How Do Peers Impact Learning? An Experimental Investigation of Peer-to-Peer Teaching and Ability Tracking[*]

Erik O. Kimbrough[†] Andrew D. McGee[‡] Hitoshi Shigeoka[§]

October 4, 2018

Abstract

Classroom peers presumably influence learning by teaching each other. Unfortunately, little is known about peer-to-peer teaching because it is never observed in field studies. The efficacy of this teaching likely depends on the ability of one's peers. We investigate the mechanisms of peer effects experimentally to establish the importance of peer-to-peer teaching and how it is affected by ability tracking—grouping students of similar ability. While peer-to-peer teaching improves learning among low-ability subjects, the positive effects are *offset* by tracking. Tracking reduces peer-to-peer teaching, suggesting that low-ability subjects suffer from the absence of high-ability peers to teach them.

Keywords: Peer-to-peer Teaching, Ability Tracking, Peer Effects, Group Composition, Education and Inequality, Laboratory Experiment
JEL codes: I24, C91, I28

Online Appendix HERE

---

[*] The authors thank Scott Carrell, Julie Berry Cullen, David Deming, Chris Muris and seminar participants at the 2017 Asian and Australasian Society of Labour Economics (AASLE) Inaugural Conference, the Kyoto Summer Workshop on Applied Economics, the University of Alaska-Anchorage Conference on Contests and Innovation, the 2016 Canadian Economic Association Annual Conference, the 2017 Economic Science Association Annual Meetings, the 2016 Southern Economic Association Annual Meetings, and the University of Arkansas for their suggestions. Catherine Michaud-Leclerc and Hanh Tong provided outstanding assistance in conducting the experiments, as did Erli Suo and Jim Sylvester in developing the experimental software and Eric Adebayo and Garrett Petersen in analyzing the audio recordings. We gratefully acknowledge the financial support from SSHRC Insight Development Grant #430-2013-1067. All remaining errors are our own.
[†] Smith Institute for Political Economy and Philosophy, Argyros School of Business and Economics, Chapman University, One University Drive, Orange, CA 92866, US. E-mail: ekimbrough@gmail.com
[‡] Department of Economics, University of Alberta, 8-14 Tory Building, Edmonton, AB T6G 2H4, Canada, and IZA. E-mail: mcgee1@ualberta.ca
[§] Corresponding Author: Department of Economics, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada, and NBER. E-mail: hitoshi_shigeoka@sfu.ca

1. Introduction

The effects of peers on educational outcomes have been studied widely, and a broad consensus exists that peers have non-trivial effects on students' learning (see Epple and Romano [2011] and Sacerdote [2011] for recent reviews). How peers influence learning, however, has proven much more difficult to answer. One key channel for the observed peer effects is that students teach each other in and out of the classroom. Indeed, first among the possible means of peer influence listed in Sacerdote [2011] is direct learning from classmates—what we refer to as peer-to-peer teaching.[5] Despite the self-evident preeminence of peer-to-peer teaching, there has been precious little research concerning this mechanism in economics.

If peer-to-peer teaching is important for learning, then the composition of a classroom may determine its effectiveness. Ability tracking—grouping students of similar abilities in classrooms—is one such policy that affects classroom composition. While widely employed, ability tracking is a contentious practice in education (see Betts [2011] for a review). An important unresolved question is how ability tracking affects peer-to-peer teaching, and whether these effects on peer-to-peer teaching depend on a student's place in the ability distribution. On the one hand, ability tracking may hurt low-ability students if these students benefit from interactions with high-ability peers that no longer occur under tracking, thereby exacerbating existing inequalities between high-ability and low-ability students (Epple et al. [2002]). On the other hand, tracking may encourage learning at all ability levels because students of more similar ability may be more effective in teaching one another (Schunk [1991]).

The effects of peer-to-peer teaching, ability tracking, and their interaction on learning are difficult to identify in schools for at least three reasons. First, ability tracking is typically implemented alongside changes to other aspects of the classroom environment such as the curriculum. Second, ability tracking may influence teacher behavior if reducing the heterogeneity in a classroom enables teachers to better tailor the pace and content of instruction. Finally and most importantly, peer-to-peer teaching is rarely—if ever—observed in field data.

Thus, to provide preliminary evidence to fill the gap and complement existing field studies, we conduct a laboratory experiment in which subjects learn to solve logic problems—in our case *Sudoku* problems—and examine how ability tracking affects learning in environments with and without peer-to-peer teaching. A laboratory experiment is useful for investigating the effects of

[5] Other mechanisms listed include motivational effects, effects on the behaviors of teachers, classroom disruptions, and preference formation among others.

peer-to-peer teaching and tracking because we can exogenously vary subjects' ability to interact with each other—which is impossible in schools—and peer group composition. Due to the well-documented econometric challenges in estimating peer effects (notably Manski [1993], Sacerdote [2001], Zimmerman [2003]), many studies exploit plausibly exogenous variation in peer group composition to account for the non-random sorting of students at the classroom level, but we know of no field studies in which the possibility for student *interaction* varies exogenously.[6] In addition, because there is no "real" teacher in our experiment, we can isolate the effects of ability tracking on peer-to-peer teaching and learning from the effects of teachers who may adapt their behavior and instruction to the classroom composition. Furthermore, we can directly measure the frequency of peer-to-peer teaching in the lab.

Of course, the benefits in terms of control and measurement in a lab setting are not without some trade-offs. Our experiment is necessarily conducted in a short period of time, with a limited "curriculum," in a simulated classroom that limits the scope of interaction and abstracts from many important contextual effects that likely operate in field settings. In our view, the benefits exceed these costs given the lack of existing evidence on this fundamental question. Moreover, we see opportunities to extend our laboratory method to field settings such as Massive Online Open Courses (MOOCs), where it may be possible to achieve similar degrees of control over peer interaction. In this sense, laboratory experiments in education can serve as a testbed for more costly implementation in the field.

Our experiments proceeded as follows. Subjects first completed as many Sudoku problems as they could in a 10-minute "Ability" block (T=0), and we use this performance as a proxy for initial ability just as test scores proxy for the same in the education literature. Subjects were then split into two groups based on their measured ability. Half of subjects were assortatively matched into groups of subjects with similar ability (the *"tracked"* treatment), while the other half of subjects were assigned to groups in which subjects of different abilities were evenly distributed (the *"untracked"* treatment). All subjects were informed of their group assignment and rank within the group. Once placed into groups, subjects participated in a 10-minute "Practice" block in which they had a single Sudoku to solve. In the "*teaching*" treatment, subjects in a given group were allowed to chat with each other concerning this practice Sudoku problem. In the "*no-teaching*" treatment, subjects worked on the practice Sudoku by themselves and could not chat

[6] See, for example, Hoxby [2000], Zimmerman [2003], Angrist and Lang [2004], Lyle [2007], Carrell et al. [2009]), Ammermueller and Pischke [2009], Imberman et al. [2012], Lavy, Paserman, and Schlosser [2012], Burke and Sass [2013], and Feld and Zölitz [2017].

with other subjects. Following the Practice block, subjects solved (on their own) as many Sudoku as they could in a 15-minute "Evaluation" block (T=1). Our measure of learning is the change in the average time taken to correctly solve a Sudoku puzzle from the baseline "Ability" block (T=0) to the "Evaluation" block (T=1).

Our objective in comparing the *teaching* and *no-teaching* treatments is to identify the impact of peer-to-peer teaching on learning. Prior studies of peer effects have been unable to identify the importance of this channel because the presence of peer-to-peer teaching is typically endogenous and—more problematically—unobserved. The *tracked* and *untracked* treatments allow us to further identify the importance of peer group composition to the effects of peer-to-peer teaching.[7]

We present three main findings. First, we find that subjects do indeed teach each other when *teaching* is possible, and this peer-to-peer teaching leads to substantial increases in learning. Allowing subjects to teach each other improves learning by 0.12 standard deviations (SD) of the average solving time in the Ability block (or a reduction in the raw average solving time by 11.6 *sec*) relative to the *no-teaching* treatment. This represents a 42 percent increase in learning compared to the mean 27.6 second reduction in average puzzle solving time in the *no-teaching* treatment. Low-ability subjects (as identified in the Ability block) drive nearly all of the gain—likely because they have more room for improvement. Given that there are no incentives for subjects to teach each other (as we did not explicitly force them to chat and the payment is at the individual and not the group level), it is remarkable that only 10 minutes of working together on a single puzzle, as opposed to working on it alone, improves learning by as much as 0.12 SD. We are not aware of any studies in the economics of education that directly show the importance of peer-to-peer teaching to learning.[8]

---

[7] We also varied the compensation scheme in the Evaluation block from *piece-rate* payments to *tournament*-style payments. This treatment manipulation, however, failed to influence observed behavior. Therefore, our analysis of this treatment is relegated to Appendix A.

[8] The importance of peer-to-peer teaching for learning has long been understood by educators (Johnson and Johnson [1975], Slavin [1983]). Earlier experimental studies of peer-to-peer teaching have shown that peers make excellent teachers (see Webb [1989] and Rohrbeck et al. [2003] for meta-analyses). These studies, however, typically compared students in pedagogical treatments in which students were either tasked with group work or offered guidance on how to help other students to students in control groups without these interventions. As such, these studies estimate the marginal contribution of the pedagogical intervention relative to a control group in which peer-to-peer teaching may also have taken place. The lone study of peer-to-peer teaching in economics, Bettinger et al. [2016], focuses on the correlations between the characteristics of peer-to-peer teaching (such as the frequency) and performance in MOOCs. In a sense, their study examines the intensive margin of peer-to-peer teaching rather than the extensive margin as in our study.

Second, we find that *tracking* in the *no-teaching* treatment has little detrimental impact on learning compared to the *untracked* treatment, suggesting there are no direct effects of ability tracking in the absence of peer-to-peer teaching. More importantly, *tracking* has a large detrimental effect on learning in the *teaching* treatment—nearly eliminating the positive effects of teaching. This negative effect of *tracking* in the *teaching* treatment is primarily experienced by low-ability subjects, who again benefitted most from peer-to-peer teaching in the first place.

Finally, we unpack this result by examining the number of instances of peer-to-peer teaching in the audio recordings of subjects' chats during the *teaching* treatment. We find that ability tracking directly reduces the frequency of peer-to-peer teaching, shining light on the mechanism through which ability tracking adversely affects low-ability subjects. That is, ability tracking negatively affects low-ability subjects because peer-to-peer teaching is less common in the absence of higher-ability peers who can teach these subjects.

Our findings have important implications regarding the amount of peer-to-peer teaching to encourage and the composition of student groups. Specifically, the findings suggest that grouping students by prior achievement may disadvantage low-ability students to the extent that these students benefit from being taught by higher-ability peers. Unless ability tracking results in benefits abstracted from in our setting (e.g., curricular customization by teachers), ability tracking alone may harm those students who stand to gain the most from peer-to-peer teaching.

Our findings may also help to reconcile seemingly conflicting findings in the tracking literature. Specifically, Duflo et al. [2011] and Booij et al. [2017]—among others—exploit random assignment to classes to find that tracking has positive effects on low-ability students in Kenya and the Netherlands, respectively, while Garlick [2016] shows that low-ability South African university students perform substantially *worse* when grouped in dormitories with students of similar ability than when they are randomly assigned to dormitories.[9] Because the treatment effects we identify are driven by changes in students' behaviors alone and not contaminated by changes in teachers' behavior or curriculum, our findings suggest that the non-instructional (residential) tracking in Garlick [2016] may harm low-ability students because they

[9] Most studies of tracking find positive or no effects of tracking on student achievement (e.g., Betts and Shkolnik [2000], Figlio and Page [2002], Zimmer [2003], Lefgren [2004], and Betts and Shkolnik [2010]). Cummins [2017] finds a negative effect of tracking on "high-ability" students assigned to a low-ability track when assigned to civil service teachers in Kenya.

no longer benefit from being taught by high-ability peers.[10] On the other hand, our findings suggest that the benefits of ability tracking to low-ability students observed in field settings must stem from the positive effects of tracking on student motivation and student-teacher interactions (consistent with evidence in Booij et al. [2017]) or the customization of instruction (consistent with evidence in Duflo et al. [2011]) that overcome any negative effects on low-ability students of losing out on interactions with higher-ability peers.

Our findings are directly relevant for the optimal design of MOOCs, which often facilitate short-term interactions between peers (e.g., via peer grading and discussion forums as documented in Luo et al. [2014] and Anderson et al. [2014]). Enrollment in MOOCs has grown tremendously over the last decade (Deming et al. [2015]). Students in MOOCs interact in virtual environments similar to that in our experiment. While our study is designed to answer fundamental questions about the way classmates affect learning, the findings also shed light on how to best structure student interaction in virtual classrooms when instruction is not being customized for students. In particular, our findings suggest that learning in MOOCs could be enhanced if these courses encouraged interaction through assignments completed by groups that include students of varied ability. Indeed, the absence of peer-to-peer teaching in virtual classrooms may be one contributing factor behind the adverse effects on student outcomes of online courses relative to traditional in-person classes documented by Bettinger et al. [2017].

Finally, this study is related to a small but recently growing literature in economics using experiments to understand classroom dynamics and education in general by answering questions that may be hard to study effectively in a field setting or using observational data. For example, Calsamiglia et al. [2013] study the effects of affirmative action on subject performance (also using Sudoku puzzles), while Andreoni and Brownback [2017] use all-pay auctions to understand the effects of grading on a curve and group size on student performance. Similar to our study, Eisenkopf [2010] and Ksoll and Lehrer [2013] use framed field experiments to examine the effects of group work on learning among high school students in Switzerland and Ghana, respectively. These studies, however, focus on the impact of peer motivation and peer distraction, and neither study examines the interaction between tracking and group work.

[10] Studies focusing on tracking outside of the classroom (e.g., Carrell et al. [2009], Carrell et al. [2013], Lyle [2007], Sacerdote [2001], and Zimmerman [2003]) also identify tracking effects unexplained by changes in curriculum and teacher behavior, but none focus on the importance of peer-to-peer teaching as in our study.

2. Experimental Design

2.1. Details of the Experimental Design

To model a classroom setting while obtaining data on *individual* performance under our various treatments, each session followed a fixed time sequence with six stages (see Table 1-A). In stage 1, subjects provided demographic information. In stage 2, they completed self-paced instructions about the tasks, and in stage 3 they were shown a common video "lecture" explaining some techniques to improve performance in the tasks. In stage 4, we collected an incentivized measure of subjects' performance individually to establish a baseline estimate of ability. In stage 5, we allowed them an opportunity to practice in an unpaid setting, and finally in stage 6, we collected a second incentivized measure of performance to quantify individual learning. By varying aspects of this environment holding constant stages 1–4, our design identifies the impact(s) of peer-to-peer teaching and ability tracking on learning.

