

Learning to Manage and Managing to Learn: The Effects of Student Leadership Service*

Michael L. Anderson
UC Berkeley and NBER
mlanderson@berkeley.edu

Fangwen Lu
Renmin University of China
lufangwen@ruc.edu.cn

February 11, 2016

Abstract

Employers and colleges value individuals with leadership service, but there is limited evidence on whether leadership service itself creates skills. Identification in this context has proved difficult because settings in which leadership service accrues to individuals for ostensibly random reasons are rare. In this study we estimate the effects of random assignment to classroom leadership positions in a Chinese secondary school. We find that leadership service increases test scores, increases students' political popularity in the classroom, makes students more likely to take initiative, and shapes students' beliefs about the determinants of success. The results suggest that leadership service may impact human capital and is not solely a signal of pre-existing skills.

JEL Codes: I21, J24

Keywords: Extracurriculars, classroom environment, student government

* Corresponding author: Fangwen Lu. We thank Stefano DellaVigna and Ted Miguel for insightful comments. Any mistakes are the authors' own.

1. Introduction

Employers and colleges place a high premium on individuals with leadership service, but little is known about what types of skills, if any, leadership service creates. Individuals with leadership or managerial experience in diverse fields – business, politics, and education – are observably different from other individuals. These differences may arise because leadership service generates human capital or because these individuals are selected for their pre-existing skills (or both). It has proven difficult to credibly measure the effects of leadership service because settings in which leadership service accrues to individuals for ostensibly random reasons are extremely rare.

In this study we estimate the effects of random assignment to classroom leadership positions in a Chinese secondary school. In most Chinese schools, teachers select students for classroom leadership positions. These positions carry considerable prestige, and some parents lobby to have their children selected for the positions. In our study, teachers nominated short lists of candidates for each classroom leadership position, and then one candidate for each position was chosen at random from the nominated list. Exploiting this randomization in leadership assignments, we evaluate the effects of leadership service on academic performance, confidence and aspirations, political popularity, and beliefs about the determinants of success.

We find that leadership service increases test scores among students most likely to serve in these positions. Some of this increase may be due to incentive effects, and much of the experience that the students gain is managerial in nature. Nevertheless, leadership service also dramatically increases political popularity in the classroom for these same students. Among students who would typically be runners-up for leadership positions, leadership service makes them more likely to take initiative and affects their beliefs about the determinants of success. Overall the results suggest that service in classroom leadership positions can impact human capital and is not solely a signal of pre-existing skills.

2. Literature Review

A large body of literature in economics focuses on whether and how leaders can affect group outcomes and on what types of traits leaders exhibit. For example, Bertrand and Schoar (2003), Chattopadhyay and Duflo (2004), and Jones and Olken (2005) present

evidence that leaders appear to affect group performance in corporations, local villages, and countries respectively. Beaman et al. (2009; 2012) demonstrate that female village leadership reduces prejudice and increases aspirations among girls. Hermalin (1998) argues that a critical trait for leaders is the ability to send credible signals to followers – they must “lead by example.” Andreoni (2006) and Güth et al. (2007) explore the effects of leadership and leading by example in public good contributions. Camerer and Lovallo (1999) and Malmendier and Tate (2005) show that leaders tend to be overconfident and can create market distortions. In contrast to this rich literature, however, there is little work that explores the effects of leadership on the leaders themselves.

One exception is research by Kuhn and Weinberger (2005) that documents a high return to high school leadership service in the United States. Individuals with high school leadership service earn from 4 to 33 percent more than individuals without high school leadership service. This return is comparable to the return on an additional 0.5 to 4 years of education (i.e., at the high end it affects wages as much as getting a college education). It is difficult to distinguish whether firms pay more for individuals with leadership service because the service itself is valuable or whether they pay more because the service signals that the worker has special skills that existed prior to the leadership service. Nevertheless, there is suggestive evidence that leadership service may have a causal effect; students attending schools with more leadership opportunities earn more than those attending schools with fewer leadership opportunities.

Outside of economics, a substantial literature in psychology studies the effects of power upon individuals. For example, individuals primed to feel powerful are more likely to take initiative (Galinsky et al. 2003), more likely to engage in risky behavior (Anderson and Galinsky 2006), and more self-oriented (Galinsky et al. 2006). This literature is related to our research in that a leadership position implies some degree of power, but our interest is in the effects of the entire leadership experience, not simply the power it entails. In summary, to the best of our knowledge our research represents the first experimental evidence of the effects of leadership service on leaders themselves.

3. Classroom Setting and Experimental Design

In most Chinese secondary schools, students sit in the same classroom with the same classmates over the entire semester, and teachers of different subjects rotate between classrooms throughout the day. Students in the same classroom not only take classes together but also conduct various extracurricular activities together, including cleaning classrooms, attending sporting events, taking part in festivals, and going on field trips.

Typically a class has a student leadership structure as shown in Figure 1. There are seven major class leadership positions: class monitor, vice class monitor, labor commissary, entertainment commissary, Chinese delegate, English delegate, and math delegate. The class monitor is at the top of the management structure, and she assumes a wide range of responsibilities – representing the class, organizing collective activities, and maintaining order in the class. At the beginning of each lecture, the class monitor calls all students to “stand up” to greet teachers. The vice class monitor assists the class monitor, especially in maintaining order in the class, i.e., keeping students from talking aloud or moving around during class time. The labor commissary assigns students to various cleaning tasks and monitors the performance of these tasks.¹ The entertainment commissary organizes singing and dance performances for school events or festivals and updates the bulletin board periodically. Course delegates are assigned for each major subject: Chinese, English, and math. They urge students to turn in homework, collect and distribute homework for the relevant subjects, and sometimes assist in grading homework. At the beginning of every morning there is a 30-minute reading-aloud section, and Chinese and English delegates rotate to lead this section. In addition to these seven major class leader positions, there may be other minor class leader positions: course delegates for less important subjects, other commissaries, and leaders of sub-classroom teams. What is common to all the major class leaders is that they have more exposure to teachers and other students and, to varying degrees, they must motivate other students in order to fulfill their responsibilities.

It is notable that the leadership positions entail a mixture of leadership opportunities and managerial responsibilities. While there is no universal definition of what qualifies as leadership, most definitions imply that leadership relates to social influence rather than direct application of authority. In this context the monitor positions may involve some exercise of

¹ In most primary and secondary schools, students are responsible for cleaning the floor, windows, and walls of their classroom. Public areas, like stairs and playgrounds, are also allocated to each class as responsibility areas. A labor commissary can ask students to redo some cleaning work if she finds it inadequate.

leadership skills; at a minimum, organizing activities and maintaining order will be much easier for students who can exercise social influence, as a student leader's authority is much more limited than, for example, a teacher's. The commissary and delegate positions, in contrast, are more managerial in nature. The organizational and planning skills necessary for these positions are important but different from leadership skills. In theory these differences represent an effect to test for effect heterogeneity by position type, but in practice sample size limits our statistical power.²

In general there is excess demand for class leader positions. In a survey among entering Grade 7 students in a school close to our study's school, students were asked to report whether they agreed with the statement, "If there is an opportunity, I want to become a class leader". Thirteen percent of students "strongly" or "somewhat" disagreed, 23% of students indicated a neutral opinion, 30% "somewhat" agreed, and 33% "strongly" agreed. Thus, though over 60% of students express interest in holding a class leadership position, less than 15% of them can hold such a position in any given year.

Teachers typically appoint class leaders, and there are no explicit rules for selection.^{3,4} Teachers usually have three goals when selecting class leaders: (1) to fulfill the responsibilities of the positions; (2) to provide role models for other students; and (3) to reward good students with leadership positions. Thus they choose students with a combination of both managerial abilities and academic performance, with more emphasis on managerial abilities for commissaries and more emphasis on academic performance for course delegates. Given that there is high demand for class leadership positions, parents sometimes lobby teachers to appoint their children to these positions.

Teachers appoint class leaders at the beginning of the semester. During the semester, as more information becomes available, teachers may adjust the appointments. For example, teachers may strip a student's leadership title if the student violates school rules by fighting with others or if they see a class leader is not performing his responsibilities. In addition

² When estimating effects separately by position, we find suggestive evidence that class monitor positions – particularly the head class monitor position – have larger effects on test scores than other positions. However, these results only achieve marginal statistical significance at best.

³ Each classroom has an administrative teacher. The administrative teacher is a regular teacher with additional managerial responsibilities, which include arranging class events, disciplining misbehavior, communicating with parents, and maintaining classroom order. The administrative teacher typically appoints the leaders.

⁴ In some schools and classrooms, class leaders are elected democratically by all students or rotate among all students. Most frequently in primary and secondary schools, however, teachers select class leaders.

teachers may also reappoint class leaders if a leader does very poorly in the midterm exam or if a non-leader does exceptionally well.

The endogenous appointment of class leaders under normal circumstances prevents a credible analysis of the effects of leadership service. Our experiment modifies how class leaders are appointed. In this experiment, teachers were asked to nominate two candidates for each leadership position using to their normal criteria. They ranked these two candidates as Candidate 1 and Candidate 2, with Candidate 1 being their first choice for the position. We told teachers before the selection process that one candidate would be randomly assigned as the class leader and that either candidate had the same chance of being appointed. We randomly assigned one out of each pair of candidates for the position.

