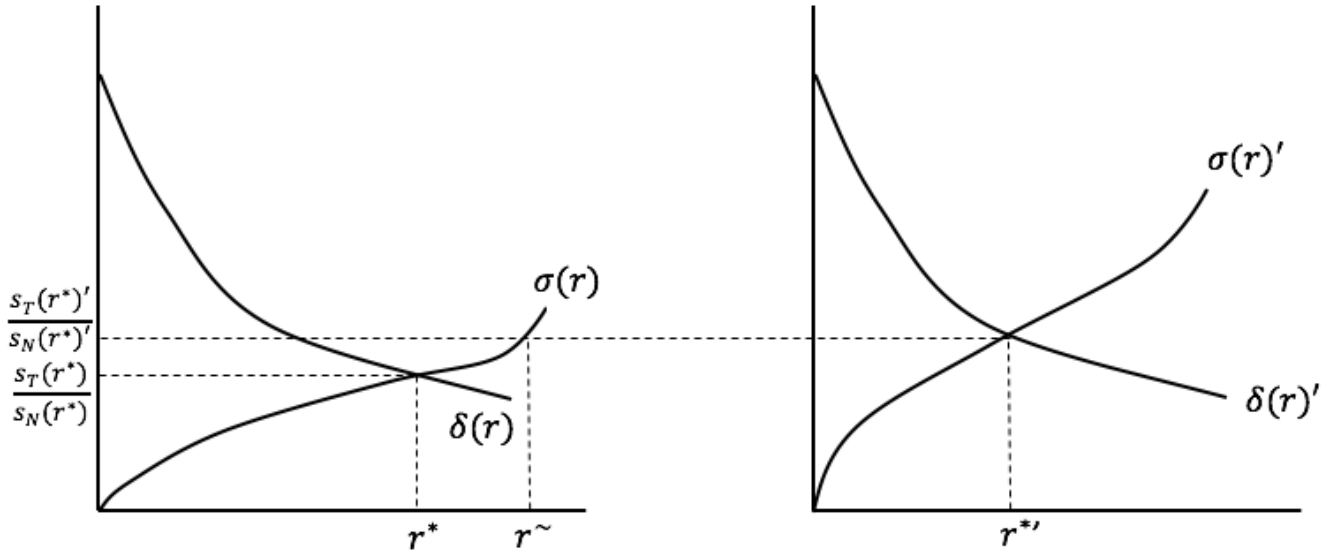


Figure A.5: Reconciling Heterogenous Negative Effects in the MTO Example with the Comparative Advantage Approach