In order to study learning and peer-to-peer teaching in a controlled setting, the experimental task must satisfy a few criteria: 1) performance must be objectively measurable; 2) participants must be able to learn (and teach) a few basic principles that will improve performance; and 3) there must be *ex ante* reason to expect substantial performance/ability differences across individuals in order to facilitate teaching.

Therefore we chose 6×6 Sudoku, logic puzzles in which the goal is to fill in numbers on a 6×6 grid such that each row, column and (pre-defined) 2×3 sub-grid contains exactly one of each integer between 1 and 6. The grid is initially partially filled as in Figure 1, and a Sudoku is correctly filled only if all the constraints are satisfied. Moreover, online searches turned up a variety of Sudoku solving "strategies" that are straightforward to teach and learn, and existing experimental evidence on 9×9 Sudoku puzzles suggests sizable variation in performance across individuals (Calsamiglia et al. [2013]). Next we describe the stages of the experiment in detail.

1. Elicitations: Each session contained 8 subjects. We collected demographic information and incentivized measures of risk attitudes and prosociality, which we include as controls in some analyses (see Appendices D & E for instructions and screenshots).[11] Risk attitudes were elicited via a multiple price list design based on Holt and Laury [2002]. Subjects made nine

[11] In our design phase, we hypothesized that there might be a correlation between these measures of preferences and individual teaching behavior in our various treatments. For example, prosocial subjects might be more willing to teach. However, our audio recordings of teaching behavior were not clear enough to distinguish individuals from one another with high probability. As such, our measures of teaching are at the group level, making such comparisons impossible.

binary choices between option A (a fixed lottery with a 50/50 chance of paying $1 and $3) and option B (a lottery between $0 and $3, where the probability of the higher payoff increases over the choice sequence from 0.1 to 0.9). Subjects were paid for one-randomly selected choice problem from the nine. Our risk measure is the number of times that a subject chose the risky option; higher scores (ranging from 0 to 9) indicate more risk-loving. Prosociality was measured using a $5 dictator game in which all subjects chose as if they were dictators and then were randomly paired, with one randomly selected subject's choice in each pair determining final payments. Our prosociality measure is the dollar amount subjects chose to give to their partners out of $5; higher values reveal more prosociality. Payoffs for these tasks were not revealed until the end of the experiment to avoid any potential influence on behavior in the main experiment.

2–3. Sudoku Instructions and Video: The simulated classroom began with instructions explaining the rules of Sudoku and a common "lecture" seen by all subjects in all treatments. For the lecture, we chose a video explaining some puzzle solving strategies and required all subjects to watch the video before starting the experiment.[12] This ensured that each "student" got the same "lecture" and thereby controlled for the instructor's influence. This is an important design feature because it has been difficult to disentangle the exogenous effect of tracking from the endogenous response of teachers in tracked classrooms. We chose this particular video because it highlights a set of strategies that subjects might also teach to (or reinforce in) one another.

4. Ability Block (T=0): Next, subjects in all treatments were given 10 minutes to work alone on Sudoku puzzles and were paid $0.50 for each puzzle completed correctly in that time. We imposed no constraints on subjects' inputs to the puzzles except that subjects could only enter whole numbers between 1 and 6. Subjects did not learn about their performance on any puzzle until the end of the Block. At that time, they learned only the total number of correctly completed puzzles and received no feedback about which puzzles (if any) were incorrectly completed or the source of such errors. We call this the "Ability Block," and it provides a measure of individual performance (ability), free from peer-influence and measured under an incentive scheme that ensures non-satiation in performance. Prior to the Ability Block, subjects only know that there will be more parts to the experiment, but they have no further information about those parts or about how the Ability Block might influence those parts. For example, subjects do not know that Ability Block performance will be used to assign them to groups.

[12] A link to the video is available here: https://www.dropbox.com/s/hmynir2fhva43z4/VideoInstructions.mp4?dl=0.

After the Ability Block, subjects were told that they had been placed in a group of 4 subjects. An individual's performance (ability) is measured by the number of Sudoku puzzles solved during the Ability Block (T=0).[13] In half of the sessions, subjects were placed into groups in which subjects of different abilities were evenly distributed (*untracked*), and in the other half were placed into groups with subjects of similar ability (*tracked*). This means that in *tracked* sessions the bottom half of performers are all in one group, while in the *untracked* sessions they are divided across both groups. We explicitly balanced ability levels across groups in the *untracked* treatments in an attempt to mimic randomly assigned classes (which are balanced in expectation) to reduce the impact of sampling variation that might arise because of our small group size.[14] See Figure 2 for details. Subjects were told the rules governing the formation of groups in their treatment in the instructions. Importantly, assignment to either group in the *tracked* treatment necessarily gives subjects information about the ability of others in the group. Thus, we displayed information about all group members' performance in both the *untracked* and *tracked* treatments to hold constant the information subjects' received about their peer group.

Next, subjects were told that they would have a chance to practice Sudoku for 10 minutes and that they would then have 15 minutes to solve another block of puzzles for which they would again be paid based on performance. Half of the sessions were told that they would be paid at the same *piece-rate* as in the Ability Block, and half of the sessions were told that they would be paid based on their relative performance in their group (where performance equalled the number of correctly solved puzzles with ties broken by average time spent on each correctly solved puzzle). In the *tournament* sessions, subjects were told that 1st place would earn \$20, 2nd place would earn \$10, 3rd place \$5 and 4th place \$0; these payments were chosen to roughly equalize expected earnings across treatments based on pilot data measuring Sudoku performance.

Thus subjects were informed about both the matching scheme and the incentive system operating during the "Evaluation Block" *prior* to starting the "Practice Block." Thus, knowledge of the matching and incentive schemes could in principle influence decisions to practice, teach and/or learn from others when doing so was possible (i.e., the *teaching* treatments).

5. Practice Block: In the Practice Block, subjects had 10 minutes to work on a single Sudoku puzzle (for which they were not paid). In the *no-teaching* sessions, this was time for

[13] The average solving time on correctly solved puzzles was the tiebreaker when ranking subjects but was not shown to subjects.

[14] One cost of this approach is that all groups are non-randomly constructed, making it difficult to look for the impacts of "bad apples" and "shining stars" explored in other settings (e.g., Lavy, Silva, and Weinhardt [2012]).

individual practice. The *no-teaching* sessions allowed us to measure both the learning that naturally takes place through individual practice and the effects of the *tracking* and *tournament* treatments in the absence of peer-to-peer teaching.

In the *teaching* sessions, the Practice Block also consisted of 10 minutes working on a single puzzle, but here all four group members worked simultaneously on the same shared puzzle, which could be edited by each group member and updated in real time on all group members' screens (Figure 1). Subjects were connected via audio chat and represented on screen by a numbered mouse cursor (from 1–4). All were told that the numbers corresponded to that subject's within-group performance rank in the Ability Block. At the start of the Practice Block they were asked by the proctor to introduce themselves to one another using their number.[15] The purpose of the Practice Block was described in the instructions for the *teaching* treatment as follows (See Appendix D for details and Appendix E for screenshots):

> You can complete this puzzle working with the people in your group. During this period, your microphone will be enabled and a voice chat room will be available in which you can discuss the puzzle you are working on. You may discuss any aspects of the experiment in the chat room, but you may not reveal your identity, make threats, or use inappropriate language (including shorthand like WTF). Other participants will be identified by a number next to their mouse cursor. This is their rank within the group. Please only speak English.

Thus, *teaching* sessions introduced the *possibility* of peer-to-peer teaching and allowed us to measure its presence (or absence) and its effect on performance in the Evaluation Block. The instructions state that this is an opportunity for group work—which we intended to encourage peer-to-peer teaching—but subjects in *teaching* sessions were not provided with any direct incentives (or disincentives) to teach each other (i.e., payment is at the individual level rather than the group level). In *tournament* sessions, there are indirect dis-incentives for teaching: if someone is too effective as a teacher, their "student" may surpass their performance in the next stage, lowering their payoff; however, as we note below we find no evidence for such an effect.

Given these considerations, observed teaching likely results from some combination of intrinsic motivation and an experimenter demand effect, with the added possibility that subjects import norms of peer-helping with them to the lab setting. We adopted this approach because

[15] To limit possible contamination from verbal communication outside of the group audio chat, in all treatments subjects were seated at desks that maximized their physical distance from one another in the lab.

actual classrooms are also usually devoid of explicit incentives for students to teach each other, but we note that we may be observing an upper bound on non-incentivized teaching behavior since subjects do not face much opportunity cost of teaching in the lab setting.

6. Evaluation Block (T=1): The final stage was a 15-minute Evaluation Block in which subjects in all sessions again worked independently to solve Sudoku puzzles. Before starting this Block, subjects were reminded briefly of the incentive scheme (*piece-rate* or *tournament* treatments) and also that they would learn their within-group rank at the end of the experiment (to hold information constant across incentive schemes).[16] The difference in performance in the Ability (T=0) and Evaluation (T=1) Blocks provides our main data on learning in our virtual classroom. After the Evaluation Block ended, subjects were informed about their earnings from the Elicitations and the Ability and Evaluation Blocks and then were called one-by-one to be paid in cash. In addition to their salient earnings from the elicitations and performance pay in the Sudoku tasks, subjects received a $7 payment for arriving to the experiment on time. Average earnings including this show-up payment were approximately $22 for a 70-minute session.

As noted above, we have three binary treatment variables: teaching (*no-teaching/teaching*), tracking (*tracked/untracked*), and incentives (*piece-rate/tournament*). Together these generate a 2×2×2 factorial experimental design, which we applied between subjects. Table 1-B summarizes the design. For each session, a randomly chosen subset of students registered in our database at Simon Fraser University were invited to attend, with restrictions excluding subjects from participating twice. In total we report data from 448 subjects in our Sudoku experiments (56 experimental sessions).[17] We collected data from 6 sessions for each combination of tracking × incentives in the *no-teaching* treatment (24 sessions) and from 8 sessions for each combination of tracking × incentives in the *teaching* treatment (32 sessions).[18]

2.2. Hypotheses

---

[16] We did this in order to reduce causal ambiguity in our design–since it is impossible to pay according to rank without also introducing rank information, but it is possible that this induces non-monetary concerns about ranking, possibly muting the additional effects of introducing monetary rewards based on ranking in our *tournament* treatments. This could account for the lack of a treatment effect and is worthy of further research.

[17] We conducted but do not report data from one pilot session with slightly different parameters, our first three *teaching* sessions in which some subjects' microphones were not working correctly for the audio chat, and one session which was lost when a subject's computer reset during the middle of the experiment.

[18] We ran more *teaching* sessions (32 sessions) than *no-teaching* sessions (24 sessions) due to our interest in the interaction between *teaching* and *tracking*.

As we show in Appendix A, our tournament incentive scheme had a negligible effect on behaviour—we find neither main effects of the treatment nor interactions with the other treatments.[19] As a result, our primary analysis pools data across incentive schemes and focuses on the effects of teaching, tracking, and their interaction—essentially reducing the study to a 2×2 factorial experimental design. Thus, although we had hypotheses about the *tournament* and *piece-rate* treatments, in this section we focus on the three remaining hypotheses of interest, ignoring the negligible effects of the incentive treatments.

Hypothesis 1 (Main Effect of *Teaching*): positive. Assuming that individuals are willing to engage in teaching (perhaps for prosocial reasons), peer-to-peer teaching will have a positive effect on performance as subjects help each other learn to solve puzzles. This effect should be largest among those who perform worst in the Ability Block because they have the most to gain.

Hypothesis 2 (Effect of *Tracked* under *No-teaching*): ambiguous. Given the evidence in the literature of psychological encouragement/discouragement effects from information about relative standing in performance, subjects in the top (bottom) group may be encouraged (discouraged) when learning their relative ranking.[20] The potential for offsetting effects makes the overall effect of tracking in the absence of teaching ambiguous.

Hypothesis 3 (Effect of *Tracked* under *Teaching*): ambiguous. Under the *teaching* treatment, subjects in the *tracked* treatment may have less to teach one another, as the difference between the best and worst students in the group is smaller on average than the difference in the *untracked* treatment. Moreover, tracking may especially hurt subjects in the bottom half of performers since they lose access to the higher-ability peers who could have taught them. In this sense, tracking may attenuate the positive effects of teaching. On the other hand, students of similar ability may be more effective in teaching one another if they find it easier to express difficulties to one another or to target suggestions to address those difficulties. Again, the potential for offsetting effects makes the effect of tracking in the presence of teaching unclear.

---

[19] The detailed analysis is summarized in Appendix A. The means of our measure of learning (described in Section 3.1 below) in the *piece-rate* and *tournament* treatments are very similar (0.306 vs. 0.331). In addition, the distributions of our learning measure—not just the means—are nearly visually indistinguishable across the *piece-rate* and *tournament* treatments (see Appendix Figure A1). We conclude that our incentive structure in the *tournament* treatment may have simply been too linear to induce treatment effects (i.e., changes in effort and teaching behavior) relative to the *piece-rate* treatment.

[20] See for example, Blanes i Vidal and Nossol [2011], Barankay [2012], and Gill et al. [2016].

3. Data

3.1. Outcome Variable

Our dependent variable capturing learning by subjects is the change in the average puzzle solving time per correctly solved Sudoku puzzle between the Ability Block (T=0) and the Evaluation Block (T=1).[21] To ease interpretation, we standardize the average solving time in both periods by subtracting the mean and dividing by the standard deviation of average solving time at T=0 so that average solving time at T=0 has a mean of zero and a standard deviation of one. In this way, we can interpret the estimates in terms of standard deviations at T=0. Formally, our measure of learning for subject *i* is written as

$$Learning_i = AST_{0i} - AST_{1i} \quad -[1]$$

where AST_{0i} and AST_{1i} are the *standardized* average solving time for the Ability Block (T=0) and Evaluation Block (T=1), respectively. Note that our measure of learning is calculated by subtracting the average solving time at T=1 from the average solving time at T=0, so that positive values indicate *improvement* in solving time. Using a standardized measure of learning also facilitates comparisons across environments in which there are differences in ex ante (baseline) proficiency. We also conducted a replication with an alternative task, and using the standardized learning measure facilitates a pooled analysis of both experiments in Appendix C.[22]

Another natural candidate for the outcome variable would be the change in the number of Sudoku puzzles solved from T=0 to T=1. However, due to the fixed length of our experiment, the observed variation in the change in the number of Sudoku puzzles solved is much smaller than that of changes in average solving time, making it difficult to use changes in the number of Sudoku solved to meaningfully identify the effects of treatments. In fact, the coefficients of variation (COV), which divide a variable's standard deviation by its mean, are 2.36 (=73.5/31.1) and 0.55 (=2.8/5.1) for changes in raw average solving time and changes in the number of

[21] Specifically, our software tracked the total time (in seconds) from initiation to completion spent on correctly solved Sudoku for each subject. Average solving time was recorded as this number of seconds divided by the number of correctly solved Sudoku.