The randomization procedure was performed without replacement as follows. Among the seven classes in Grade 7, Classes 1, 4, 5, and 7 were assigned to Group 1, and Classes 2, 3, and 6 were assigned to Group 2. For Group 1 classes, we chose Candidate 1 to be actual leaders for class monitor, entertainment commissary, and Chinese delegate, and Candidate 2 for vice class monitor, labor commissary, English delegate, and math delegate. For Group 2 classes, we did the reverse. The goal of this procedure was to ensure that: (1) Candidate 1 and Candidate 2 students were perfectly balanced across “appointed” and “non-appointed” groups; (2) each class had as close to 50% Candidate 1 students appointed leaders as possible; (3) in each class both Candidate 1 and Candidate 2 students were represented in class/vice class monitor positions, commissary positions, and course delegate positions. While all of these results would occur in expectation with a simple random assignment procedure, our strategy of sampling without replacement ensures that balance is achieved on all three of these dimensions. This is notable given the small size of our sample.

We asked teachers to avoid adjustments in leadership positions unless the appointed leaders performed very poorly. In that case we advised teachers to avoid appointing students who were not leadership candidates, either by letting existing class leaders take more than one position or by appointing other students who were not initial candidates as new leaders. Since teachers tend to adjust class leaders after the midterm exams, we called the school principal around the midterm and asked him to emphasize the importance of not adjusting class leaders and reiterate the strategies for avoiding selection of non-leader candidates. We

estimate an overall compliance rate of approximately 85% based on surveys of 21 students (3 from each class) shortly before the final.⁵

To avoid the possibility of gaming, teachers were not made aware of the assignment algorithm.⁶ We told teachers that we were interested in the effect of leadership positions on the “development” of students. They may have reasonably expected test scores as one of the outcomes. Beyond this, they should not have anticipated the survey outcomes we used in the study, which were unfamiliar to them. Students were unaware of the research project and the candidate list, and there were no financial incentives for teachers or students.

Given these details, some sections of our study represent a “natural field experiment” while other sections represent a “framed field experiment” (Harrison and List 2004). The sections of our study that estimate the effects of leadership service on test scores qualify as a natural field experiment. The subjects (students) work in a natural setting, they are unaware that their leadership positions have been randomly assigned, and the leadership activities and exams proceed exactly as they would have absent the study. The sections of our experiment that estimate the effects of leadership service on confidence, beliefs, and social networks qualify as a framed field experiment. Subjects remain unaware that their leadership positions are randomly assigned, and the activities and responsibilities of these positions are identical to what they would have been absent the study. Thus there is field context in the “task” (leadership service) that the subjects perform. These sections are not a pure natural field experiment, however, because the outcome measure – survey responses – is not a measure that would naturally occur if the experiment had not been run. During the survey the subjects can likely infer that we wish to gather some type of information from them, even though they do not know the specific purpose of the study.

4. Data and Empirical Framework

⁵ To conduct this survey the school principal directly called the students – chosen by us – from the classroom and took them to a specific room to take the short survey. Teachers had no knowledge about the survey and thus no opportunity to coach students on how to answer.

⁶ Indeed, although we told them all candidates had equal chance of being selected, teachers urged us to select Candidate 1 if possible.

The experiment was conducted in a “rural” middle school in a coastal city in Jiangsu, China during the fall semester of 2009.⁷ The types of leadership positions in this school are typical of those found in most Chinese schools. The characteristics of children in our school, however, may not be representative of the average Chinese child. Table 1 presents summary statistics from the 2000 Chinese Census comparing households in our study’s area to the average Chinese household. The school in our study is located in a rural area, so the first two columns of Table 1 compare households living in all Chinese rural areas to households living in our study’s rural area. Households in our study’s rural area are more educated and slightly larger than households in the average Chinese rural area. They are also more likely to have running water or toilets. However, most of these differences are modest in magnitude even when statistically significant (e.g., less than 0.25 standard deviations). The last two columns of Table 1 compare all Chinese areas to our study’s overall area. The differences between the last two columns are even smaller than the differences between the first two columns. For all measures except toilet availability and household size, the urban-rural gap is much larger than the gap between our study area and all Chinese areas. This suggests that the main issue for generalizing our results to other areas in China may be the urban-rural divide rather than the specific area in which we conducted our study. We discuss issues of generalizability in more detail in Section 6.

Our study’s middle school had seven classes in Grade 7, the school’s entry grade. Each class had between 52 to 56 students. At the end of the second week, teachers selected 14 candidates (2 candidates for each of the seven major class leader positions) out of a total of 52 to 56 students. As two candidates were selected for more than one position by mistake, and one candidate transferred to another school, three candidate pairs are dropped from the analysis. In total, 46 pairs of candidates (92 total) remain in the study, with Candidates 1 and Candidates 2 each having a 50% chance of being selected.

Background characteristics for students include baseline test scores, gender, age, height, birth order, father’s education, mother’s education, and relative income of the family. The baseline test score comes from an exam administered during the first week of the semester, and it is normalized to have mean 0 and standard deviation 1 across all students. Panel A of Table 2 reports summary statistics for leadership candidates and all other

⁷ The school is in a rural administrative unit, but it is only a 20-minute drive from the nearest city center. Using United States’ classifications it would probably classify as “suburban” rather than “rural.”

students in the grade. Candidates have baseline scores that are 0.5 standard deviations higher than non-candidates, which confirms that academic performance is one of the factors teachers use to select candidates. Due to a preference for sons in the area, 59% of all students are boys. However, only 45% of the candidate group is male, as teachers appear more likely to select girls as candidates for leadership positions. Candidates are on average 13.3 years old and about 159 centimeters tall (5'3") at the beginning of the semester. Relative to non-candidates, candidates are more likely to be first born. Fathers' education is around 8.3 years, and mothers' education is around 6.9 years.⁸ As students were only around 13 years old and were unlikely to know their household income, the survey asked students to compare their families with their classmates' families, and to report on a 1-5 scale with 1 indicating "far below the average," 3 "around the average," and 5 "far above the average".⁹ The mean of the comparison scores among candidates is 2.8, with 72% reporting around the average, 22% below the average, and 6% above the average.

We collected outcome data for four different domains – academic performance, confidence, perceived determinants of success, and social networking or popularity. Panel B of Table 2 reports mean outcomes for leadership candidates and all other students in the class. Academic performance is measured with test scores from the midterm and the final. Students were tested in three major subjects – Chinese, English, and math – and two minor subjects – history and politics. Grading was rigorously conducted. Teachers in the same subjects allocated exam questions among themselves so that the same question was always graded by the same teacher. In addition, students' names were hidden during the grading process, diminishing the likelihood that teachers might grade leaders on a more favorable scale.¹⁰ The three major subjects account for 150 points each in raw scores, history accounts

⁸ We do not have age information for parents, so it is a bit difficult to compare our sample with the national average because education has been trending upwards over time. If we assume that parents are 24 years older than students, then most parents were born around May 1972. The average education level in rural areas is 8.2 years for males born between May 1970 and May 1974, which is similar to the average fathers' education in our sample. The corresponding figure for females is 7.5 years, which is higher than that in our sample. (Data come from the 0.1% sample of the 2000 Chinese population census.)

⁹ The data on relative income was collected after the experiment, but we do not expect that leadership service will affect perceptions of household income. Indeed, this characteristic is balanced across treated and control students.

¹⁰ During midterms and finals students from the seven classrooms were mixed together and randomly seated under the school's effort to prevent cheating. Before grading, answer sheets from the same exam room were stapled together with students' names sealed. Subject teachers cycle through classrooms and were exposed to dozens of leaders and hundreds of students. It would thus be very difficult to identify a particular student's

for 60 points, and politics accounts for 40 points. We standardize the test scores for our analysis. On average, leadership candidates score approximately 0.7 standard deviations above the mean of non-candidates on the midterm and final.

We are interested in the effects of leadership on confidence and perceived determinants of success because the psychology literature finds that individuals that are primed to feel powerful display greater initiative and are more “self oriented.” We measure confidence in several ways. One way is to ask students to evaluate their own abilities relative to other students; more confident individuals are more likely to evaluate their own abilities as higher than others’ even if there are no disparities in actual abilities. Several days before the midterm students were asked to rank their own learning abilities (abilities to understand and apply concepts and methods) relative to their peers and were instructed to evaluate themselves on a percentile scale of 0 to 100. 100 means the student believes that she is the best in the entire grade, 0 means she believes that she is the worst, and 50 means she believes that she is average.¹¹ The average reported percentile is 77 among candidates and 69 among non-candidates, strongly suggesting the presence of a “Lake Wobegon effect.” We also created a measure of “overconfidence” by taking the difference between a student’s perceived percentile ranking in the grade and her actual percentile ranking on the midterm, and we asked students how many years of education they aspire to. Finally, we asked students whether they were willing to be the first contributor in a sequential public goods game and how much they would contribute.¹² Going first signals a degree of confidence or leadership in that subsequent contributors observe the contributions of the first contributor. 76% of candidates and 66% of non-candidates were willing to go first.

handwriting from more than 350 answer sheets. In addition, as teachers of the same subject divided the grading, a teacher was only in charge of one-half or one-third of the questions in a specific subject.