a) Boys, Old Neighborhood

b) Boys, New Neighborhood



c) Girls, Old Neighborhood

c) Girls, New Neighborhood

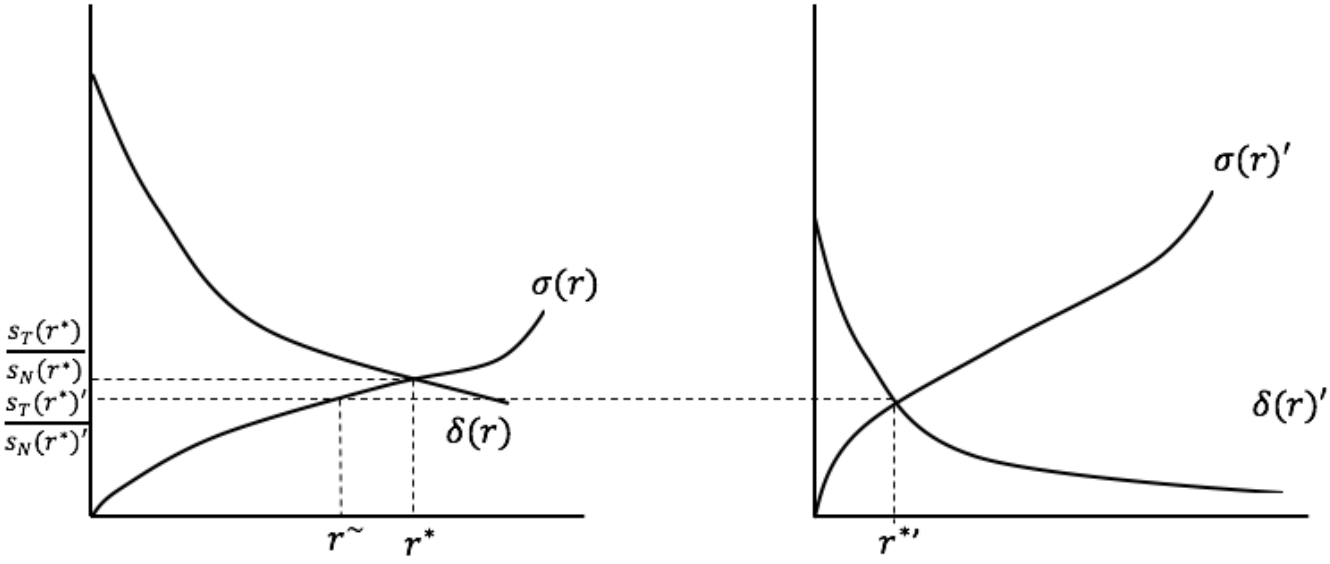
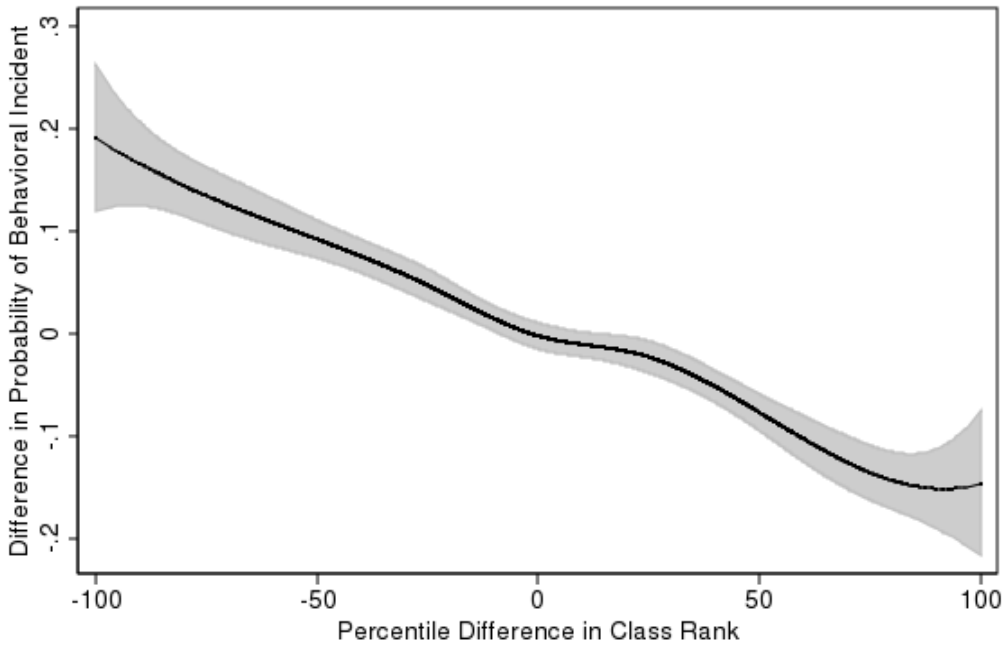


Figure A.6: Evidence from the National Educational Longitudinal Study



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). Estimates are obtained using cubic b-splines with three nodes that divide the sample equally. Appendix D.2 and the Data Appendix E provide additional information on the exact econometric specification as well as the sample.