[22] Because our learning measure is skewed to the right especially among subjects at the lower tail of the distribution, we also examine the change in logged learning as a robustness check in which we take the difference in the logs of (non-standardized) average solving time in the Ability Block (T=0) and the Evaluation Block (T=1). Estimates using the change in logged learning as our dependent variable can be interpreted in terms of percentage changes in average solving time.

Sudoku puzzles solved, respectively.[23] Moreover, using the number of Sudoku puzzles correctly solved as an outcome measure makes no distinction between subjects who barely finish N puzzles and those who run out of time just before correctly completing the N+1th puzzle, while also implying a difference in performance between subjects who barely complete N puzzles and those who were about to finish the Nth puzzle when time expired. Using average solving time for correctly solved problems allows us to distinguish between these subjects.

3.2. Summary Statistics and Balance Checks

Table 2-A reports summary statistics for the subject-level variables that we collected, while Table 2-B reports the results of balance tests examining whether our randomization of subjects to treatments was successful. Column (1) of Table 2-A presents the means of our control variables and the outcome variable. In total, 68 percent of the subjects had some prior experience with Sudoku. Raw average solving time in the Ability Block (AST_{0i}) is roughly two minutes (119 *sec*) per puzzle solved, while in the Evaluation Block (AST_{1i}) it is 88 *sec*, yielding average raw "learning" of 31 *sec*. To illustrate the learning by subjects, Figure 3 plots the relationship between standardized AST_{0i} and AST_{1i} for each subject. We confirm that most of the subjects indeed learn. The solid line indicates the 45-degree line, and thus subjects *below* the 45-degree line exhibit improvement in their average solving time. Out of 448 subjects, 371 subjects (84 percent) exhibit positive learning. In addition, only eight subjects (2 percent) could not solve any Sudoku at T=0.[24] Finally, the mean of our main outcome—learning measured in standard deviations of AST_{0i}—is 0.32.

---

[23] We observe more variation in our learning measure than in the change in the number of problems solved for a straightforward reason. Consider the improvement in average solving time necessary to produce a one-puzzle improvement in the number solved for subjects with differing initial performance. Going from solving two puzzles to solving three puzzles in 15 minutes requires a much larger improvement in average solving time in both absolute terms and as a proportion of initial average solving time than going from solving 12 puzzles to 13. In other words, to observe an equivalent change in the number of puzzles solved requires much more *learning* for subjects in the lower end of the ability distribution. This in turn makes it much more likely that learning (improvement) among low-ability subjects would go undetected using the change in number of problems solved.

[24] We impute the raw average solving time for the 8 subjects who could not solve any Sudoku puzzles in the Ability Block (T=0) to be 600 seconds (= 10 minutes), the length of Ability Block. Only one subject also could not solve any Sudoku puzzles in the Evaluation Block (T=1) as well. We cap the raw average solving time for this subject at 600 seconds, so that learning is equal to zero. We have also assigned this subject a value of 900 (= 15 minutes) for his/her raw average solving time, the length of Evaluation Block, so that raw learning is –300, but the estimates are almost identical given that there is only one such subject. As a robustness check, we exclude the 8 subjects who could not solve any Sudoku puzzles at T=0 from the sample and find quantitatively the same results.

As we are especially interested in the heterogeneity of learning by initial performance, Columns (2) and (3) report the means of the control and outcome variables for subjects in the top half and bottom half of their session's ability distribution (as defined by performance in the Ability Block). Column (4) presents the differences between these two columns. As expected, subjects in the top half had more prior experience with Sudoku than subjects in the bottom half (88 percent vs. 48 percent). In addition, the subjects in the top half were slightly less prosocial insofar as they give less in the dictator game than subjects in bottom half.

Importantly, the raw average solving time during the Ability Block (T=0) is much larger for subjects in the bottom half than those in the top half (168 vs. 71 *sec*), implying that there is more room for improvement among subjects in the bottom half. In fact, raw average learning ($=AST_0 - AST_1$) is much larger for subjects in the bottom half than those in the top half (53 vs. 9 *sec*). Also, the minimal raw learning (9 *sec*) by subjects in the top half suggests that high-ability subjects achieve near-peak performance even during the Ability Block (T=0), which may limit the scope for *any* treatment to affect their performance. We investigate this issue in Section 5 using a game less familiar to subjects than Sudoku called Nonograms. Interestingly, while subjects in the bottom half "learn" much more than subjects in the top half, the raw average solving time for the bottom half during the Evaluation Block (AST_1) is still larger than the raw average solving time for the top half during the Ability Block (AST_0), suggesting that the subjects in the bottom half could only close about half of the initial performance gap. The bottom line is that we expect to see heterogeneous treatment effects across the ability distribution given that subjects in the top half already perform at a very high level at T=0, while subjects in the bottom half have far more room for improvement.

Table 2-B reports the results of two types of balancing tests. Columns (1) and (2) report estimates from bivariate regressions that test how each variable in the far-left column is related to the *teaching* treatment (Column 1) and the *tracking* treatment (Column 2). Only one out of the 14 estimates is statistically significant at the 10 percent level, suggesting that the random assignment of subjects to treatments was successful.

As an alternative test of random assignment, Columns (3) and (4) reports p-values for F-tests of the null hypothesis that the means across treatments are equal. Column (3) tests the null that the means across the eight treatments (2×2×2) are equal. Each cell reports the p-value for a separate test for each variable in the far-left column. We fail to reject the null hypotheses of no

differences across treatments for all of our controls.[25] As mentioned earlier, we pool data across incentive schemes in our main analysis and focus on teaching (*no-teaching/teaching*), tracking (*tracked/untracked*) and their interaction, which essentially reduces our design to four treatments (2×2). Therefore, Column (4) tests the null that the means across these four treatments are equal. While we still fail to reject the null hypothesis for each variable, some of the p-values are less than 0.20. Furthermore, the potential lack of balance in covariates becomes more relevant when we replicate our design using another game due to the small sample in that robustness exercise (Section 5). As such, in much of the analysis we control for these subject characteristics.

4. Main Results

As noted above, our incentive treatments had no perceptible impact on behavior. Thus our main analysis pools the data over the incentive schemes and focuses on the impact of peer-to-peer teaching on learning, as well as the interaction of teaching with ability tracking.

4.1. Effects of the *Teaching* Treatment on Learning

To test whether the peer-to-peer teaching has any positive impact on learning, we first simply compare the treatment means while controlling for individual characteristics. Because the assignment of subjects to each treatment is random, the estimation equation is straightforward:

$$Learning_i = \alpha + \beta Teaching_i + \gamma X_i' + \varepsilon_i \quad -[2]$$

where $Teaching_i$ is a dummy equal to one for subjects in the *teaching* treatment and zero otherwise. β is our coefficient of interest.[26] The inclusion of individual controls X_i' is, in principle, not necessary for estimation given that random assignment to treatment appears to have been successful as shown above, but we nonetheless include them to gain efficiency. Specifically, our controls are a dummy for male, a dummy for prior experience, the number of risky choices made in the risk preference elicitation, the amount offered in the dictator game, and a dummy for each of the eight subjects who could not solve any Sudoku puzzles during the Ability Block (T=0). For all of the regression results that we report, the standard errors are clustered at the group level, where each session consists of two groups, because all interaction

[25] We also performed Kruskal-Wallis equality-of-populations tests as checks of our randomization and obtained similar results (results available upon request). We do not report them here because that test assumes the variables are measured on an ordinal or continuous scale, an assumption which does not apply for our binary variables.
[26] Note that any measured effect of the *teaching* treatment on performance is actually an intent-to-treat effect because some groups may not (and in fact did not) do much teaching.

among subjects (chatting in the *teaching* treatment and the revelation of subjects' performance and rank information in all treatments) occurs at the group level.

To evaluate whether learning is heterogeneous across the ability distribution, we estimate:

$$Learning_i = \alpha_0 + \beta_1 Teaching_i + \beta_2 Bottom_i + \beta_3 Teaching_i \times Bottom_i + \gamma X_i' + \varepsilon_i \quad \text{–[3]}$$

where $Bottom$ is a dummy equal to one for subjects in the bottom half of their session's ability distribution. The bottom half consists of the subjects who ranked 5–8 out of eight subjects in the Ability Block (T=0), while the top half consists of the subjects ranked 1–4. Because our interest is in the heterogeneous treatment effects by initial ability (i.e., which parts of the ability distribution are affected by peer-to-peer teaching) instead of the *difference* in treatment effects by initial ability, we report the treatment effects separately for subjects in top half and the bottom half using the outputs of estimating equation [3]. Specifically, β_1 captures the effect of peer-to-peer teaching on subjects in the top half, while the sum of β_1 and β_3 captures the effect of peer-to-peer teaching on subjects in the bottom half.

Testing Hypothesis 1 (Main Effect of *Teaching*):

According to Hypothesis 1, teaching should have a positive impact on learning as subjects help each other figure out how to solve Sudoku puzzles. This effect might be larger for subjects in the bottom half of the ability distribution because they have more scope for improvement.

Finding 1: The peer-to-peer *teaching* treatment significantly increases learning.

Evidence: Table 3 summarizes the relevant statistics from the outputs of estimating equations [2] and [3]; the coefficient estimates themselves are reported in Appendix Table B1.[27] Note that all of the regressions include a dummy for the eight subjects who could not solve any Sudoku puzzles in the Ability Block (T=0)—even those labeled as including "no" controls. Columns (1) and (2) in Table 3 are based on the outputs from estimating equation [2]. Column (1) shows that teaching improves learning by 0.11 SD (p-value<0.05). The additional controls in Column (2) barely affect the estimate (0.12 SD), reconfirming that randomization was successful.[28] This improvement by 0.12 SD corresponds to a reduction in raw average solving

[27] As expected, the coefficient of experience is negative—suggesting that subjects with experience solving Sudoku puzzles had little (or less) room for improvement. Gender, risk attitudes and prosociality have no impact on learning. Once we add the dummy indicating whether subjects were in the bottom half (equation [3]), the coefficient of experience is substantially reduced and no longer statistically significant due to the high negative correlation between the dummies for being experienced and being in the bottom half. Finally, the large estimate for the constant term indicates that subjects learn even *without* peer-to-peer teaching.

[28] Note that this is an average treatment effect pooling the *untracked* and *tracked* treatments given that both types of environments (tracked and untracked) exist in practice. The estimated treatment effects of teaching can be separately derived for the *tracked* and *untracked* treatments from the estimates in Appendix Table B2.

time of 11.6 *sec*. Given that the mean learning in the no-teaching treatment is 27.6 *sec*, this translates into a 42 percent increase in learning.

Importantly, the average effects reported in Columns (1) and (2) may not capture differences in the distributions of learning across treatments. Figure 4-A presents the kernel densities of learning for the *no-teaching* (solid line) and *teaching* (dashed line) treatments. The two-sample Kolmogorov-Smirnov test for equality of the distributions yields a p-value of 0.01, suggesting that the two distributions are quite different. This is further evidence that peer-to-peer teaching substantially reduces the average solving time.

It is important to reiterate that although we did not explicitly ask the subjects to teach other or give them incentives to do so, we did suggest that the Practice Block could be used to work together on a puzzle. Subjects simultaneously edited the same puzzle on the screen, were equipped with headphones, and were allowed to chat for 10 minutes, so the design directly encourages teaching by (and learning from) peers. Indeed, as we show later in the analysis of the audio recordings (Section 4.3), substantial peer-to-peer teaching occurred during the Practice Block (with considerable variation across sessions). Further, it is remarkable that being given only 10 minutes to work *together* on a single Sudoku puzzle as opposed to working on it alone (as in the *no-teaching* treatment) increases learning by 0.12 SD or 42 percent. We are not aware of any past studies in the economics of education that directly document the importance of peer-to-peer teaching to learning.[29]

A significant caveat, however, applies to our estimates. The estimated effect of peer-to-peer teaching on learning in our experiment is non-trivial relative to other treatment effects in the education literature, which likely reflects the scope for basic instruction from peers to resolve confusion and improve performance. In actual classrooms, however, teachers likely provide much of this instruction, and as such the effect of peer-to-peer teaching may smaller in practice than it is in our experiment. Of course, the potential for substitution between instruction from peers and instruction from teachers also implies that peers may attenuate the negative effects of

[29] Li et al. (2014) find that low-performing students experience large gains between achievement tests when seated next to a high-performing peer who was being paid for improvements in the low-performing student's test scores in Chinese middle schools. Significantly, the gains experienced by low-performing students were much larger than when seated next to a high-performing student who was not being paid for improvements in the low-performing student's score. Their experiment, however, does not identify how the high-performing students influence the low-performing students' scores. Nonetheless, their findings suggest that even larger treatment effects may have emerged had we provided subjects with group incentives.

"bad" teachers. In any case, the large experimental estimate of the effect of peer-to-peer teaching highlights the need for further investigation of peer-to-peer teaching in other contexts.

Finding 2: The positive effect of peer-to-peer teaching is primarily on those individuals in the bottom half of the ability distribution.

Evidence: Columns (3) and (4) in Table 3 present the estimated treatment effects of peer-to-peer teaching for subjects in the top and bottom halves, respectively, based on the outputs from estimating equation [3]. Columns (3) and (4) show that the positive effect of teaching on learning is driven entirely by subjects in the bottom half. In Column (4), which also controls for individual characteristics, the estimated teaching effect for subjects in top half is very small (–0.02 SD) and far from statistically significant. On the other hand, the estimated teaching effect for subjects in the bottom half is 0.24 SD (p-value<0.01).[30]

Figure 4-B presents kernel densities of the learning distribution for the *no-teaching* and *teaching* treatments in which we restrict the sample to subjects in the bottom half. The figure clearly shows that the distribution of learning is shifted to the right in the *teaching* treatment compared to the *no-teaching* treatment.