¹¹ One candidate and several non-candidate students reported scores between 100 and 150, suggesting that they may have used the scale for test scores in making their evaluations. In response we top coded the self-evaluations at 100.

¹² The sequential public goods game was presented as follows: Individuals in a group voluntarily contribute money to form a collective fund, and the fund will be invested such that its value doubles. An equal share of the revenue will be distributed to each individual no matter how much he contributes. Each group has 10 people, and the other nine are your classmates. The amount of contribution is between 0 and 10. Before the investment, everyone has 10 yuan initially. For example, in one group 4 students contribute 2 yuan, 3 students contribute 5 yuan, 2 students contribute 6 yuan, and 1 student contributes 0 yuan. The collective fund is then 35 yuan ($4*2 + 3*5 + 2*6 + 1*0 = 35$). The collective revenue will be 70 after investment ($35*2=70$), and member receives 7 yuan ($70/10 = 7$). The students who contribute 6 yuan ultimately have 11 yuan ($10 - 6 + 7 = 11$), while the one who contributes nothing has 17 yuan ($10 - 0 + 7 = 17$).

We measure perceived determinants of success by asking students what the most important factor in academic success is. Possible choices include effort, good teachers, fostering by parents, innate talent, help from classmates and friends, and home study environment. Effort was the most frequent choice (51% of candidates and 36% of non-candidates), and good teachers was the next most common choice (24% for both groups). No other choice exceeded 15% frequency in either group.

We surveyed students on their social networks by asking each student to list their three closest friends and their total number of friends. We then constructed two measures of social networks: the number of students that student i reports, and the number of other students that list student i as a close friend. To gauge political popularity we asked students to list three names that they would vote for if class leaders were democratically elected during the following semester. Due to the significant amount of work needed to match names across hundreds of students, we only constructed the political popularity measure and the second social network measure for leadership candidates (i.e., students that actually appear in our regressions).

Figure 2 presents a timeline of the study's different phases. The total time lapsed between random assignment of leadership positions and the final survey is approximately five months (20 weeks). We designed the surveys containing the measures discussed above to collect data explicitly for the study. The midpoint survey, conducted shortly before the midterm, was framed as a short learning exercise for students. In this survey we asked each student to rank him or herself in the class and emphasized that accurate prediction is important for the growth of a student. Administrative teachers handed out the questionnaire during a break in a normal weekday, and students were instructed to fill it out independently. The post survey was entitled "Survey on middle school students' study and life" and it was conducted right after students finished the final exam. Students had two hours to complete it. All the students in the cohort, including non-candidate students, took both surveys. Thus the surveys will not cause leaders and non-leader candidates to feel special relative to other students.

There is little or no attrition for most outcomes. The exceptions are the questions relating to the sequential public goods game. Only 76% of candidates responded to the question about whether they would be willing to contribute first, and only 70% of candidates reported how much they would be willing to contribute. However, the rates of missing

responses to these questions do not significantly differ across leaders and non-leaders; the treatment group is 7.9 percentage points more likely to be missing responses for the first question ($t = 1.0$) and 5.9 percentage points less likely to be missing responses for the second question ($t = -1.0$). Thus we do not expect this attrition to cause any bias.

We test for random assignment of leadership positions by limiting the sample to class leader candidates and estimating regressions of the form:

$$X_i = \alpha_0 + \alpha_1 \text{Leader}_i + \varepsilon_i \quad (1)$$

In this regression, X_i is a predetermined characteristic such as baseline test score, gender, or father’s education. If the random assignment procedure was successful, then $\hat{\alpha}_1$ should be close to zero and statistically insignificant in these regressions. However, statistical inference in equation (1) (and all other regressions in our paper) is complicated by the fact that the standard errors are clustered at the classroom level and our sample contains only seven classrooms. This means that there are very few clusters (seven), potentially biasing the clustered standard errors. To address this problem we use our knowledge of the randomization procedure to perform exact permutation tests. These tests are derived solely from the actual randomization procedure and thus have the appropriate size regardless of the dependence structure of the data (Rosenbaum 2007).

Recall that our experiment contains seven classrooms and seven leadership positions per class. We chose four classrooms to designate as Group 1. In these four classrooms we chose three positions in which Candidate 1 would be the leader (class monitor, entertainment commissary, and Chinese delegate). In the remaining four positions Candidate 2 was assigned as leader. The remaining three classrooms we designated as Group 2. In Group 2 classrooms we reversed the candidate assignments (i.e., Candidate 2 was chosen as the leader in the class monitor, entertainment commissary, and Chinese delegate positions, and Candidate 1 was chosen as leader for the remaining four positions). Using this procedure of drawing four classrooms and three positions, there are 35 ways (seven choose four) to assign classrooms and 35 ways (seven choose three) to assign leadership positions.¹³ There are thus a total of 1,225 (35^2) ways in which the leadership positions could have been assigned. Under the sharp null hypothesis of no treatment effect we can map out the

¹³ One might also consider drawing three classrooms and four positions instead of four classrooms and three positions. However, since the labels “Group 1” and “Group 2” are arbitrary, it makes no difference in the permutations whether we draw three classrooms and four positions for “Group 1” or four classrooms and three positions for “Group 1.”

distribution of t -statistics for all 1,225 combinations and use this distribution for statistical inference. These tests remain valid even for small numbers of clusters since they are derived from the randomization procedure itself.

To implement the exact permutation tests we calculate all 1,225 possible ways of assigning leadership positions; we call these placebo assignments. For each placebo assignment we estimate equation (1) and record the t -statistic on α_1 . We then compare the t -statistic from a regression using the actual leadership assignment variable to the distribution of t -statistics from the 1,225 placebo leadership assignments. The p -value, reported in italics in all tables, represents the fraction of placebo assignment t -statistics that are larger than the actual t -statistic.

Table 3 reports results from estimating equation (1) on the sample of class leader candidates. Panel A presents estimates for students who were teachers' first choices for leadership positions (Candidates 1), and Panel B presents estimates for students who were teachers' second choices for leadership positions (Candidates 2). Baseline test scores, gender, height, age, birth order, parental education, and family income are all balanced across leaders and non-leaders, suggesting that the randomization procedure was successful. All of the coefficients in both panels are statistically insignificant, with p -values ranging from 0.170 to 0.973 (median p -value = 0.793). We simultaneously test whether all eight characteristics are correlated with class leader assignments by regressing the leadership assignment variable on all eight characteristics and testing whether the coefficients on all characteristics jointly equal zero. This test returns p -values of 0.778 for Candidates 1 and 0.689 for Candidates 2.¹⁴ Finally, for completeness we check that covariates are balanced across leaders and non-leaders when pooling both candidate types together in the regression sample (not shown in Table 3). The coefficients for the eight characteristics remain statistically insignificant when we pool both candidate types together, with p -values ranging from 0.355 to 0.982 (median p -value = 0.787).

¹⁴ We cannot perform a conventional F -test because the number of clusters (seven) is less than the number of parameters in the regression (eight coefficients and an intercept). Furthermore, the accuracy of this F -test would be questionable given the small number of clusters. Instead we use the exact permutation test to compare the actual R^2 in the regression of leader assignments on all eight characteristics to the distribution of R^2 statistics from identical regressions that use placebo leadership assignments as their dependent variables. A p -value of 0.778, for example, reveals that the actual R^2 is smaller than 78 percent of the placebo-generated R^2 statistics.

In our main analysis we test for effects on many outcomes across two candidate types. Multiple testing is thus a concern in our context; even if the treatment has no effects, we may find several significant results simply from running so many tests. To address this concern we report false discovery rate adjusted “ q -values” in all of our main results tables. The false discovery rate (FDR) represents the expected proportion of rejections that are type I errors. FDR formalizes the trade-off between correct and false rejections. If all hypotheses are true (i.e., the treatment has no effects), then controlling the false discovery rate at level q also controls the familywise error rate (FWER) at level q . When some false hypotheses are correctly rejected, however, FDR control affords additional power over FWER control (Anderson 2008). We define the family of tests under consideration as all outcomes across Tables 4 through 7 and use the “sharpened” FDR control procedure presented in Benjamini, Krieger, and Yekutieli (2006) to achieve FDR control. In each table we report FDR-adjusted p -values in square brackets. Controlling FDR at the level of $q = 0.05$ implies that we expect one false rejection for every 19 correct rejections. To be conservative we also report FWER-adjusted p -values computed using an algorithm described in List, Shaikh, and Xu (2015). This algorithm accounts for dependence between test statistics and thus can afford additional power in contexts such as ours. For the most significant outcomes the FWER-adjusted p -values are slightly smaller than the FDR-adjusted q -values (because the List et al. algorithm accounts for dependence), but for moderately significant outcomes the FDR adjusted q -values are generally smaller than the FWER-adjusted p -values (because FDR control is willing to tolerate an occasional false rejection in return for many correct rejections).