Table A.1
Estimates of the Impact of Rank on Test Scores in the Control Group of Duflo et al. (2011)

	Endline Test Score				
	(1)	(2)	(3)	(4)	(5)
Percentile (÷100)	.418	.613**	.630**	.459**	.431*
	(.269)	(.273)	(.269)	(.229)	(.235)
Test Score at Baseline	.374***	.319***	.322***	.371***	.378***
	(.084)	(.085)	(.084)	(.073)	(.074)
Squared Test Score at Baseline	.015	.020	.021	.022	.022
	(.016)	(.016)	(.016)	(.015)	(.015)
Contract Teacher	.142***	.179***	.173***	.173***	.174***
	(.051)	(.019)	(.019)	(.019)	(.019)
Peers' Mean Test Score		1.183***	1.228***	1.116***	1.456***
		(.377)	(.389)	(.342)	(.422)
Squared Peers' Mean Test Score		2.072	1.118	1.382	1.396
		(3.783)	(4.065)	(3.161)	(3.273)
Cubed Peers' Mean Test Score		-18.609	-19.658	-19.889	-24.118
		(10.124)	(.038)	(7.071)	(7.158)
Variance of Peers Test Score		-.001	-.008	-.014	.076
		(0.108)	(.109)	(.109)	(.155)
10th Percentile of Peer Skill Distribution					-.106
					(.212)
25th Percentile of Peer Skill Distribution					.119
					(.248)
75th Percentile of Peer Skill Distribution					-.296
					(.193)
90th Percentile of Peer Skill Distribution					.127***
					(.138)
Constant	-.309*	-.429**	-.229	-.064	-.159
	(0.168)	(.216)	(.298)	(.286)	(.387)
Additional Controls	No	No	Yes	Yes	Yes
School Fixed Effects	No	No	No	Yes	Yes
R-Squared	.240	.247	.257	.392	.396
Number of Observations	2190	2190	2188	2188	2188

Notes: Entries are OLS coefficients and standard errors fromestimating equation (6). Heteroskedasticity robust standard errors are clustered on the school level and reported in parentheses. The sample consists of all students who attend nontracking schools and have nonmissing baseline test scores. 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. ***, **, and * denote statistical significance at the 1%-, 5%-, and 10%-levels, respectively.

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

	Endline Test Score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
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 sample consists of 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. ***, **, and * denote statistical significance at the 1%-, 5%-, and 10%-levels, respectively.

Table A.3: Summary Statistics for NYCPS Data by School Zoning Attendance

	Not Comply	Comply	p-value (1) - (2)
Behavioral Indicators:			
Behavioral Incident in 5th Grade	.086	.075	.065
Rank Indicators:			
Predicted Change in Rank (Percentiles)	-3.441	-3.258	.299
Test Scores:			
5th Grade ELA Test Score	662	661	.672
5th Grade Test Score	668	669	.593
Demographics:			
White	.110	.197	.003
Black	.379	.234	.000
Hispanic	.407	.379	.312
Asian	.100	.187	.000
Other race	.004	.004	.172
Male	.496	.519	.000
Female	.504	.481	.000
Free lunch	.844	.810	.156
English Language Learner	.082	.105	.009
Special education	.098	.072	.000
Observations	134,266	115,752	250,018

Notes: Entries are means and p-values for difference in means for predetermined student characteristics in the NYCPS data, by compliance with school zoning regulations. p-values for difference in means account for clustering at the school level. For further details and precise definitions of all variables tsee he Data Appendix.

Appendix Table A.4: Exploring the Correlation between
5th Grade Behavior and Predicted Rank Changes

	Compliance with Middle School Zoning		
	Not Comply	Comply	Full Sample
Predicted Change in Rank	-.00000	-.00011	-.00005
	(.00008)	(.00009)	(.00006)
R-Squared	.0941	.0984	.0914
Number of Observations	257,094	221,297	478,391

Notes: Entries are OLS coefficients and standard errors from regressing a an indicator variable for behavior incidents in 5th grade on the instrument described in section 3.2. The set of controls consists of the covariates in (8), including school fixed effects. Heteroskedasticity robust standard errors are clustered at school level and reported in parentheses. See the Data Appendix for the precise definition and source of each variable.

Table A.5: 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. ***, **, and * denote statistical significance at the 1%-, 5%-, and 10%-levels, respectively.

Table A.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 A.7: 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.8: 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. ***, **, and * denote statistical significance at the 1%-, 5%-, and 10%-levels, respectively.