4.2. Ability Tracking and Peer-to-peer Teaching

In the previous subsection, we established that peer-to-peer teaching encourages learning in our experimental setting, though the effect of teaching is concentrated among subjects in the bottom half of the ability distribution who have more scope for improvement. In this section, we examine how ability tracking interacts with peer-to-peer teaching and hence learning.

Testing Hypothesis 2 (Effect of *Tracked* under *No-teaching*)

As a benchmark, we estimate the direct effect of tracking in the *absence* of teaching. High-performing subjects may be encouraged when learning their relative ranking while low-performing subjects may be discouraged.[31] In addition, learning about the performance of other

[30] To assess the robustness of our estimates to the influence of skewness, we estimate the same equations as [2] and [3] but replace the outcome by the difference in the logs of (non-standardized) average solving time in the Ability Block (T=0) and the Evaluation Block (T=1). Appendix Table B3 shows that the general message is the same as our baseline estimates in Table 3: the large gain in learning in the peer-to-peer teaching treatment is concentrated among subjects in the bottom half of the ability distribution.

[31] Note that subjects were informed of their relative rank in all treatments within their matched group of 4 (see screenshots in the Appendix E), but we did not tell them their overall rank among the 8 subjects of their session. Perhaps it was easier for low-performers to infer their overall rank due to the zero lower bound on performance. Our results are robust to including indicators for within-group rank.

subjects in their group may increase the pressure felt by subjects and affect their performance. The potential for these offsetting effects makes the effect of tracking ambiguous.[32] We estimate

$$Learning_i = \alpha_0 + \beta_1 Tracked_i + \beta_2 Teaching_i + \beta_3 Tracked_i \times Teaching_i + \gamma X_i' + \varepsilon_i \text{ –[4]}$$

where the reference group is the *untracked* and *no-teaching* treatment. β_1 is the effect of tracking in the *no-teaching* treatment (Hypothesis 2), while the sum of β_1 and β_3 captures the effect of tracking in the *teaching* treatment (Hypothesis 3 as shown later).

To examine heterogeneous effects by initial ability, we estimate

$$Learning_i = \alpha_0 + \beta_1 Tracked_i + \beta_2 Teaching_i + \beta_3 Bottom_i + \beta_4 Tracked_i \times Teaching_i + \beta_5 Tracked_i \times Bottom_i + \beta_6 Teaching_i \times Bottom_i + \beta_7 Tracked_i \times Teaching_i \times Bottom_i + \gamma X_i' + \varepsilon_i \text{ –[5]}$$

which adds to equation [4] a dummy indicating subjects ranked in the bottom half of their session in the Ability Block and interactions between this dummy and the treatment indicators. In the *no-teaching* treatment, β_1 captures the effect of tracking for subjects in the top half and the sum of β_1 and β_5 captures the effect of tracking for subjects in the bottom half. We report the coefficient estimates themselves from equations [4] and [5] in the Appendix Table B2.

Finding 3: We observe no significant impact of ability *tracking* in the absence of teaching.

Evidence: Column (1) in Table 4 shows that tracking reduces learning by 0.04 SD on average in the absence of teaching, but the estimate is far from statistically significant and small in magnitude. Figure 5-A plots the kernel densities of learning in the *no-teaching* × *untracked* and *no-teaching* × *tracked* treatments. The p-value of a Kolmogorov-Smirnov test between the *untracked* and *tracked* treatments is 0.56, and thus we cannot reject the null that the two distributions are the same.

This overall null result, however, might mask heterogeneity among subjects. Column (2) of Table 4 shows that ability tracking may reduce the improvement in average solving time by 0.095 SD among the subjects in the bottom half, which is sizable and consistent with a discouragement effect from learning their relative ranks but far from statistically significant at conventional levels. We conclude that we do not find a direct negative effect of tracking *per se*.

Testing Hypothesis 3 (Effect of *Tracked* under *Teaching*)

[32] In education, Murphy and Weinhardt [2014] and Elsner and Isphording [2017] find that primary and secondary school rank has large effects on subsequent academic outcomes even after controlling for ability. They attribute their findings to the development of confidence and to the formation of expectations and perceptions about ability, respectively. These long-term effects of rank information are unlikely to be important in our short experiment.

On the one hand, subjects in the *teaching* treatment who are *tracked* may have less to teach one another as the difference between the best and worst performer in a given group is smaller on average than in the *untracked* treatment. In fact, the standard deviations of raw average solving time in the Ability Block (AST_0) at the group level are 0.85 and 0.48 SD (82.6 and 46.1 *sec*) in the *untracked* and *tracked* sessions, respectively.[33] In addition, tracking may especially hurt subjects in the bottom half as they lose access to high-ability peers who could have taught them in the *untracked* sessions. In this sense, ability tracking may attenuate the positive effects of teaching. On the other hand, subjects of more similar ability may be more effective in teaching one another. The potential for offsetting effects makes the effect of ability tracking in the teaching treatment ambiguous. Columns (3) and (4) in Table 4 present the estimated treatment effects of tracking in the *teaching* treatment from equations [4] and [5].

Finding 4: *Tracked* groups exhibit less learning than *untracked* groups in the *teaching* treatment.

Evidence: Figure 5-B plots the kernel densities of learning in the *teaching* × *untracked* and *teaching* × *tracked* treatments. The figure shows that distribution of learning in the *tracked* treatment is shifted to the left compared to the *untracked* treatment, suggesting that *tracked* group exhibit less learning than *untracked* groups in the *teaching* treatment. A two-sample Kolmogorov-Smirnov test yields a p-value of 0.015. Another way to visualize this shift is presented in Figure 6, which displays the empirical CDFs of learning by treatment. When peer-to-peer teaching is allowed, Figure 6-B shows that learning in the *tracked* treatment is stochastically dominated by learning in the *untracked* treatment.

Column (3) in Table 4 shows that *tracking* reduces learning by 0.15 SD (p-value<0.10) on average in the *teaching* treatment, substantially offsetting the positive effects of teaching (0.17 SD in the *untracked* treatment from Appendix Table B2). This offsetting effect of tracking is driven mainly by subjects in the bottom half—the subjects who benefitted most from teaching in the first place—while having little effect on subjects in the top half.[34] For subjects in the bottom half, the estimates in Column (4) indicate that as *tracking* reduces learning by as much as 0.28

[33] In the *teaching* treatment, there are 32 groups each in the *untracked* and *tracked* sessions.

[34] The sum of β_1 and β_4 from equation [5] captures the effect of *tracking* on subjects in the top half in the *teaching* treatment, while the sum of β_1, β_4, β_5, and β_7 captures the effect on the subjects in the bottom half. The original estimates from equation [5] are reported in Appendix Table B2.

SD (p-value<0.10) relative to the estimated effect of teaching on learning among these subjects of 0.33 SD in the *untracked* treatment.[35,36]

In summary, while *tracking* does not have large negative effects on learning for subjects in the bottom half without teaching, it has large detrimental effects on them in the *teaching* treatment—probably because they lose access to high-ability peers who could have taught them. In Section 4.3, we examine the frequency of teaching to test this conjecture.

4.3. Mechanism for the Negative Impact of *Tracking* under *Teaching*

So far, we have documented that the positive effect of *teaching* on learning is offset by a negative effect of *tracking* on subjects in the bottom half of the ability distribution. In this section, we investigate the mechanism underlying this finding by analyzing actual instances of peer-to-peer teaching behavior in our experiments.

Specifically, our software recorded subjects' conversations with members of their group during the Practice Block. Two research assistants (who were unaware of our research question and the particulars of our experimental design) transcribed these conversations and then *independently* counted the number of teaching related statements and the number of non-teaching related statements for each group. A teaching statement is defined to be any utterance in which subjects are engaged in trying to teach each other how to do Sudoku such as "You can't have a five there; there is already one in that column." After each research assistant counted instances of teaching in each group independently, the two research assistants cross-checked their counts and resolved the few disagreements.[37]

Figure 7 plots the frequency of "teaching related statements" (hereafter just "teaching") at the group level. The graph on the left plots the number of groups with given teaching frequencies in the *untracked* treatment (N=32) as well as the same distribution for groups composed of the

---

[35] Appendix Figure B1 reproduces the empirical CDFs in Figure 6-B for subjects in the top and bottom halves separately, recognizing the risk of splitting the sample on too many dimensions. The figure indicates that only subjects in the bottom half are negatively affected by tracking in the *teaching* treatment.

[36] Appendix Table B4 presents estimated treatment effects analogous to those presented in Table 4 using log learning as the dependent variable. Our findings that *tracking* negatively impacts learning only in the *teaching* treatment and only for subjects in the bottom half remain robust. Column (4) shows that in the *teaching* treatment, *tracking* reduces learning among the subjects in the bottom half by 15.2 percent (p-value<0.05), while tracking has a negligible impact on the subjects in the top half.

[37] See Appendix F for a description of the scheme used by the research assistants to categorize statements as "teaching."

bottom half of Ability Block performers in the *tracked* treatment (N=16). The graph on the right plots (again) the number of groups with given teaching frequencies in the *untracked* treatment (N=32) along with the same distribution for groups composed of the top half of Ability Block performers in the *tracked* treatment (N=16).

Table 5 reports the results of several regression specifications in which the outcome is the number of teaching statements exchanged by a group. The hypothesis is that the number of teaching statements will be consistent with the results for learning in Columns (3) and (4) of Table 4: namely that subjects in the bottom half experienced less peer-to-peer teaching in the *tracked* treatment than in the *untracked* treatment.

Finding 5: Ability tracking reduces the frequency of instances of peer-to-peer teaching.

Evidence: Table 5 presents our analysis of teaching frequency, which is by construction limited to 32 *teaching* sessions with 16 sessions each in the *untracked* and *tracked* treatments. The unit of analysis here is a group, and in each session there are two groups. Column (1) shows that on average we observe 4.8 instances of teaching per group in the *untracked* treatment. Columns (2) and (3) present the means for the *tracked* groups containing the bottom half of subjects in the Ability Block (ranks 5–8) and the top half (ranks 1–4), respectively.

Columns (4) and (5) of Table 5 report the differences between Columns (1) and (2) estimated via different econometric models. Column (4) presents the estimates from OLS, and Column (5) estimates from a zero-inflated Poisson model in order to account for both the discrete nature of teaching frequencies and the fact that there are a number of groups without any teaching; the Vuong test of the zero-inflated versus the standard Poisson reported in the table indicates that the zero-inflated models are preferred in all specifications in Table 5. Columns (6) and (7) report corresponding estimates for the difference between Columns (1) and (3).

This analysis provides evidence that ability tracking reduces the number of teaching statements in both the high- and low-ability groups. The reduction in teaching frequency due to ability tracking is much more substantial among subjects in the top half (Columns (6) and (7) of Table 5) than the reduction among subjects in the bottom half (Columns (4) and (5) of Table 5). We believe this is because high-ability subjects have no one to teach under ability tracking when they are surrounded by other high ability subjects who already know how to do Sudoku and do

not need to be taught by others. The histograms of teaching frequency by treatment in Figure 7 provide further evidence to this effect.[38]

To summarize, the *tracked* treatment, which assortatively groups subjects based on ability, reduces the frequency of peer-to-peer teaching compared to the *untracked* treatment. While we observe a reduction in the instances of teaching among subjects in both the top and bottom halves of the ability distribution, the reduction is much larger among subjects in the top half. The reduction in teaching frequency among subjects in the top half may reflect the fact that high ability subjects have little to teach each other. Subjects in the bottom half, however, still try to teach each other, but their peer-to-peer teaching is evidently ineffective.[39]

5. Replication Exercise Using a Different Game: Nonograms

So far, we have shown that most learning in our experiment occurs among low-ability subjects, while high-ability subjects are hardly affected by any treatment combination. The positive effect of *teaching* on learning is concentrated among subjects in the bottom half of the ability distribution, and the negative effect of *tracking* in the *teaching* treatment is also concentrated on the subjects in the bottom half. One possible reason for the absence of treatment effects among subjects in the top half is a "ceiling effect." High-ability subjects may already achieve near-peak performance in solving Sudoku puzzles even during the baseline (T=0) and

---

[38] One potential explanation for tracking's effect on teaching frequencies is that tracking may reduce the variance of initial ability within a group. On the one hand, a more homogenous group may facilitate teaching if subjects of similar ability find it easier to express their difficulties to one another. On the other hand, it is also possible that some heterogeneity is necessary to generate a meaningful exchange of information in the form of questions and (correct) answers. Appendix Table B5 correlates the number of teaching statements with the group mean and group standard deviation (SD) of standardized average solving time at T=0. Because group mean is the group average of standardized average solving time, the higher the group mean is the *worse* the group's performance in the Ability Block. To account for the discrete nature of teaching frequencies, we report the results from a zero-inflated Poisson model in Columns (1)–(4); the Vuong test reported in the table indicates that the zero-inflated model is preferred to the standard Poisson in all specifications. Throughout Columns (2)–(4), the most robust result is that the coefficient estimate on group SD is positive and statistically significant. Notably, Column (3), including both group mean and group SD, indicates that greater group heterogeneity is associated with more peer-to-peer teaching even when comparing groups with subjects of similar ability on average. Column (4) shows that, while imprecisely estimated, the coefficient estimate of the interaction between group mean and group SD is negative—suggesting that the positive effect of group SD on peer-to-peer teaching is larger in groups with better performing subjects on average (i.e., a lower group mean) than in groups with worse performing subjects. This result seems plausible if the positive effect of ability heterogeneity on teaching frequency is mitigated when a group consists of lower ability subjects with less to teach each other.

[39] That students in the bottom half of the ability distribution need higher ability students to teach them is consistent with Lavy, Silva, and Weinhardt's [2012] finding that girls (although not boys) in the bottom half of the ability distribution benefit from the presence of very bright peers.

thus have no scope for improvement no matter what the treatment. Indeed, mean raw learning by subjects in the top half is only 9 *sec* per puzzle (vs. 53 *sec* per puzzle in the bottom half).

To investigate this claim, we replicate our experiment substituting a less popular logical puzzle called a *Nonogram*. Nonograms are similar to Sudoku in the sense that subjects need to fill a 5×5 grid while satisfying a set of logical constraints as shown in Appendix Figure B2 (and detailed in the instructions in Appendix D). Moreover, there are a numerous puzzle solving "strategies" for Nonograms that are also straightforward to teach and learn, and an instructive video also exists as in the case of Sudoku.[40] In fact, this game has been used in another study as an alternative to Sudoku (Charness et al. [2015]). The most important difference between Sudoku and Nonograms is that most of the subjects in our experiment have no prior experience solving Nonograms: unlike Sudoku with which 68 percent of subjects have some experience, only 2 percent of our subjects have some experience with Nonograms.