5. Results

We estimate the effects of leadership service – i.e., service in a classroom leadership position – by limiting the sample to class leader candidates and estimating regressions of the form:

$$Y_i = \beta_0 + \beta_1 \text{Leader}_i + X_i \gamma + \varepsilon_i \quad (2)$$

Since class leaders are randomly assigned within the population of candidates, estimates of β_1 will be unbiased even absent controlling for baseline covariates X_i . Nevertheless, including X_i in the regression increases precision by reducing the regression’s

mean squared error. We thus include baseline test scores, gender, height, and age in our regressions. Since compliance with leadership assignments may have been less than 100%, we interpret our regression estimates as “intent to treat” effects. For readability we often refer below to the “effects of leadership service,” but we implicitly mean the “effects of intending to treat a student with leadership service.” If we were to apply an instrumental variables estimator to account for noncompliance with leadership assignments, the IV estimates would be approximately 18% larger in magnitude than the estimates we report in Tables 4 through 7 (this figure is based on an estimated compliance rate of 85%). However, since our data on noncompliance rates were collected somewhat informally from students – we were not certain teachers would accurately report noncompliance – we feel it is safer to report “intent to treat” estimates.

In all tables we separately estimate effects for teachers’ first choices for each position (Candidates 1) and second choices for each position (Candidates 2). The effects on Candidates 1 represent the effects of the “treatment on the treated” that we would expect to observe absent the randomization procedure, since under normal circumstances Candidate 1 would be appointed as leader. The effects on Candidates 2 represent the effects that we might expect to observe if we increased the supply of leadership positions, since Candidates 2 would be the natural candidates to fill newly created positions. This interpretation of course assumes there is no dilution effect from increasing the number of leaders.

Table 4 presents estimates of the effects of leadership service on test scores and the availability of study time. In Tables 4 and all subsequent tables, the estimates in Panel A control for gender, baseline test score, height, and age to increase precision. Panel B reports corresponding unadjusted estimates as a robustness check. Columns (1), (3), and (5) demonstrate that being appointed as a class leader increases Candidate 1 exam performance by approximately 0.33 standard deviations (Panel A). All three results – midterm, final, and combined score – are statistically significant, with p -values between 0.001 and 0.008. The effects on the midterm and combined exam remain significant even when controlling FDR, and the final exam effect remains near significance ($q = 0.061$). Columns (2), (4), and (6) demonstrate no positive effects of leadership service on Candidate 2 exam performance. Candidate 2 coefficients are less precisely estimated however, so we cannot reject the hypothesis that coefficients in both sets of columns might be equal.

Column (7) demonstrates that for Candidates 1, appointment to a leadership position increases his or her rank on the combined midterm and final score by 7.4 percentage points. The effect is statistically significant ($p = 0.004$) but narrowly loses significance when controlling FDR ($q = 0.052$). There is no comparable effect for Candidates 2 in Column (8). Column (9) reveals that among Candidates 1, class leaders are 22 percentage points more likely than non-leaders to believe that serving as a class leader reduces time available for studying ($p = 0.015$, $q = 0.065$). These leaders perform better on tests in spite of the demands of their positions. Among Candidates 2, there is no evidence of an effect on this belief.

What mechanisms may plausibly generate the test score effects? We suspect for two reasons that increased studying is the main mechanism through which test scores improve. First, studying is a factor over which students have direct control and can easily manipulate during the course of a semester. Other factors determining test scores tend to be less malleable. Second, the group with test score effects (Candidates 1) also reports that leadership service interferes with time available for studying. In contrast the group with no test score effects (Candidates 2) reports no conflict between leadership service and time available for studying. These patterns are consistent with a model in which Candidate 1 students devote more time to studying when assigned to leadership positions – improving test scores and intensifying their scarcity of time – while Candidate 2 students do not, though this is not the only possible explanation for these patterns.

Several factors may motivate class leaders to increase their study efforts. One possibility is that students appointed as class leaders are afraid that they will be replaced mid-semester if they underperform the rest of the class on the midterm exam. The effects of leadership on midterm scores may therefore be specific to the incentive structure facing class leaders in many Chinese schools. However, the fact that the effect of leadership service on the final exam – at which point the class leaders generally know that they are performing above average – is as large as the effect of leadership service on the midterm (0.328 standard deviations versus 0.322 standard deviations) suggests that direct incentives are not the only mechanism at work. A second possibility is a Rosenthal effect – people may perform better when others place higher expectations on them. In this case, class leaders may not want to disappoint the teachers that appointed them, or they may hope to boost their reputation with an eye towards leadership appointments in future years. Furthermore, their higher

classroom profile may increase the embarrassment associated with a poor performance on an exam.

Table 5 reports effects of leadership service on measures of confidence and aspirations. Columns (1) and (2) demonstrate that appointment as a class leader does not appear to increase a student's perception of where she falls in the class distribution of learning abilities (abilities to understand and apply concepts and methods), though the coefficient for Candidates 2 is non-trivial in magnitude (7.5 percentage points) but imprecisely estimated. Since leadership service increases Candidate 1 performance but does not increase perceived class rank, it seems likely that leadership service may reduce "overconfidence" among Candidates 1. Column (3) reveals that leadership service reduces "overconfidence" – which we construct as the difference between a student's perceived percentile ranking in the grade and his actual percentile ranking on the midterm – among Candidates 1 ($p = 0.013$, $q = 0.065$). Leadership service has no significant effect on overconfidence for Candidates 2 in Column (4). Columns (5) and (6) report results for educational aspirations; students appointed as class leaders aspire to a statistically insignificant 0.2 to 0.3 years more education than non-leaders. Columns (7) through (10) report the effects on students' responses as to how they would behave in a sequential public goods game. Among Candidates 2, Column (8) presents evidence that appointment as a class leader increases a student's willingness to contribute first in the public goods game (34 percentage point effect, $p = 0.016$). However, this effect narrowly loses significance after FDR adjustment ($q = 0.065$). There is no evidence in Columns (9) and (10) that leadership service affects the amount that either candidate type would contribute in the public goods game (the coefficient estimates are 0.5 yuan or less, or approximately 10%).

Table 6 reports the effects of leadership on students' perceived determinants of success. We asked students to rank the three most important factors in determining academic success among the following six options: effort, good teachers, parental fostering, innate talent, help from friends and classmates, and home study environment.¹⁵ If individuals suffer from self-serving bias – a bias to skew one's beliefs to support one's own interests – then we expect that they may attribute success to their own efforts and failure to external factors (Duval and Silvia 2002). Of particular salience is a form of self-serving bias known as

¹⁵ The survey also offered a seventh choice, luck during exams. In practice, however, virtually no candidates ranked this choice among the top three.

attribution bias, in which agents attribute success to their own efforts and failure to exogenous factors (Van den Steen 2004). Thus, in the presence of this bias, those appointed as class leaders may be convinced that their success is a product of their own internal efforts, while those who are not appointed may believe that their lesser success is due to external factors. Empirical evidence of self-serving bias varies by context. Babcock and Loewenstein (1997) document cases in which it appears to be a factor in impeding bargaining, and Billett and Qian (2008) find that it may lead to overconfidence among CEOs. However, Dahl and Ransom (1999) find no evidence that it plays a role in determining one's fair contribution to charity.

Columns (1) and (2) report results from regressing an indicator for listing effort as the most important determinant of success on assignment to a leadership position.¹⁶ Leadership service increases the probability of listing effort as the most important determinant of success by a statistically insignificant 21 percentage points among Candidates 1 ($p = 0.158$) and a statistically significant 26 percentage points among Candidates 2 ($p = 0.015$). The latter result narrowly loses significance after FDR adjustment ($q = 0.065$). It is notable that the coefficients for Candidates 2 are larger than the coefficients for Candidates 1. This suggests that class leaders incorrectly claim credit for their success in that Candidates 2 would not have been the first choice for leader absent the randomization procedure.

Columns (3) through (12) report results of regressions with indicators for listing other factors as the most important determinants of success. The coefficient estimates in these columns are generally negative, but none achieve statistical significance. In an accounting sense, most of the increase in the probability of reporting effort as the greatest determinant of academic success comes from decreases in reporting teachers (third and fourth columns) and parental fostering (fifth and sixth columns) as the most important determinants of academic success.

A natural question is whether the effect of leadership service on the perception of effort as the most important factor in success is due to “self-aggrandizement” by candidates appointed as class leaders or resentment by candidates not appointed as class leaders. On the one hand, candidates appointed as class leaders may view their appointment as earned rather than due (in part) to the randomization procedure, which they are unaware of. On the other

¹⁶ We find similar results if we instead measure whether effort is listed as one of the two most or three most important factors in determining academic success.

hand, candidates not appointed as class leaders may be disappointed with being passed over and may believe (correctly) that external factors are to blame. To distinguish between these possibilities, we asked students about the most important determinants of success at the beginning of the semester as well as at the end of the semester. Relative to students that were not selected as leadership candidates at all, Candidate 1 leaders experienced a 7.0 percentage point increase in the probability of reporting effort as the most important determinant of success, and Candidate 2 leaders experienced a 21.3 percentage point increase in the probability of reporting effort as the most important determinant of success. In comparison, Candidate 1 non-leaders experienced a 2.5 percentage point decrease in the probability of reporting effort as the most important determinant of success, and Candidate 2 non-leaders experienced a 9.7 percentage point decrease in the probability of reporting effort as the most important determinant of success.¹⁷ The majority of the effect thus appears to be the result of leadership appointment causing an increase in the belief that effort determines success rather than failure to be appointed causing a decrease in the belief that effort determines success.

Table 7 presents the effects of leadership service on social networks and popularity. Columns (1) and (2) regress the number of close friends that a candidate reports on assignment to a class leadership position. Leadership service doubles the number of reported close friends among Candidates 2, but this result does not achieve statistical significance ($p = 0.066$, $q = 0.183$). Furthermore, Column (4) demonstrates that Candidates 2 that serve as leaders are 27 percent less likely to be listed as close friends by other students ($p = 0.053$, $q = 0.163$). Columns (5) and (6) of Table 7 investigate potential incumbency effects. Candidates 1 appointed to leadership positions receive 9.5 more votes in a hypothetical election – a 262% increase – than non-leader candidates ($p = 0.008$, $q = 0.061$). This effect is not mirrored among Candidate 2 leaders, who receive only 2.2 additional votes in a hypothetical election ($p = 0.171$, $q = 0.364$). This suggests that the incumbency advantage is strongest for the most qualified candidates (Candidates 1). Leadership service alone may be insufficient for generating an effect on popularity. Rather, it is the interaction between serving as a class leader and a student’s qualifications that generates large effects on future electability, perhaps because “voters” (other students) learn about the competency of the leader. Alternatively,

¹⁷ The overall trend among non-candidate students during this period was a 7 percentage point decrease in the probability of reporting effort as the most important determinant of success.

even if most second candidates are sufficiently competent, they may still face a “backlash” from students who, like the teacher, believe that Candidate 1 was the best pick. In that case the incumbency advantage for Candidates 2 would be minimal, as we find.

6. Discussion

Overall we find effects of leadership service along several dimensions. For students most likely to become class leaders (first candidates), leadership service increases their test scores by 0.3 standard deviations and more than triples their political popularity. These effects represent the effects of the treatment on the treated under the typical class leader selection process. They persist for at least five months (the duration of the semester), though we cannot be certain how long they endure following that period. For students who would be runners-up for class leadership positions (second candidates), leadership service appears to enhance confidence. For these students there is evidence of increases in their willingness to lead by example and their belief that effort is the most important determinant of success, and weak evidence suggesting that their perceived number of friends and overconfidence may increase. These effects represent the effects that might occur if class leadership opportunities were expanded to the next tier of qualified students.

Our use of nonparametric permutation tests for statistical inference, combined with false discovery rate control, increases our confidence that the statistically significant coefficients represent real effects. Nevertheless, our sample sizes are small, and our point estimates in general are not precise. The magnitudes of our estimates should thus be interpreted with caution; the true effects could differ meaningfully in magnitude from our point estimates.

An alternative way of evaluating the evidence in our paper is to consider the “post-study” (i.e., posterior) probabilities that our results imply about the likelihood that β_1 , the causal effect of leadership service on test scores, is nonzero (Maniadis, Tufano, and List 2014). For simplicity we assume within a Bayesian framework that β_1 either equals zero with probability p or equals an alternative value with probability $(1 - p)$. The post-study probabilities naturally depend on one’s prior about the probability that leadership service affects test scores (p), so we calculate post-study probabilities for a variety of priors. To calculate post-study probabilities we also need to know the value of β_1 when it is nonzero,

which we take from the data to be 0.325 (see Column (5) of Table 4), and the standard error of $\hat{\beta}_1$, which we set at 0.10.¹⁸ We adjust the post-study probabilities to reflect the fact that we test for effects on two candidate types by setting the required threshold for statistical significance to 0.025 (or 0.005) instead of 0.05 (or 0.01).¹⁹

Table 8 reports post-study probabilities that β_1 is nonzero under different assumptions. In Column (1) we set the threshold for declaring a result significant at $\alpha=0.05$, and in Column (2) we set the threshold at $\alpha=0.01$. In first row of Column (1) we assume that one’s prior is that there was only a 5% chance that leadership service could affect academic performance (i.e., the paper’s main result is very surprising).²⁰ In that case, following the experiment, the probability that our finding represents a true effect of leadership experience on academic performance is 46%. If one’s prior increases to a 10% chance that leadership service could affect academic performance, the post-study probability rises to 65%. Further increases in the prior to 25% and 50% raise the post-study probability to 85% and 94% respectively. Column (2) reports similar figures when the threshold for declaring a result significant is set to 0.01 (this correspond to a t -statistic of approximately 2.6). The post-study probability now rises to 76% when one’s prior is only 5%, and it reaches 87% when one’s prior is 10%. A prior of 25% implies a post-study probability of 95%.²¹

Many economists may be unfamiliar with interpreting post-study probabilities. In particular, a post-study probability of 46% in the first row, corresponding to a skeptical prior of 5%, suggests that the chance of a false positive is quite high (54%). However, it is mechanically impossible for a post-study probability to exceed 51% if one’s prior is 5% and α is set at 0.05, because even as power approaches 100% the probability of a true positive cannot exceed 5% – the approximate chance of a false positive. This fact illustrates a deeper

¹⁸ The relevant standard error in Column (5) of Table 4 (Panel A) is only 0.047. However, as we discuss in Section 4, we believe that clustering issues may seriously affect the analytic standard errors. Thus we instead use a more conservative estimate of 0.10, which reflects the standard error that would generate a p -value equivalent to the permutation-based p -value that we compute in Column (5) of Table 4 (Panel A).

¹⁹ Since there are two tests, the per-study probability of at least one false positive, conditional on β_1 being zero, remains at 0.05 when we set the significance threshold at 0.025.

²⁰ This statement naturally assumes a Bayesian perspective. The equivalent statement from a frequentist perspective would be that, on average, only one in twenty hypotheses that are as “surprising” as this one turn out to be true.

²¹ We also experimented with “shrinking” the value of β_1 that we use towards zero to reflect the fact that we likely would not be conducting this exercise if $\hat{\beta}_1$ were not statistically significant. This had only a modest effect on the post-study probabilities.

point, which is that it makes little sense to apply a lenient significance threshold of $\alpha=0.05$ if one has a skeptical prior; a skeptic would instead demand stronger evidence of an effect and apply a more conservative threshold of $\alpha=0.01$ or even $\alpha=0.005$ (not shown in Table 8). At these thresholds, the post-study probabilities for a 5% prior rise to 76% and 85% respectively. In summary, the post-study probabilities imply that even someone who was deeply skeptical that classroom leadership service could have any effect on student test scores would revise their views to the point at which they think it is quite plausible that leadership service affects test scores.

A key question in applying our results is whether they might generalize to other contexts. We consider this question in three parts. First, how likely are the effects to generalize beyond the school in our experiment? Second, do the effects that we measure – test scores, hypothetical vote share, willingness to lead by example, and belief that effort is the most important determinant of success – have meaning outside the classroom context in which they were generated? Finally, could leadership service in other contexts produce similar effects?

We consider the first question – whether our results might generalize to other (Chinese) schools – through the lens of the theory of generalizability developed in Al-Ubaydli and List (2012). Al-Ubaydli and List note that we may specify heterogeneous treatment effects as a function of a random vector of characteristics Z_s , some of which may be unobservable. Let z_1 be the realization of Z_s for our study school. Results are “informative” for school s if they cause us to update our priors on school s ’s treatment effects. Results are globally generalizable if they are informative for schools with realizations of Z_s outside a neighborhood around z_1 ; they are locally generalizable if they are informative for schools with realizations of Z_s within a neighborhood around z_1 ; and they have zero generalizability if they are never informative unless $Z_s = z_1$.²²

We view global or local generalizability as the most plausible possibilities in our context. Zero generalizability requires that the treatment effects change discontinuously at z_1 , or that z_1 be completely outside the real-world support of Z_s . Neither case seems plausible in our context; the latter case is impossible since our experiment occurs in a natural field

²² In the fully abstract case, generalizability also depends on the values of the treatment in the experiment, and how those values compare to values in the “target space” that we might be interested in. For example, the treatment could be an income tax rate that might realize values from 0 to 100 percent. Since our treatment is binary, however, we do not consider this issue here.

context. When considering local versus global generalizability, we hypothesize that the most important dimensions of Z_i are the leader selection process and the importance that students attach to class leadership positions, as these directly affect students' incentives and the prestige of the positions. In that sense we expect that our results are globally generalizable, rather than locally generalizable, because the leader selection process is broadly similar across Chinese schools, and there is high demand for leadership positions in most Chinese schools, including the one we study.²³ Global generalizability does not imply that our results are informative for all Chinese schools, but it does imply that they are informative for a nontrivial fraction of them.