Table 6 summarizes the design of Nonograms experiments. We conducted 4 sessions for each combination of *tracking* × *incentives* in the *no-teaching* treatment (16 sessions) and the *teaching* treatment (16 sessions) with a total of 256 subjects (participants were restricted to those who had not previously participated in the Sudoku experiment). As a result, we often lack the statistical power to precisely detect treatment effects, and thus we view the results reported below as only complementary to our main findings using Sudoku. In fact, none of the Kolmogorov-Smirnov tests for equality of the distributions are statistically significant even though it is visually apparent that the distributions of learning are different. The procedures for the Nonogram experiments were identical to those for Sudoku as described in Table 1-A.

Appendix Table B6 presents summary statistics from the Nonogram experiments. The mean of our main outcome—learning measured in standard deviations of AST_0—is 0.48. Out of 256 subjects, 234 (91.4 percent) exhibited positive learning—in part because 20 percent of subjects could not solve any Nonograms at T=0. Thus unlike in the Sudoku experiments, even subjects in the top half may have some scope for improvement. The p-values for tests of the null hypothesis that the covariates are balanced across four treatments (2×2) (we again pool sessions from different incentive schemes) for several variables are just above conventional significance levels suggesting that some variables are not perfectly balanced across treatments due to the

[40] A link to the video is available here: https://www.dropbox.com/s/vpsti0kpsepp0kh/NonogramTutorial.mp4?dl=0

small sample.[41] Thus, we estimate equations [4] and [5] using the data from the Nonograms sessions controlling for subject characteristics as we did for the Sudoku sessions.

We first examine whether teaching has a positive effect on learning in Nonograms. Figure 8-A presents the cumulative distributions of learning for the *no-teaching* (solid line) and *teaching* (dashed line) treatments, while Figure 8-B focuses on subjects in the bottom half. These figures show that subjects' learning in the *teaching* treatment stochastically dominates subjects' learning in the *no-teaching* treatment—especially among subjects in the bottom half.

Table 7, which corresponds to Table 3 for Sudoku, confirms this visual inspection.[42] Column (2) shows that teaching increases subjects' learning by 0.20 SD (p-value<0.01), which is larger than in Sudoku (0.12 SD). This estimate suggests that having subjects with a high degree of ex ante proficiency as in the Sudoku experiments is not a necessary condition for successful peer-to-peer teaching. Column (4) shows that most of the gains from *teaching* are again concentrated among subjects in the bottom half of their session (0.36 SD with p-value<0.01). Teaching does not seem to positively impact learning for subjects in the top half despite their lack of familiarity with Nonograms (0.02 with p-value of 0.29). Given that subjects in the top half improve by almost 27 *sec* on average as reported in Appendix Table B6 in the Nonogram experiments compared to 9 *sec* for top half subjects in the Sudoku experiments, "ceiling effects" are less likely to be an issue in the Nonogram experiments.[43,44] As such, we conjecture that subjects in the top half may not be exposed to sufficiently many higher ability subjects who have something to teach them. In addition, the absence of improvements in performance among subjects in the top half suggests that the effects of *teaching* in the Sudoku experiments were unlikely to have been driven by peers motivating each other during the Practice Block. In the Nonogram experiments, top half subjects had room for improvement and would presumably be

---

[41] Again, our tournament incentive scheme had a negligible effect on behavior (either through main effects of the treatment or interactions). The means of learning in the *piece-rate* and *tournament* treatments are very similar (0.479 vs. 0.484 SD), and the p-value for the test of the equality of means is 0.956 (see Appendix Table A1).

[42] The original estimates from equation [5] are reported in Appendix Table B7.

[43] The only notable difference between Sudoku and Nonograms is that the estimate on "bottom half" for Sudoku in Column (4) of Appendix Table B1 is statistically significant and positive (0.17) while that for Nonograms in Column (4) of Appendix Table B7 is very small (–0.01). The Sudoku estimate suggests that subjects in the top half did not have room for improvement, and thus learning among these subjects is substantially less than the learning among subjects in the bottom half (ceiling effects). By contrast, the Nonogram estimate for subjects in the bottom half is close to zero, suggesting that subjects in the top and the bottom halves improve by similar amounts. Interestingly, peer-to-peer teaching does not *additionally* improve learning among subjects in the top half in Nonograms despite the fact that they have room for improvement.

[44] Following standardization, the gain for subjects in the top half is small compared to subjects in the bottom half, but these gains are large relative to initial solve times among top half subjects.

susceptible to positive motivational effects from the *teaching* treatment, but we observe no such improvement.

Finally, we examine how ability tracking interacts with peer-to-peer teaching and hence learning. We first examine the effect of *tracking* in the *no-teaching* environment to see if tracking *per se* has any discouragement or encouragement effects when subjects learn their ranks within the group. Figure 9-A plots the empirical CDFs of learning in the *no-teaching* × *untracked* and *no-teaching* × *tracked* treatments. The distributions are almost identical suggesting that tracking per se does not seem to affect learning as in the Sudoku experiments.

We now turn to the main effect of interest to us, the effect of *tracking* when peer-to-peer *teaching* is possible. Figure 9-B plots the empirical CDFs of learning in the *teaching* × *untracked* and *teaching* × *tracked* treatments. The figure shows that learning in the *tracked* treatment is nearly stochastically dominated by learning in the *untracked* treatment, suggesting that tracking when *teaching* is possible diminishes learning. This result is consistent with that in the Sudoku experiments.

Table 8 summarizes the estimates based on the outputs of equations [4] and [5] to formalize the inference from the visual inspection of Figure 9.[45] Columns (1) and (2) show that ability tracking had no significant impact on learning in the *no-teaching* treatment regardless of whether we consider the full sample or subjects in the top and bottom halves of their sessions. Unlike in the Sudoku experiment, the effect of *tracking* in the *no-teaching* treatment is positive in all of these samples, but the estimates are far from statistically significant.

Column (3) in Table 8 shows that ability tracking reduces learning by 0.063 SD on average in the *teaching* treatment, but this effect is not statistically significant. Column (4), however, shows that ability tracking reduces the average solving time by 0.063 SD (p-value<0.05) among the subjects in *top* half. For subjects in the bottom half, the estimated treatment effect is similar in magnitude (–0.081) but imprecisely estimated. Thus we infer that *tracking* may reduce the learning of *all* subjects in the *teaching* treatment of the Nonogram experiment.[46]

That tracking has a negative effect on learning for subjects in the top half is consistent with evidence from the education literature that students who teach their peers learn more as a result

[45] The coefficient estimates themselves from equations [4] and [5] are reported in Appendix Table B8.
[46] Appendix Figure B3 reproduces Figure 9-B separately for subjects in the top and bottom halves. Despite the small sample, the figure clearly shows that both groups of subjects are negatively affected by *tracking* in the *teaching* environment in contrast to the corresponding figures for Sudoku in Appendix Figure B1 in which only subjects in the bottom half are negatively affected by *tracking* in the *teaching* treatment.

(Bargh and Schul [1980]). *Tracking* groups subjects with similar levels of understanding; as a consequence there may be fewer opportunities to engage in mutually beneficial teaching.[47] Our failure to observe a similar effect in the Sudoku treatment may have resulted from the high levels of Sudoku proficiency already evident in the Evaluation Block. Indeed, the potential for such "ceiling effects" is precisely why we ran the Nonogram experiments.

Exploring further this difference between the Sudoku and Nonogram experiments, Appendix Table B9 compares the number of "teaching related statements" in the *tracked* and *untracked* sessions. In contrast to the Sudoku experiment, we do not see meaningful differences in teaching intensity between the *tracked* and *untracked* treatments partly because of the small number of groups (we only have 32 data points).[48] Subjects in the bottom half—regardless of whether they were in the *tracked* or *untracked* sessions—may ask more questions about this unfamiliar game. As such, the increase in *teaching* in the *tracking* treatment in the bottom half group may simply reflect a compositional effect. Regardless of the explanation, the means of teaching *frequency* in Appendix Table B9 cannot rationalize the negative effects of tracking on all subjects in the Nonogram experiments. One possibility is that the frequency of teaching abstracts from the quality or nature of the exchanges. Top half students may benefit from having bottom half students asking basic questions (the answers to which benefit all subjects), while bottom half subjects may benefit from having better peers with answers to these questions.[49]

6. Discussion & Conclusion

Our study provides the first estimates of the importance of peer-to-peer teaching: enabling this interaction for only 10 minutes leads to a 42 percent increase in our measure of learning—an increase predominantly driven by low-ability subjects. While highlighting the potentially sizable effect of peer-to-peer teaching, our study also suggests that the effects of these interactions are shaped by the composition of peer groups. We see that the positive effect of peer-to-peer

[47] Song et al. [2017] similarly find that Chinese middle school students serving as tutors showed gains in achievement—even while the students being tutored enjoyed no achievement gains.

[48] For completeness, Columns (5)–(8) in Appendix Table B5 report the estimates from regressing teaching frequency on the group mean and SD of standardized average solving time at T=0. While group SD is always positive throughout the specifications as it was for Sudoku (except for Column (8)), the estimates are imprecise given the small sample size (N=30).

[49] Our use of a common, standardized learning measure also allows us to conduct an analysis of the pooled data from both the Sudoku and Nonogram treatments. Such an analysis maximizes the value of the replication exercise and provides increased statistical power—especially given that the results are similar across the treatments. In Appendix C, we report the pooled analysis, which confirms our main results.

teaching on low-ability subjects is substantially offset when subjects are tracked by ability. This implies that ability tracking based on prior achievement can potentially disadvantage low-ability students who may miss out on interactions with high-ability peers who can teach them. These insights into the role of peers as teachers come—like most estimates—with qualifying remarks.

Like field studies of tracking exploiting random assignment to peer groups, we cannot rule out the possibility that ability tracking may have little effect on low-ability students relative to an untracked setting in practice if students in untracked settings segregate themselves by ability on their own accord as documented in Carrell et al. [2013]. Furthermore, a laboratory experiment such as ours necessarily misses some important features of classrooms (e.g., the fact that most classes go on for weeks rather than hours, interactions among peers outside of the classroom).

Nevertheless, using a laboratory experiment allows us to estimate in a credible fashion the effect of peer-to-peer teaching—typically unmeasured in other settings. The virtue of the laboratory experiment is the extent of experimental control: our design allows us to exogenously vary both subjects' ability to teach other and peer group composition while also shutting down potential competing channels through which tracking may influence learning. Our laboratory setting allows us to observe the counterfactual world in which students are completely unable to teach one another. It is possible to observe peer-to-peer teaching in the field, but no (ethical) design can eliminate the possibility that students engage in peer-to-peer teaching outside the classroom—making a study such of ours documenting the significance of peer-to-peer teaching for learning essential to the peer effects literature.

Assessing the relevance of our findings for actual classrooms and academic disciplines, we note that learning Sudoku consists of learning rules and when to apply these rules. Likewise, learning addition, subtraction, calculus and even many languages similarly involves learning rules and when to apply them. Students stumbling over basic rules may benefit enormously from simple clarification to eliminate confusion as these basic rules form the building blocks for more challenging concepts. In our experiment, we suspect the large treatment effect from peer-to-peer teaching results from precisely this sort of elimination of confusion over basic rules among subjects with little familiarity with Sudoku. In classrooms, peers may have daily opportunities to resolve similar confusion when learning rule-based subjects, and thus we argue that the potential for significant learning gains from peer-to-peer teaching exists in many classrooms.

Differences in contexts, measures of learning, and timeframes make comparisons between our study and others in the peer effects literature challenging, but two recent studies bear special mention as they provide additional insight into how our findings are likely to apply in other

settings. Booij et al. [2017] and Feld and Zolitz [2017] exploit random assignment to classroom peer groups in Dutch universities and find that low-achieving students benefit from being placed in groups with students of similar ability—the opposite of what our findings might suggest.

Importantly, both studies also examine students' responses to surveys to explore the channels through which peers influence each other. Feld and Zolitz [2017] find that having peers with higher average GPA is positively related to "group interaction" as measured by agreement with the statements "My tutorial group has functioned well" and "Working in tutorial groups with my fellow-students helped me to better understand the subject matters of this course." Certainly the positive correlation between the latter statement and peer GPA is consistent with our inference that higher ability peers make better teachers for low-ability subjects.[50]

Booij et al. [2017] find that while students benefit from better peers, low-ability students are negatively affected by peer heterogeneity. The dominance of the latter effect makes tracking students by ability optimal. Using student surveys, they show that tracking positively affects interactions (i.e., whether students study together or help each other) and involvement (i.e., whether students ask questions in the tutorial and whether classmates' ability has an effect on students' motivation) in the tutorial group. These effects of tracking, however, are surely absent in our design as subjects neither study together outside of the lab nor work together on homework. Likewise, inspection of the chat transcripts provides scant evidence of subjects motivating each other in our experiment. Booij et al. [2017] and Feld and Zolitz [2017] examine peer groups that were central to students' studies over an extended period in which interactions outside of the classroom (e.g., studying together, homework collaboration) and the motivational effects of peers are likely to be far more important than in our experiment. By contrast, peers in our experiment influence each other exclusively through peer-to-peer teaching.

This should not be taken to imply that our findings are unlikely to generalize as the positive social and motivational effects of tracking identified in Booij et al. [2017] may be less likely to dominate the negative effects of tracking on peer-to-peer teaching in other contexts. In many North American universities, students are exposed to more peer groups as they take, on average, more courses per year and are unlikely to take the same courses with a single group of

[50] Feld and Zolitz [2017] find that low-ability students benefit from better peers as our findings would suggest, but they infer that low ability students benefit from being in groups with students of more similar ability because they find a negative coefficient on the fraction of one's peers in the top-third of the ability distribution. Increasing the fraction of peers in the top-third comes at the expense of students in the middle third—students more similar to students in the bottom third of the ability distribution.

students. As such, the peer interactions outside of the classroom and motivational effects of any given peer group are likely to be much smaller, and the effects of direct instruction from peers in the classroom may be more significant than in these Dutch universities. Indeed, we would suggest that future research on the role of peers as teachers should focus on how to structure classrooms and student interactions so as to maximize the benefits from peer-to-peer teaching.