The second question – whether our outcomes have meaning outside the specific classroom context – is similar to considering whether results from laboratory experiments might generalize to real-world contexts. Levitt and List (2007) note six major issues in generalizing results from experiments: selection of participants into the experiment, differences in market experience, short-run versus long-run effects, size of stakes, endogenous institutions, and group differences. Of these factors, the first two are least likely to pose problems in our experiment. The selection of participants into the study was unaffected by the experimental design since all students were potentially eligible to hold leadership positions and teachers were free to rank leadership candidates as they naturally would. Participants' "market experience" upon entry to the experiment – leadership experience in this case – was also typical for students of this age. This holds most strongly for the Candidate 1 students, who should have the same leadership experience as a typically selected classroom leader.

The relevance of the remaining four issues varies by outcome. For test scores, the issue of short-run versus long-run effects is most relevant. We would characterize our test score outcomes as "medium run" in that the interval between initiation of the treatment and measurement of the outcome measures in weeks or months. This interval is considerably longer than the intervals in many laboratory experiments, which typically measure in minutes or hours. However, as with many educational interventions, we do not have long-run test scores or achievement measures from many years after the treatment.

²³ We also note that the individual participation decision causes fewer issues in our study than in many field experiments. Students' participation decisions are not a factor because we have a natural field experiment, and the group of students participating is representative of the group of students that would naturally participate (particularly when focusing on Candidates 1).

For the hypothetical vote share measure, which is similar to a poll, the issue of endogenous institutions is most relevant. Since there is no actual election, none of the class leaders or potential candidates engages in any campaigning. If the main incumbency advantage is name recognition, this advantage might diminish if other candidates began to campaign seriously. However, since the incumbency advantage appears larger for Candidates 1 than Candidates 2, it seems unlikely that the advantage derives solely from name recognition.

In the public goods game (willingness to lead by example) size of stakes is the primary issue; it is possible that class leaders volunteered more quickly to go first than they would have had they been playing with real money. Size of stakes is also an issue in the determinants of success question, but the bias could go in either direction in that case. On the one hand, students might be more willing to claim credit for their own success in a private survey than they would be publicly. On the other hand, students might be more prone to self-serving bias when there are clear economic gains (e.g., when claiming credit for success during contract negotiations) than in a survey question.

The last generalizability issue – group differences – does not apply in the typical sense because we do not document differences in effects by gender, race, or age. However, we do find differences in effects by the propensity of leadership appointment. The effects for first candidates appear in test scores and political popularity, while the effects for second candidates appear in confidence-related measures. Whether these differences might change in the real world depends in part upon whether expanding leadership opportunities to second candidates would “dilute” the effects of leadership service for either group. In economic terms, the general equilibrium effects could be different than the partial equilibrium effects that we estimate here.

Finally we consider the question of whether the effects of leadership service in a schooling environment might generalize to other environments, particularly workplace environments. Test score effects cannot directly generalize because cognitive tests are infrequently used outside of academic settings. Nevertheless, to the degree that the test score effects represent performance incentives that class leaders face in maintaining their leadership positions, then we might expect that individuals in other contexts would respond along dimensions that are important for maintaining their leadership or managerial positions. Workplace promotions in particular might incentivize employee performance, even after

conditioning on salary raises. The incumbency advantage could plausibly generalize to other contexts as well. Indeed, Lee (2008) finds large incumbency advantages in Congressional elections. Confidence-related effects could generalize since confidence is a product of multiple factors in a variety of settings, but the magnitude would likely vary by measure and context. Future experimental research might extend these results to other contexts or test for persistence of effects using long-term follow-up data.

References

- Al-Ubaydli, Omar, and John A. List. 2012. *On the Generalizability of Experimental Results in Economics*. National Bureau of Economic Research. Working Paper. <http://www.nber.org/papers/w17957> (February 4, 2015).
- Anderson, Cameron, and Adam D. Galinsky. 2006. "Power, Optimism, and Risk-Taking." *European Journal of Social Psychology* 36(4): 511–36.
- Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103(484): 1481–95.
- Andreoni, James. 2006. "Leadership Giving in Charitable Fund-Raising." *Journal of Public Economic Theory* 8(1): 1–22.
- Babcock, Linda, and George Loewenstein. 1997. "Explaining Bargaining Impasse: The Role of Self-Serving Biases." *The Journal of Economic Perspectives* 11(1): 109–26.
- Beaman, Lori et al. 2009. "Powerful Women: Does Exposure Reduce Bias?" *The Quarterly Journal of Economics* 124(4): 1497–1540.
- Beaman, Lori, Esther Duflo, Rohini Pande, and Petia Topalova. 2012. "Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India." *Science* 335(6068): 582–86.
- Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. "Adaptive Linear Step-up Procedures That Control the False Discovery Rate." *Biometrika* 93(3): 491–507.
- Bertrand, Marianne, and Antoinette Schoar. 2003. "Managing with Style: The Effect of Managers on Firm Policies." *The Quarterly Journal of Economics* 118(4): 1169–1208.
- Billett, Matthew T., and Yiming Qian. 2008. "Are Overconfident CEOs Born or Made? Evidence of Self-Attribution Bias from Frequent Acquirers." *Management Science* 54(6): 1037–51.
- Camerer, Colin, and Dan Lovallo. 1999. "Overconfidence and Excess Entry: An Experimental Approach." *The American Economic Review* 89(1): 306–18.

- Chattopadhyay, Raghavendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72(5): 1409–43.
- Dahl, Gordon B., and Michael R Ransom. 1999. "Does Where You Stand Depend on Where You Sit? Tithing Donations and Self-Serving Beliefs." *The American Economic Review* 89(4): 703–27.
- Duval, Thomas Shelley, and Paul J. Silvia. 2002. "Self-Awareness, Probability of Improvement, and the Self-Serving Bias." *Journal of Personality and Social Psychology* 82(1): 49–61.
- Galinsky, Adam D, Deborah H Gruenfeld, and Joe C Magee. 2003. "From Power to Action." *Journal of Personality and Social Psychology* 85(3): 453–66.
- Galinsky, Adam D., Joe C. Magee, M. Ena Inesi, and Deborah H. Gruenfeld. 2006. "Power and Perspectives Not Taken." *Psychological Science* 17(12): 1068–74.
- Güth, Werner, M. Vittoria Levati, Matthias Sutter, and Eline van der Heijden. 2007. "Leading by Example with and without Exclusion Power in Voluntary Contribution Experiments." *Journal of Public Economics* 91(5–6): 1023–42.
- Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42(4): 1009–55.
- Hermalin, Benjamin E. 1998. "Toward an Economic Theory of Leadership: Leading by Example." *The American Economic Review* 88(5): 1188–1206.
- Jones, Benjamin F., and Benjamin A. Olken. 2005. "Do Leaders Matter? National Leadership and Growth Since World War II." *The Quarterly Journal of Economics* 120(3): 835–64.
- Kuhn, Peter, and Catherine Weinberger. 2005. "Leadership Skills and Wages." *Journal of Labor Economics* 23(3): 395–436.
- Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142(2): 675–97.
- Levitt, Steven D., and John A. List. 2007. "Viewpoint: On the Generalizability of Lab Behaviour to the Field." *Canadian Journal of Economics* 40(2): 347–70.
- List, John A., Azeem M. Shaikh, and Yang Xu. 2015. *Multiple Hypothesis Testing in Experimental Economics*. University of Chicago. Working Paper.
- Malmendier, Ulrike, and Geoffrey Tate. 2005. "CEO Overconfidence and Corporate Investment." *The Journal of Finance* 60(6): 2661–2700.
- Maniadis, Zacharias, Fabio Tufano, and John A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review* 104(1): 277–90.

Rosenbaum, Paul R. 2007. "Interference Between Units in Randomized Experiments." *Journal of the American Statistical Association* 102: 191–200.

Van den Steen, Eric. 2004. "Rational Overoptimism (and Other Biases)." *The American Economic Review* 94(4): 1141–51.

Figure 1: Class Management Structure

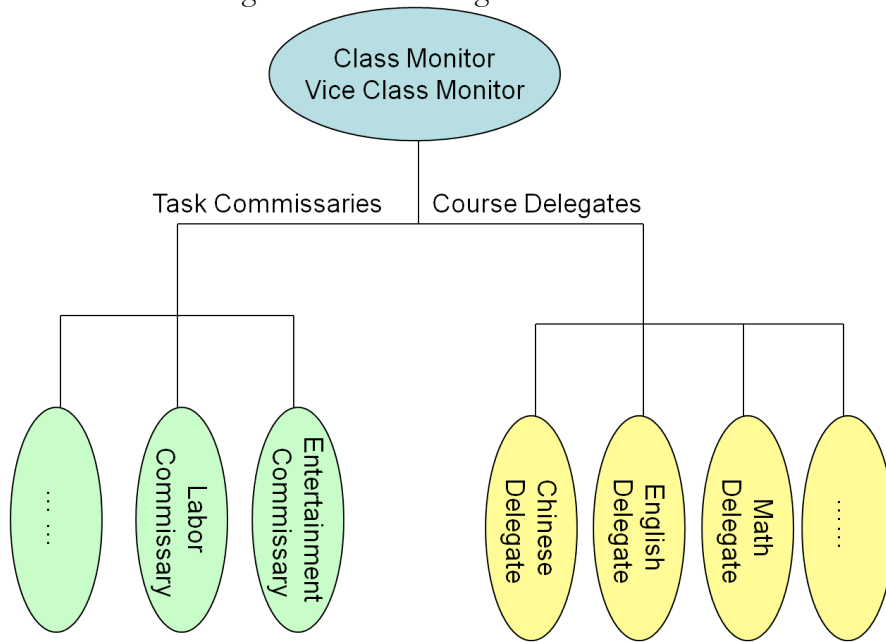


Figure 2: Study Timeline

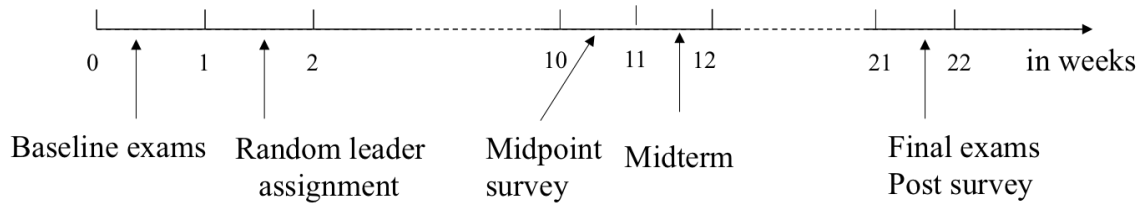


Table 1: Comparisons Between Study Areas and All Areas

Variable	All Rural	Study Rural	All Areas	Study Area
Years of Education	7.9 (2.4)	8.2* (2.0)	8.8 (2.8)	8.8 (2.4)
Education \geq 9 years	0.62 (0.49)	0.68* (0.47)	0.72 (0.45)	0.75* (0.44)
HH size	4.3 (1.5)	4.4 (1.4)	4.2 (1.5)	4.3 (1.4)
Running water	0.23 (0.42)	0.30 (0.46)	0.40 (0.49)	0.42 (0.49)
Toilet available	0.69 (0.46)	0.81* (0.39)	0.70 (0.46)	0.76* (0.43)
Households	36,436	135	53,300	186

Notes : Standard deviations in parentheses. * denotes statistically different than All Rural/All Areas average at 5% level. Data come from a 0.1% sample of the 2000 Chinese Census. Education statistics are for households with parents born between 1960 and 1980. Other characteristics are for households with children born between 1995 and 1999.

Table 2: Summary Statistics

	Leadership Candidates			Leadership Non-Candidates			Difference in Means	Difference <i>t</i> -statistic
	Mean	Std Dev	N	Mean	Std Dev	N		
<i>Panel A: Covariates</i>								
Baseline Score	0.40	0.97	90	-0.13	0.98	279	0.53	4.5
Male	0.45	0.50	92	0.63	0.48	285	-0.19	-3.2
Age	13.3	0.9	92	13.5	0.9	281	-0.2	-1.4
Height (cm)	158.5	8.6	92	157.7	8.3	281	0.8	0.8
Birth Order	1.64	0.86	91	1.95	1.02	260	-0.31	-2.6
Father's Edu (years)	8.27	2.35	89	8.51	2.23	220	-0.24	-0.9
Father's Edu (years)	6.87	2.71	87	7.44	2.50	203	-0.56	-1.7
Relative Income	2.80	0.70	86	2.89	0.70	184	-0.08	-0.9
<i>Panel B: Outcomes</i>								
Midterm Score	0.51	0.98	92	-0.17	0.95	272	0.68	5.9
Final Score	0.50	1.00	92	-0.17	0.94	270	0.67	5.8
Combined Exam Score	0.50	0.98	92	-0.17	0.92	269	0.67	5.9
Combined Percentile Rank	64.9	28.4	92	45.1	27.3	269	19.8	5.9
Leadership Affects Studies	0.26	0.44	82	0.27	0.44	187	-0.01	-0.2
Perceived Percentile Rank	76.6	12.2	88	69.1	14.0	262	7.5	4.4
Overconfidence	11.3	23.8	88	23.5	25.1	260	-12.2	-4.0
Edu Aspirations	16.7	3.2	88	15.0	3.2	242	1.7	4.3
Contribute First	0.76	0.43	70	0.66	0.47	160	0.09	1.4
Amount Contributed (yuan)	5.28	2.81	64	5.38	3.04	138	-0.10	-0.2
Strongest Determinant of Success:								
Effort	0.51	0.50	87	0.36	0.48	234	0.14	2.3
Good Teachers	0.24	0.43	87	0.24	0.42	234	0.01	0.1
Parental Fostering	0.05	0.21	87	0.15	0.36	234	-0.11	-2.6
Innate Talent	0.06	0.23	87	0.15	0.36	234	-0.09	-2.2
Help from Classmates	0.06	0.23	87	0.05	0.21	234	0.01	0.4
Home Study Environment	0.08	0.27	87	0.04	0.20	234	0.04	1.3
Number of Friends	6.1	5.1	84	7.7	12.5	223	-1.6	-1.2
Times Listed as Close Friend	1.9	1.4	92			0		
Votes Received	5.6	8.4	92			0		

Notes: Leadership candidates are students who were nominated by teachers to enter the random assignment lottery for leadership positions. Leadership non-candidates are all other students in the seven study classrooms.

Table 3: Tests for Random Assignment of Baseline Characteristics

<i>Dependent Variable:</i>	Baseline Test Score	Male	Height	Age	Birth Order	Father's Education	Mother's Education	Family Income
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Independent Variable:</i>								
<i>Panel A: Candidate 1</i>								
Leadership Position	0.266 (0.336) <i>0.375</i>	-0.043 (0.144) <i>0.750</i>	-0.4 (1.9) <i>0.798</i>	-0.3 (0.2) <i>0.170</i>	(0.06) (0.22) <i>0.789</i>	-0.091 (0.66) <i>0.931</i>	-0.23 (0.95) <i>0.834</i>	0.09 (0.23) <i>0.719</i>
N	45	46	46	46	45	44	42	41
<i>Panel B: Candidate 2</i>								
Leadership Position	-0.415 (0.307) <i>0.250</i>	0.000 (0.154) <i>0.950</i>	0.3 (2.3) <i>0.859</i>	0.0 (0.3) <i>0.973</i>	0.35 (0.35) <i>0.335</i>	-1.069 (0.79) <i>0.227</i>	-0.08 (0.47) <i>0.855</i>	-0.03 (0.24) <i>0.912</i>
N	45	46	46	46	46	45	45	45
Mean of Dependent Variable (Controls)	0.438	0.457	158.5	13.4	1.57	8.57	6.95	2.79

Notes: Each column represents a separate linear regression of the dependent variable on an indicator for assignment to a leadership position. The sample in Panel A (Panel B) is limited to students who were the first (second) candidate nominated by teachers for leadership positions. Parentheses contain standard errors clustered by classroom. Permutation-based *p*-values are reported in italics. The mean of the dependent variable is reported for leadership candidates (both first and second) who were not assigned to leadership positions.

Table 4: Effects of Leadership Positions on Test Scores and Studying

<i>Dependent Variable:</i>	Midterm Score		Final Score		Combined Exam Scores		Combined Score Percentile		Leadership Negatively Affects Studying	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
<i>Leadership candidate:</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
<i>Independent Variable:</i>										
<i>Panel A: Covariate-adjusted regressions</i>										
Leadership position	0.322 (0.050) <i>0.001</i> [0.016] {0.013}	-0.050 (0.209) <i>0.815</i> [0.865] {1.000}	0.328 (0.067) <i>0.008</i> [0.061] {0.105}	-0.011 (0.209) <i>0.960</i> [0.948] {1.000}	0.325 (0.047) <i>0.000</i> [0.001] {0.000}	-0.031 (0.202) <i>0.878</i> [0.865] {1.000}	7.41 (1.62) <i>0.004</i> [0.052] {0.027}	-0.98 (5.73) <i>0.878</i> [0.865] {1.000}	0.224 (0.061) <i>0.015</i> [0.065] {0.150}	0.003 (0.092) <i>0.973</i> [0.948] {0.986}
<i>Panel B: Unadjusted regressions</i>										
Leadership Position	0.307 (0.060) <i>0.005</i> [0.085] {0.080}	0.073 (0.129) <i>0.623</i> [0.829] {0.866}	0.309 (0.096) <i>0.028</i> [0.125] {0.357}	0.097 (0.176) <i>0.621</i> [0.829] {1.000}	0.308 (0.072) <i>0.011</i> [0.085] {0.161}	0.085 (0.145) <i>0.594</i> [0.829] {0.895}	5.80 (2.21) <i>0.036</i> [0.147] {0.239}	2.45 (4.53) <i>0.645</i> [0.829] {0.943}	0.218 (0.063) <i>0.025</i> [0.125] {0.258}	0.000 (0.112) <i>0.960</i> [1.000] {0.973}
Mean of dependent variable (controls)	0.467	0.448	0.400	0.494	0.434	0.471	63.86	64.12	0.100	0.300
N	46	46	46	46	46	46	46	46	42	40

Notes: Each column represents a separate regression of the dependent variable on an indicator for assignment to a leadership position. The sample is limited to students nominated by teachers for leadership positions. Test score measures are differenced relative to the relevant score at baseline. Covariate-adjusted regressions control for gender, baseline test score, height, and age. Parentheses contain standard errors clustered by classroom. Permutation-based p -values are reported in italics, FDR adjusted q -values are reported in brackets, and FWER adjusted p -values are reported in set braces. The mean of the dependent variable is reported for leadership candidates who were not assigned to leadership positions.