To close, laboratory experiments such as ours—in spite of their limitations—have a role to play as "mechanism experiments" (Ludwig et al. [2011]) to investigate basic but fundamental issues such as the effects of peers. Given that lab experiments are smaller and less expensive than field experiments, such studies can be more easily replicated and the robustness of findings to differences in context and experimental design tested. Indeed, effects of the magnitude we report are so large as to demand replication and interest from education researchers.

Furthermore, we view laboratory experiments as a natural complement to more burdensome and potentially disruptive field experiments—perhaps as a precursor to guide and inform the design of such interventions. For example, anecdotally it has been suggested to the authors that tracking is as prevalent *within* classrooms as it is *across* classrooms with teachers matching students for group work. The laboratory could be used to investigate whether "nearest neighbor" matching rules assigning similar students to work together lead to better outcomes than alternative assignment rules. Alternatively, Carrell et al. [2013] speculate that having middle-ability students in a classroom may be important for low-ability students if middle-ability students serve as mediators or bridges between low and high-ability students. The laboratory could be used to investigate the importance of peers who can serve as "bridges" between groups of students with different abilities. Experiments could also shed light on whether peer ability is a complement or substitute in the education production function. We leave intriguing questions such as these for future research.

References

Andreoni, James, and Andy Brownback. 2017. "All-Pay Auctions and Group Size: Grading on a Curve and Other Applications." *Journal of Economic Behavior & Organization*, 137: 361–373.

Ammermueller, Andreas, and Jorn-Steffen Pischke. 2009. "Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study." *Journal of Labor Economics*, 27(3): 315–348.

Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg and Jure Leskovec. 2014. "Engaging with Massive Online Courses." In *Proceedings of the 23rd international conference on World wide web*, 687–698.

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

Barankay, Iwan. 2012. "Rank Incentives Evidence from a Randomized Workplace Experiment." *Unpublished manuscript.*

Bargh, John A., and Yaacov Schul. 1980. "On the cognitive benefits of teaching." *Journal of Educational Psychology*, 72: 593–604.

Bettinger, Eric P., Lindsay Fox, Susanna Loeb, and Eric S. Taylor. 2017. "Virtual Classrooms: How Online College Courses Affect Student Success." *American Economic Review*, 107(9): 2855–2875.

Bettinger, Eric, Jing Liu, and Susanna Loeb. 2016. "Connections Matter: How Interactive Peers Affect Students in Online College Courses." *Journal of Policy Analysis and Management*, 35(4): 932–954.

Betts, Julian R. 2011. "The Economics of Tracking in Education." in *Handbook of the Economics of Education*, Volume 3, ed. by E. Hanushek, S. Machin, and L. Woessmann, 341–381, Elsevier.

Betts, Julian R., and Jamie L. Shkolnik. 2000. "Key Difficulties in Identifying the Effects of Ability Grouping on Student Achievement." *Economics of Education Review*, 19(1): 21–26.

Betts, Julian R., and Jamie L. Shkolnik. 2010. "The Effects of Ability Grouping on Student Achievement and Resource Allocation in Secondary Schools." *Economics of Education Review*, 19: 1–15.

Blanes i Vidal, Jordi, and Mareike Nossol. 2011. "Tournaments Without Prizes: Evidence from Personnel Records." *Management Science*, 57(10): 1721–1736.

Booij, Adam S., Edwin Leuven, and Hessel Oosterbeek. 2016. "Ability Peer Effects in University: Evidence from a Randomized Experiment." *Review of Economic Studies*, 0: 1–32.

Burke, Mary A. and Tim R. Sass. 2013. "Classroom Peer Effects and Student Achievement." *Journal of Labor Economics*, 31(1): 51–82.

Calsamiglia, Caterina, Jorg Franke, and Pedro Rey-Biel. 2013. "The incentive effects of affirmative action in a real-effort tournament." *Journal of Public Economics*, 98(C): 15–31.

Carrell, Scott E., Richard L. Fullerton, and James E. West. 2009. "Does Your Cohort Matter? Measuring Peer Effects in College Achievement." *Journal of Labor Economics*, 27: 439–464.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West. 2013. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica*, 81(3): 855–882.

Charness, Gary, David Cooper, and Zachary Grossman. 2015. "Silence is Golden: Communication Costs and Team Problem Solving." *Unpublished manuscript.*

Cummins, Joseph. 2017. "Heterogeneous Treatment Effects in the Low Track: Revisiting the Kenyan Primary School Experiment." *Economics of Education Review*, 56: 40–51.

Deming, David, Claudia Goldin, Lawrence Katz, and Noah Yuchtman. 2015. "Can Online Learning Bend the Higher Education Cost Curve?" *American Economic Review*, 105(5): 496–501.

Duflo, Esther, Pascaline Dupas, and Michael 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.

Eisenkopf, Gerald. 2010. "Peer effects, motivation, and learning." *Economics of Education Review*, 29(3): 364–374.

Elsner, Benjamin, and Ingo Isphording. 2017. "A Big Fish in a Small Pond: Ability Rank and Human Capital Investment." *Journal of Labor Economics*, 35(3): 787–828.

Epple, Dennis, Elizabeth Newlon, and Richard Romano. 2002. "Ability Tracking, School Competition, and the Distribution of Educational Benefits." *Journal of Public Economics*, 83(1): 1–48.

Epple, Dennis, and Richard Romano. 2011. "Peer Effects in Education: A Survey of the Theory and Evidence." In *Handbook of Social Economics*, Vol. 1B, ed. Jess Benhabib, Alberto Bisin, and Matthew Jackson, 1053–1163. Amsterdam: North-Holland, Elsevier.

Feld, Jan, and Ulf Zölitz. 2017. "Understanding Peer Effects: On the Nature, Estimation, and Channels of Peer Effects." *Journal of Labor Economics*, 35(2): 387–428.

Figlio, David N., and Marianne E. Page. 2002. "School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?" *Journal of Urban Economics*, 51(3): 497–514.

Garlick, Robert. 2016. "Academic Peer Effects with Different Group Assignment Rules: Residential Tracking versus Random Assignment." *Unpublished manuscript.*

Gill, David, Zdenka Kissova, Jaesun Lee, and Victoria Prowse. 2016. "First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision." *Unpublished manuscript*.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5): 1644–1655.

Hoxby, Caroline. 2000. "Peer Effects in the Classroom: Learning from Gender and Race Variation." *NBER Working Paper 7867*.

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

Jain, Tarun, and Mudit Kapoor. 2015. "The Impact of Study Groups and Roommates on Academic Performance." *Review of Economics and Statistics*, 97(1): 44–54.

Johnson, David, and Roger Johnson. 1975. *Learning together and alone.* Englewood Cliffs, NJ: Prentice-Hall.

Ksoll, Christopher and Kim Lehrer. 2013. "Learning from Peers: Experimental Evidence of Group Learning in Senior High School in Ghana." *Unpublished manuscript.*

Lavy, Victor, M. Daniele Paserman, and Analia Schlosser. 2012. "Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom." *The Economic Journal*, 122(559): 208–237.

Lavy, Victor, Olmo Silva, and Felix Weinhardt. 2012. "The Good, the Bad and the Average: Evidence on the Scale and Nature of Ability Peer Effects in Schools." *Journal of Labor Economics*, 30(2): 367–414.

Lefgren, Lars. 2004. "Educational Peer Effects and the Chicago Public Schools." *Journal of Urban Economics*, 56(2): 169–191.

Li, Tao, Li Han, Linxiu Zhang, and Scott Rozelle. 2014. "Encouraging classroom peer interactions: Evidence from Chinese migrant schools." *Journal of Public Economics*, 111: 29–45.

Ludwig, Jens, Jeff Kling, and Sendil Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives*, 3(25): 17–38.

Luo, Heng, Anthony C. Robinson and Jae-Young Park. 2014. "Peer Grading in a MOOC: Reliability, Validity and Perceived Effects." *Journal of Asynchronous Learning Networks*, 18(2): n2.

Lyle, David. 2007. "Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point." *Review of Economics and Statistics*, 89(2): 289–299.

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

Murphy, Richard, and Felix Weinhardt. 2014. "Top of the Class: The importance of ordinal rank." *Unpublished manuscript*.

Rohrbeck, Cynthia, Marika Ginsburg-Block, John Fantuzzo, and Traci Miller. 2003. "Peer-assisted learning interventions with elementary school students: A meta-analytic review." *Journal of Educational Psychology*, 95(2): 240–257.

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

Sacerdote, Bruce. 2011. "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" in E. Hanushek, S. Machin, and L.Woessmann, eds., *Handbook of the Economics of Education*, Dordrecht: Elsevier, 249–277.

Schunk, Dale. 1991. *Learning Theories: An Educational Perspective*. New York: Merrill.

Slavin, Robert. 1983. *Cooperative learning.* New York: Longman.

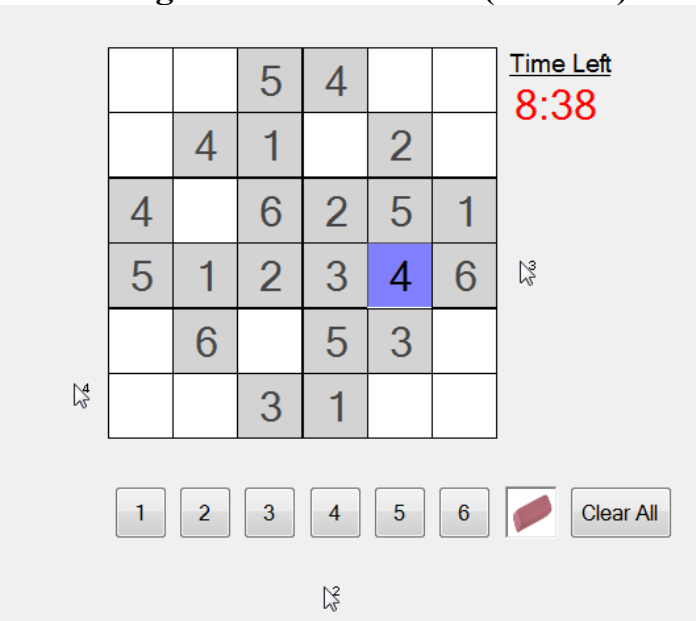
Song, Yang, George Loewenstein, and Yaojiang Shi. 2017. "Heterogeneous effects of peer tutoring: Evidence from rural Chinese middle schools." *Unpublished manuscript.*

Webb, Noreen. 1989. "Peer interaction and learning in small groups." *International Journal of Educational Research*, 13: 21–40.

Zimmer, Ron. 2003. "A New Twist in the Education Tracking Debate." *Economics of Education Review*, 22(3): 307–315.

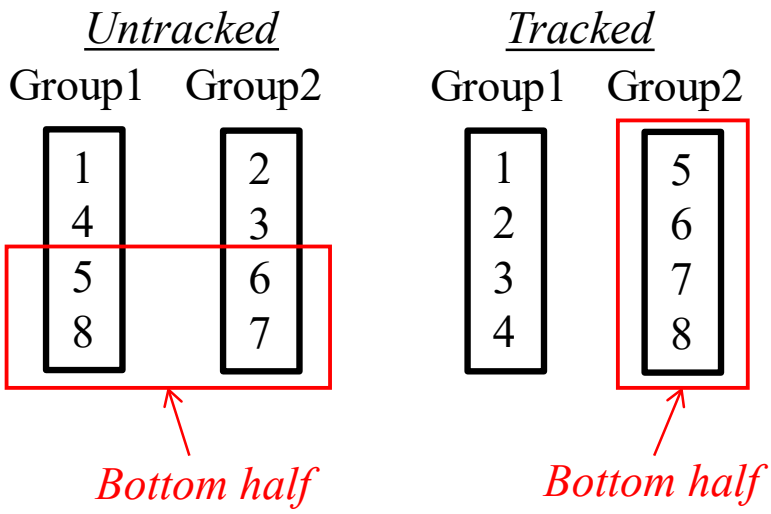
Zimmerman, David J. 2003. "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment." *Review of Economics and Statistics*, 85(1): 9–23.

Figure 1: Screenshot from the *Teaching* Treatment during the Practice Block (Sudoku)



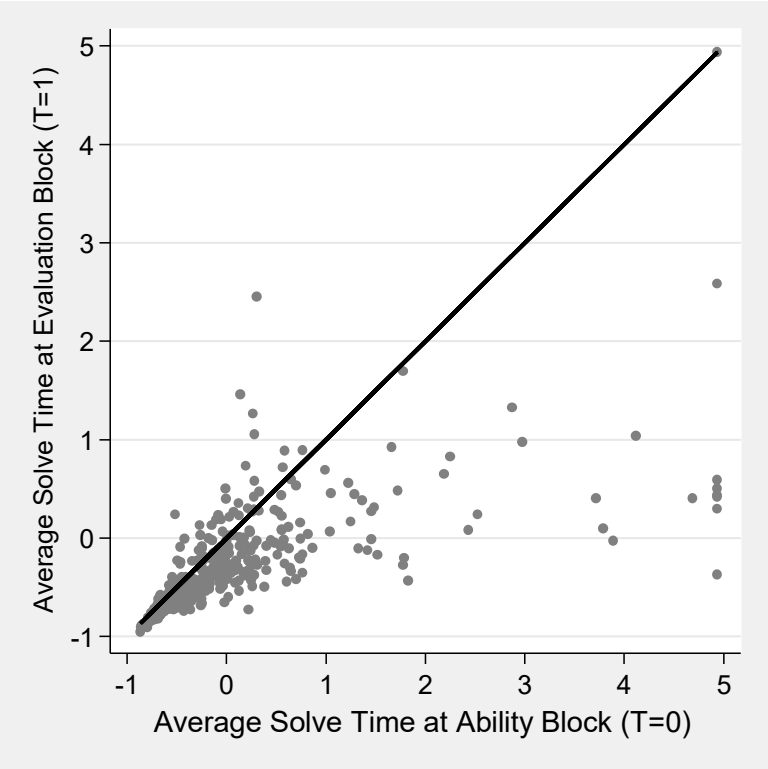
Notes: Shown from the perspective of the subject ranked 1 out of 4 in his group. Subjects are able to simultaneously edit a common 6×6 Sudoku puzzle during the Practice Block. Each other mouse arrow is labeled with the within-group performance rank of the person in the Ability Block (T=0). Performance is measured by the number of Sudoku puzzles solved with the average solving time serving as a tie-breaker. In the *no-teaching* treatment, the three arrows of other subjects would not have been visible, as each subject worked independently.