Table 5: Effects of Leadership Positions on Confidence and Aspirations

	Perceived Percentile		Overconfidence (Perceived		Educational		Willing to Go First in		Contribution in Public	
<i>Dependent Variable:</i>	Rank in Class		Rank vs Actual Rank)		Aspirations (Years)		Public Goods Game		Goods Game	
<i>Leadership Candidate:</i>	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
<i>Independent Variable:</i>										
<i>Panel A: Covariate-adjusted regressions</i>										
Leadership Position	1.12	7.49	-6.69	9.52	0.31	0.22	0.135	0.342	-0.27	0.51
	(2.00)	(3.81)	(1.65)	(4.80)	(1.03)	(1.06)	(0.139)	(0.120)	(0.79)	(1.45)
	<i>0.622</i>	<i>0.104</i>	<i>0.013</i>	<i>0.082</i>	<i>0.776</i>	<i>0.857</i>	<i>0.374</i>	<i>0.016</i>	<i>0.755</i>	<i>0.740</i>
	[0.814]	[0.253]	[0.065]	[0.214]	[0.865]	[0.865]	[0.682]	[0.065]	[0.865]	[0.865]
	{1.000}	{0.767}	{0.094}	{1.000}	{1.000}	{0.857}	{1.000}	{0.307}	{1.000}	{1.000}
<i>Panel B: Unadjusted regressions</i>										
Leadership Position	2.90	5.69	-9.34	15.97	0.41	0.00	0.213	0.353	-0.57	0.83
	(3.52)	(3.72)	(5.47)	(8.53)	(1.02)	(1.27)	(0.118)	(0.123)	(0.60)	(1.12)
	<i>0.430</i>	<i>0.171</i>	<i>0.120</i>	<i>0.107</i>	<i>0.695</i>	<i>1.000</i>	<i>0.095</i>	<i>0.007</i>	<i>0.410</i>	<i>0.499</i>
	[0.647]	[0.356]	[0.266]	[0.250]	[0.839]	[1.000]	[0.244]	[0.085]	[0.643]	[0.722]
	{0.981}	{1.000}	{0.861}	{1.000}	{1.000}	{1.000}	{0.498}	{0.129}	{0.694}	{1.000}
Mean of Dependent Variable (Controls)	76.14	72.52	9.62	9.58	16.68	16.45	0.600	0.647	5.63	4.80
N	44	44	44	44	44	44	36	34	33	31

Notes: Each column represents a separate regression of the dependent variable on an indicator for assignment to a leadership position. The sample is limited to students nominated by teachers for leadership positions. Covariate-adjusted regressions control for gender, baseline test score, height, and age. Parentheses contain standard errors clustered by classroom. Permutation-based p -values are reported in italics, FDR adjusted q -values are reported in brackets, and FWER adjusted p -values are reported in set braces. The mean of the dependent variable is reported for leadership candidates who were not assigned to leadership positions.

Table 6: Effects of Leadership Positions on Perceived Determinants of Success

The Greatest Factor in Academic Success Is:														
<i>Dependent Variable:</i>	Effort				Parental Fostering				Innate Talent		Help from Classmates		Home Study Environment	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd		
<i>Leadership Candidate:</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)		
<i>Independent Variable:</i>														
<i>Panel A: Covariate-adjusted regressions</i>														
Leadership Position	0.207	0.261	-0.240	-0.099	-0.053	-0.146	-0.011	-0.035	0.055	-0.041	0.092	0.059		
	(0.134)	(0.074)	(0.153)	(0.156)	(0.060)	(0.093)	(0.075)	(0.061)	(0.103)	(0.033)	(0.091)	(0.085)		
	<i>0.158</i>	<i>0.015</i>	<i>0.198</i>	<i>0.520</i>	<i>0.829</i>	<i>0.223</i>	<i>0.661</i>	<i>0.642</i>	<i>0.620</i>	<i>0.211</i>	<i>0.331</i>	<i>0.423</i>		
	[0.357]	[0.065]	[0.396]	[0.805]	[0.865]	[0.396]	[0.814]	[0.814]	[0.814]	[0.396]	[0.661]	[0.682]		
	{1.000}	{0.204}	{1.000}	{1.000}	{1.000}	{1.000}	{0.678}	{1.000}	{0.648}	{0.464}	{1.000}	{0.706}		
<i>Panel B: Unadjusted regressions</i>														
Leadership Position	0.255	0.318	-0.268	-0.136	-0.045	-0.136	0.002	-0.045	0.004	-0.045	0.097	0.045		
	(0.121)	(0.081)	(0.133)	(0.121)	(0.047)	(0.068)	(0.077)	(0.040)	(0.106)	(0.046)	(0.095)	(0.088)		
	<i>0.051</i>	<i>0.008</i>	<i>0.098</i>	<i>0.286</i>	<i>0.936</i>	<i>0.215</i>	<i>0.983</i>	<i>0.304</i>	<i>0.941</i>	<i>0.819</i>	<i>0.305</i>	<i>0.707</i>		
	[0.179]	[0.085]	[0.244]	[0.471]	[1.000]	[0.379]	[1.000]	[0.471]	[1.000]	[0.986]	[0.471]	[0.839]		
	{0.479}	{0.113}	{1.000}	{0.935}	{1.000}	{1.000}	{1.000}	{0.511}	{0.983}	{1.000}	{1.000}	{1.000}		
Mean of Dependent Variable (Controls)	0.364	0.364	0.364	0.318	0.045	0.136	0.045	0.091	0.091	0.045	0.045	0.045		
N	43	44	43	44	43	44	43	44	43	44	43	44		

Notes: Each column represents a separate regression of the dependent variable on an indicator for assignment to a leadership position. The sample is limited to students nominated by teachers for leadership positions. Covariate-adjusted regressions control for gender, baseline test score, height, and age. Parentheses contain standard errors clustered by classroom. Permutation-based p -values are reported in italics, FDR adjusted q -values are reported in brackets, and FWER adjusted p -values are reported in set braces. The mean of the dependent variable is reported for leadership candidates who were not assigned to leadership positions.

Table 7: Effects of Leadership Positions on Social Networks and Popularity

	Number of Close		Number Who Report		Number Who Would	
<i>Dependent Variable:</i>	Friends		You as Close Friend		Vote for You	
<i>Leadership Candidate:</i>	1st	2nd	1st	2nd	1st	2nd
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Independent Variable:</i>						
<i>Panel A: Covariate-adjusted regressions</i>						
Leadership Position	0.75	4.40	0.76	-0.55	9.45	2.23
	(0.88)	(1.81)	(0.26)	(0.25)	(2.17)	(1.33)
	<i>0.390</i>	<i>0.066</i>	<i>0.038</i>	<i>0.053</i>	<i>0.008</i>	<i>0.171</i>
	[0.682]	[0.183]	[0.123]	[0.163]	[0.061]	[0.364]
	{0.470}	{1.000}	{0.456}	{0.263}	{0.106}	{0.391}
<i>Panel B: Unadjusted regressions</i>						
Leadership Position	0.24	4.86	0.74	-0.52	10.30	1.30
	(0.87)	(1.28)	(0.32)	(0.26)	(2.77)	(1.74)
	<i>0.792</i>	<i>0.007</i>	<i>0.069</i>	<i>0.073</i>	<i>0.014</i>	<i>0.481</i>
	[0.980]	[0.085]	[0.216]	[0.216]	[0.086]	[0.722]
	{0.953}	{0.138}	{0.834}	{0.364}	{0.180}	{1.000}
Mean of Dependent Variable (Controls)	5.29	4.29	1.74	2.04	2.61	2.70
N	42	42	46	46	46	46

Notes: Each column represents a separate regression of the dependent variable on an indicator for assignment to a leadership position. The sample is limited to students nominated by teachers for leadership positions. Covariate-adjusted regressions control for gender, baseline test score, height, and age. Parentheses contain standard errors clustered by classroom. Permutation-based p -values are reported in italics, FDR adjusted q -values are reported in brackets, and FWER adjusted p -values are reported in set braces. The mean of the dependent variable is reported for leadership candidates who were not assigned to leadership positions.

Table 8: Post-Study Probabilities that $\beta \neq 0$

<i>Required significance level</i>	$\alpha = 0.05$	$\alpha = 0.01$
	(1)	(2)
<i>Prior on Probability $\beta \neq 0$:</i>		
0.05	0.46	0.76
0.10	0.65	0.87
0.25	0.85	0.95
0.50	0.94	0.98

Notes: Post-study probabilities that $\beta \neq 0$ represent the probability that $\beta \neq 0$ for a given prior on the probability that $\beta \neq 0$ (which varies by row) and a given significance threshold (which varies by column). All entries assume a standard error of 0.10 and a value of $\beta = 0.325$ if β is nonzero.