Figure 2: Group Assignment Rules in *Untracked* vs. *Tracked* Treatment



Notes: This figure describes the procedure for assigning subjects to groups in the *untracked* and *tracked* treatments. Rank is based on performance in the Ability Block (T=0). Performance is measured by the number of Sudoku puzzles solved with the average solving time serving as a tie-breaker. We define subjects ranked 5–8 in the Ability Block to be the bottom half and those ranked 1–4 to be the top half.

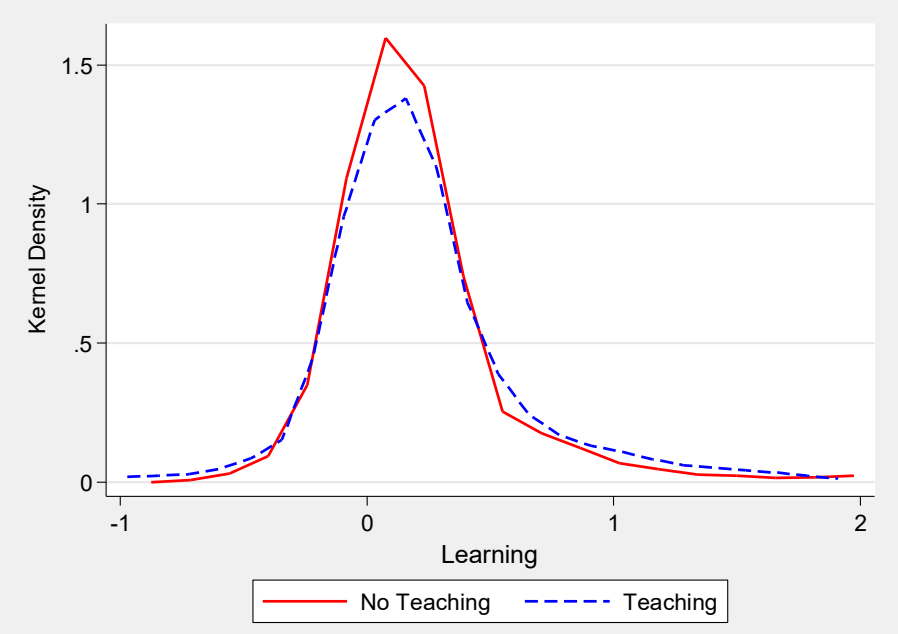
Figure 3: Standardized Average Solving Time in the Ability and Evaluation Blocks (Sudoku)



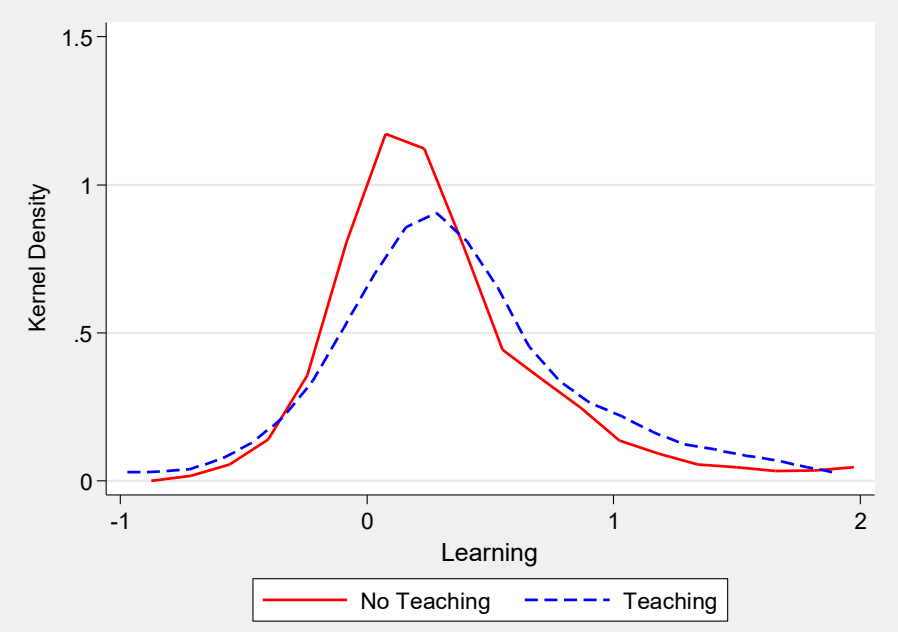
Notes: Scatter plots of the standardized average solving times (AST) in the Ability Block (T=0) and the Evaluation Block (T=1) are displayed. Raw average solving times (in seconds) for both blocks are standardized by the mean and standard deviation of average solving time at T=0 so that standardized AST at T=0 has a mean of zero and standard deviation of 1. The mean, median, and standard deviation of raw AST at T=0 are 119.14, 93.33, and 97.39 *sec*, respectively. The solid line represents the 45-degree line. Subjects below the 45-degree line show improvement in their standardized average solving time for Sudoku puzzles. "Learning," which is our main outcome, is calculated by subtracting the standardized AST at T=1 from standardized AST at T=0, so that higher values indicate *improvement* in average solving time. There are 448 subjects in total, and 371 subjects (84.4 percent) exhibited positive learning.

Figure 4: Kernel Densities of Learning in the *No-teaching* and *Teaching* Treatments (Sudoku)

A. Full Sample



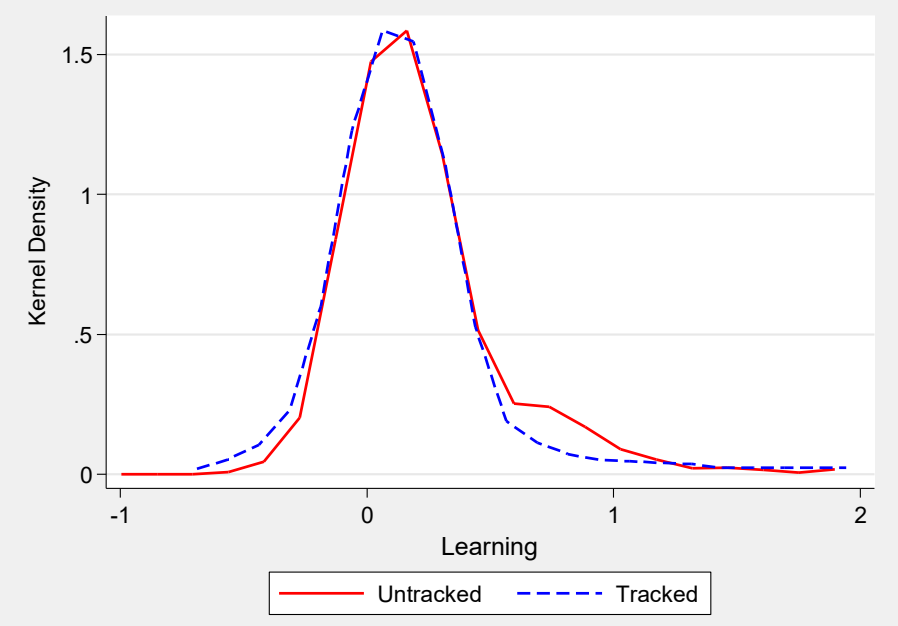
B. Among Subjects in the Bottom Half



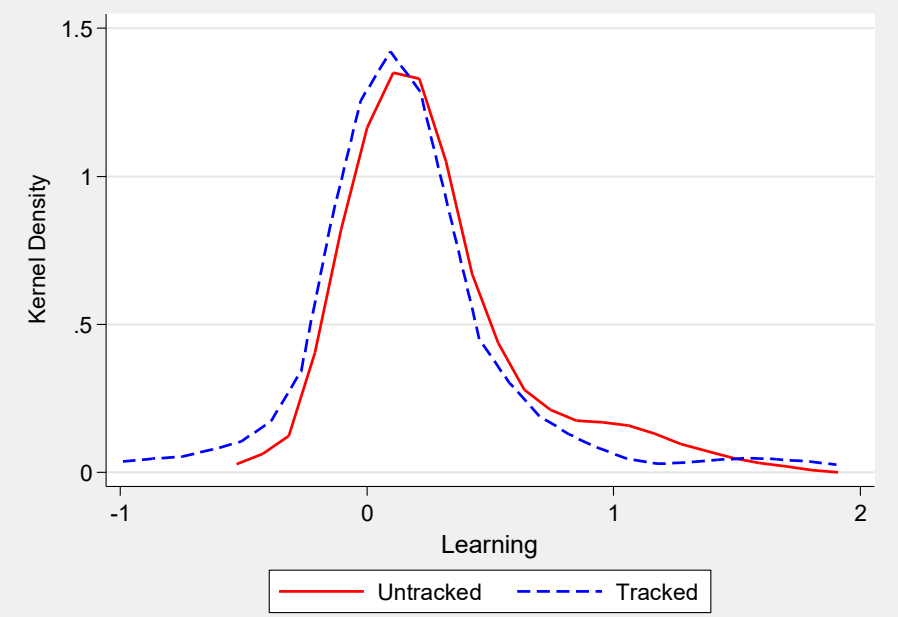
Notes: Kernel density plots of learning in the *no-teaching* and *teaching* treatments are displayed. Panel A is for the full sample while Panel B is restricted to subjects in the bottom half who ranked 5–8 in the Ability Block (T=0). Learning is calculated by subtracting the standardized average solving time in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in average solving time. For Panel A, the p-value for the two-sample Kolmogorov-Smirnov test for the equality of the distributions between the *no-teaching* and *teaching* treatments is 0.014, while that for Panel B is 0.004. There are a total of 448 subjects in Panel A and 224 subjects in Panel B.

Figure 5: Kernel Densities of Learning in the *Tracked* and *Untracked* Treatments (Sudoku)

A. In the *No-teaching* Treatment



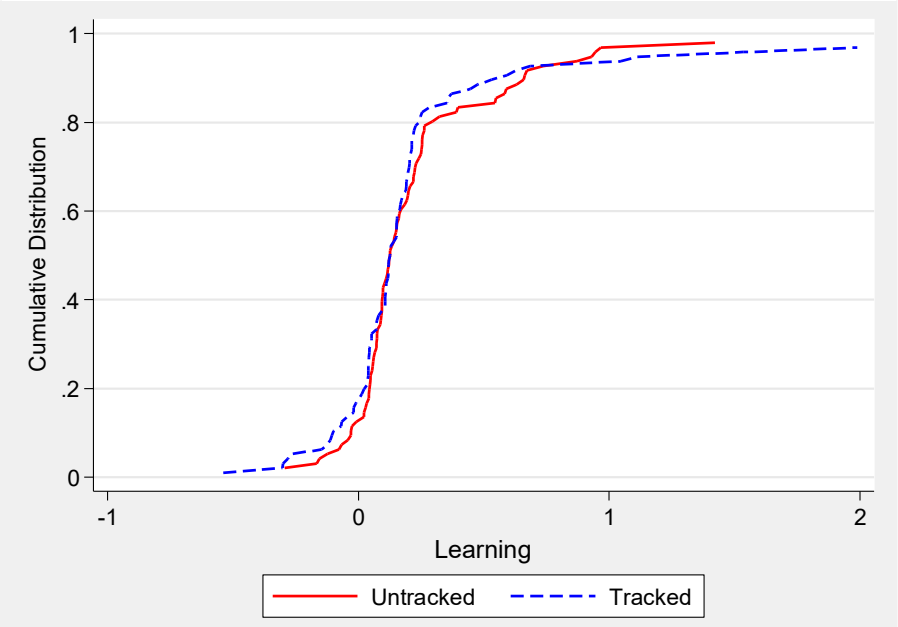
B. In the *Teaching* Treatment



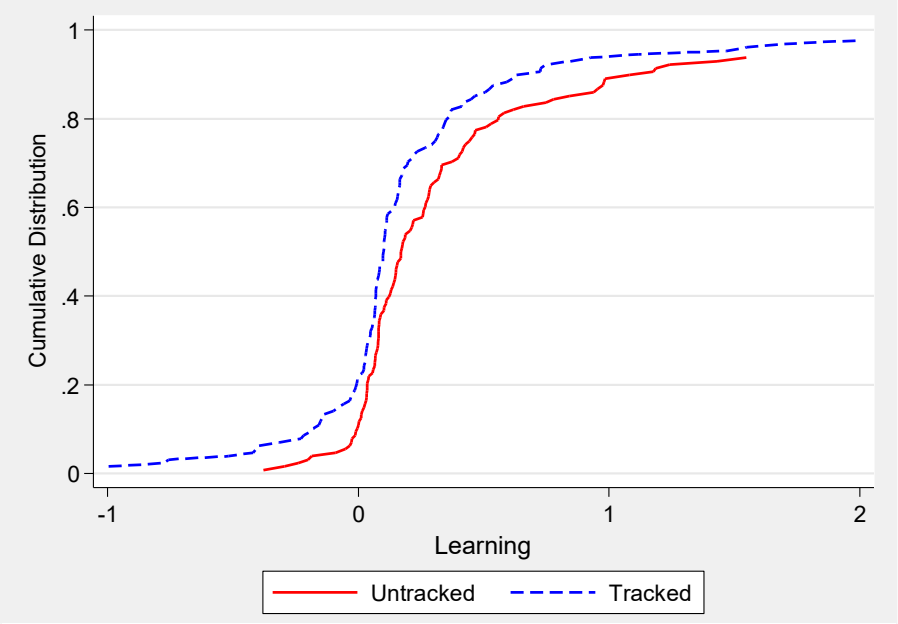
Notes: Kernel density plots of learning for the *untracked* and *tracked* treatments are displayed. Learning is calculated by subtracting the standardized average solving time in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in average solving time. Panel A plots the kernel densities in the *no-teaching* treatment where the p-value for a two-sample Kolmogorov-Smirnov test for equality of the distributions between the *untracked* and *tracked* treatments is 0.557. Panel B plots the kernel densities in the *teaching* treatment where the p-value for the Kolmogorov-Smirnov test is 0.015. There are 24 *no-teaching* sessions with 192 subjects, and 32 *teaching* sessions with 256 subjects.

Figure 6: Cumulative Distributions of Learning in the *Tracked* and *Untracked* Treatments (Sudoku)

A. In the *No-teaching* Treatment

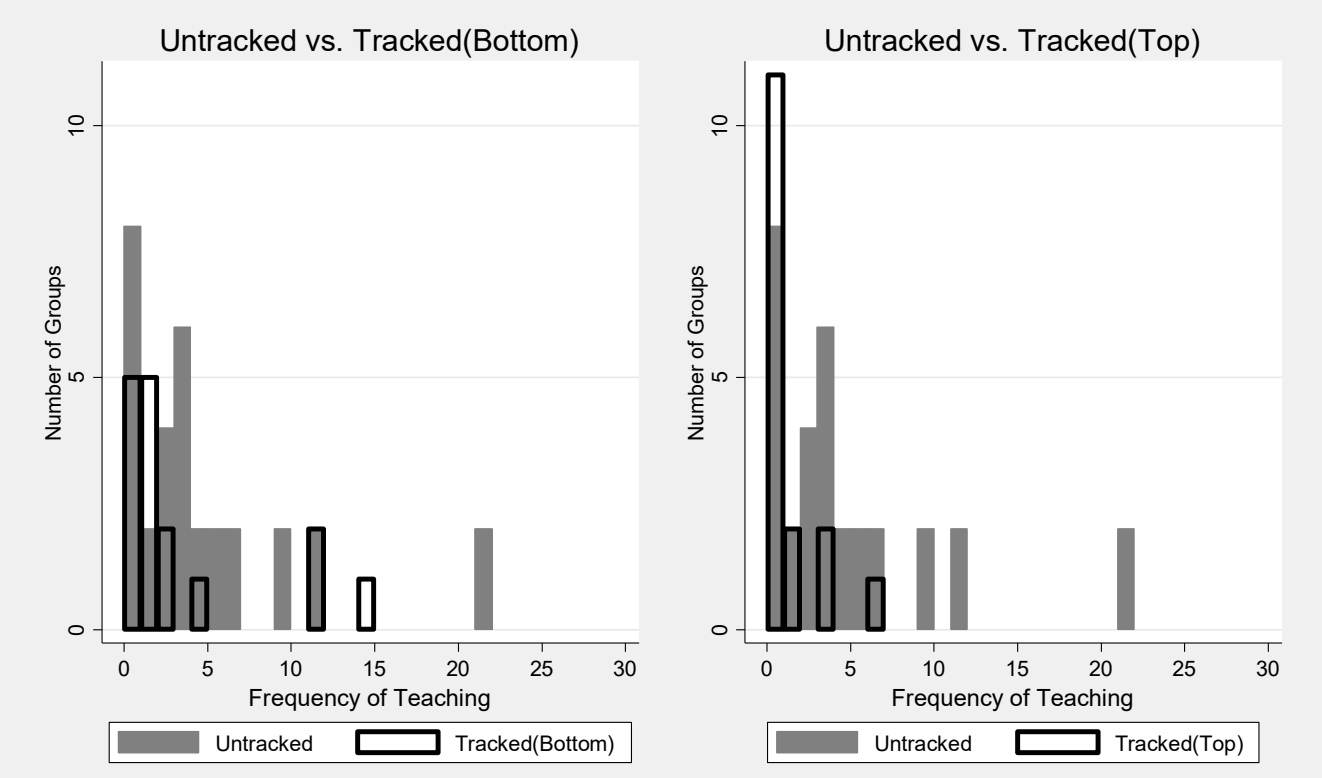


B. In the *Teaching* Treatment



Notes: Cumulative distributions of learning for the *untracked* and *tracked* treatments are displayed. Learning is calculated by subtracting the standardized average solving time in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in average solving time. Panel A plots the cumulative distributions in the *no-teaching* treatment where the p-value for a two-sample Kolmogorov-Smirnov test for equality of the distributions between the *untracked* and *tracked* treatments is 0.557. Panel B plots the cumulative distributions in the *teaching* treatment where the p-value for the Kolmogorov-Smirnov test between the *untracked* and *tracked* treatments is 0.015. There are 24 *no-teaching* sessions with 192 subjects, and 32 *teaching* sessions with 256 subjects (8 subjects per each session).

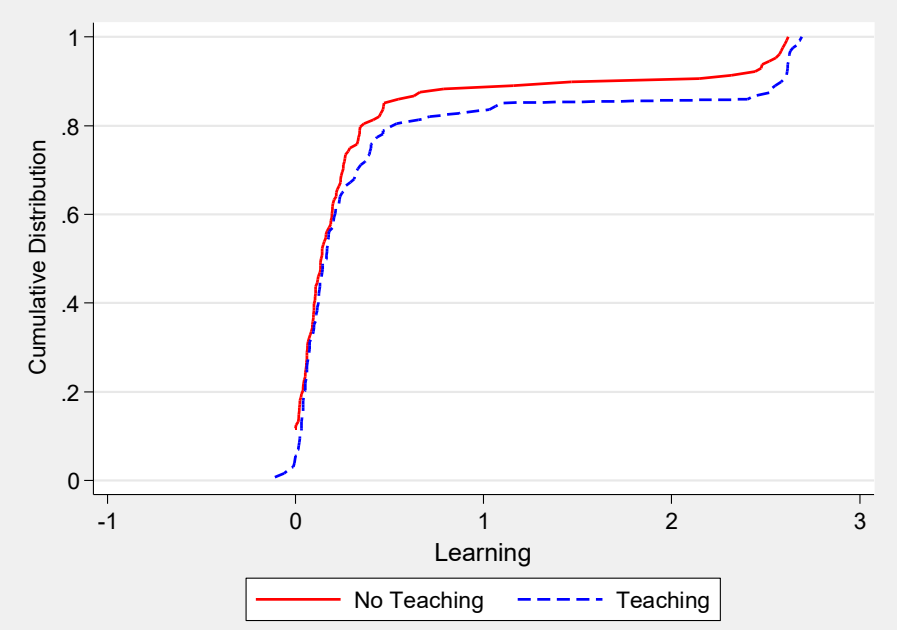
Figure 7: Frequency of Teaching (Sudoku)



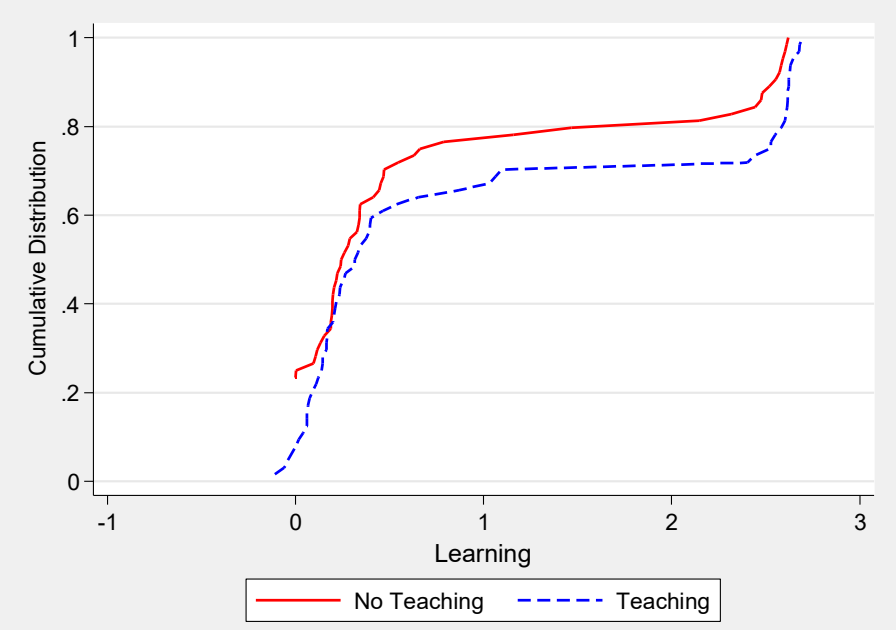
Notes: The unit of observation is a group. The sample is limited to the 32 *teaching* treatment sessions with 16 sessions each for the *untracked* and *tracked* treatments. In the *untracked* treatment, there are total of 32 groups (two groups for each session), while for the *tracked* treatment there are 16 groups each for subjects in the bottom half (Group 2 in the *tracked* treatment in Figure 2 consisting of subjects ranked 5–8 in the Ability Block (T=0)) and for subjects in the top half (Group 1 in the *tracked* treatment in Figure 2 consisting of subjects ranked 1-4 in the Ability Block (T=0)). The left graph plots the number of groups on the vertical axis exhibiting a given frequency of teaching on the horizontal axis for groups in the *untracked* treatment (N=32) and for groups consisting of subjects in the bottom half in the *tracked* treatment (N=16). The right graph similarly plots the number of groups by teaching frequency for groups in the *untracked* treatment (N=32) again and groups consisting of subjects in the top half in the *tracked* treatment (N=16). A teaching statements is defined to be any utterance in which subjects are engaged in trying to teach each other how to do Sudoku such as "You can't have a five there; there is already one in that column.

Figure 8: Cumulative Distributions of Learning in the *No-teaching* and *Teaching* Treatments (Nonograms)

A. Full Sample



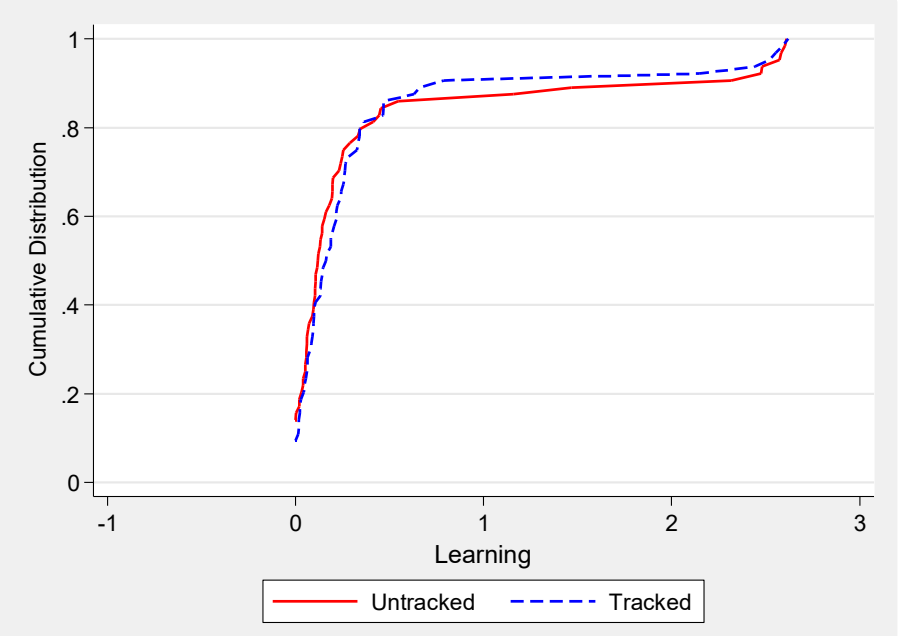
B. Among Subjects in the Bottom Half



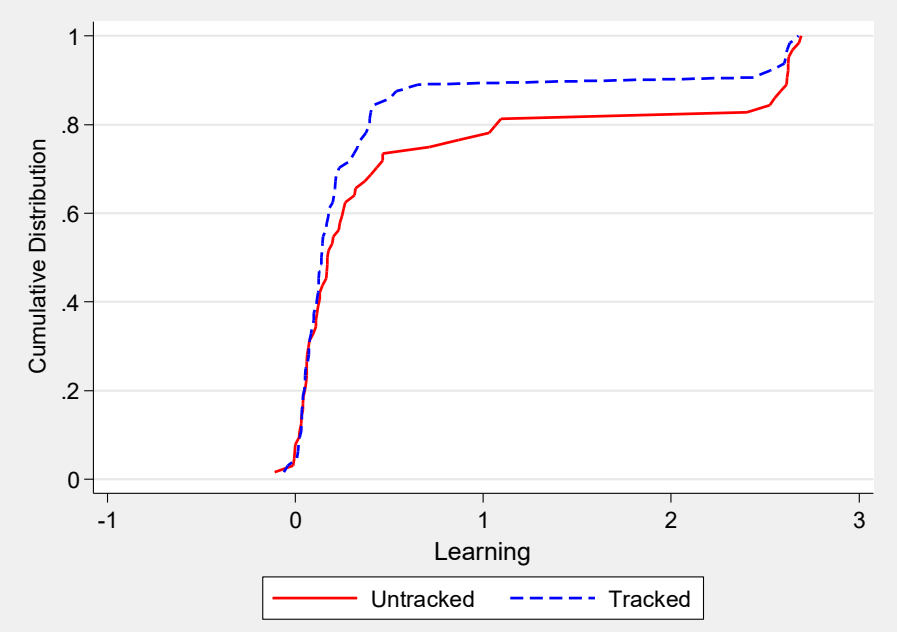
Notes: Cumulative distributions of learning in the *no-teaching* and *teaching* treatments are displayed. Panel A is for the full sample while Panel B is restricted to subjects in the bottom half who ranked 5–8 in the Ability Block (T=0). Learning is calculated by subtracting the standardized average solving time in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in average solving time. For Panel A, the p-value for the two-sample Kolmogorov-Smirnov test for the equality of the distributions between the *no-teaching* and *teaching* treatments is 0.629, while that for Panel B is 0.303.

Figure 9: Cumulative Distributions of Learning in the *Tracked* and *Untracked* Treatments (Nonograms)

A. In the *No-teaching* Treatment



B. In the *Teaching* Treatment



Notes: Cumulative distributions of learning in the *untracked* and *tracked* treatments are displayed. Learning is calculated by subtracting the standardized average solving time in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in solving time. Panel A plots the cumulative distributions in the *no-teaching* treatment while Panel B plots the cumulative distributions in the *teaching* treatment. There are 16 sessions with 128 subjects (8 subjects per each session) in both the *teaching* and *no-teaching* treatments. For Panel A, the p-value for the two-sample Kolmogorov-Smirnov test for the equality of the distributions between the *untracked* and *tracked* treatments is 0.704, while that for Panel B is 0.418.

Table 1: Structure of the Experiment

A. Overview of an Experimental Session

Elicitations	**Instructions**	**Video**	**Ability Block (T=0)**	**Practice Block**	**Evaluation Block (T=1)**
Elicit risk preferences via MPL and prosociality via a dictator game. Collect demographic data.	Basic Sudoku instructions displayed on screen, self-paced.	Provides a common "lecture" including a set of puzzle solving strategies. (9 min)	10 minutes to solve 6×6 Sudoku, paid at a piece-rate of $0.50 per correct puzzle. Performance used for *tracking*.	10 minutes to work on a single Sudoku, w/ or w/out chat for peer-to-peer *teaching*. Sorted into groups.	15 minutes to solve 6×6 Sudoku with *incentives* varied.

Notes: The table shows the time sequence for a single session. At the conclusion of each session, subjects are paid for their performance in both the Ability and Evaluation Blocks as well as for one of their choices in the risk elicitation task and, with equal probability, either their own or another person's allocation in the dictator game. See Appendix D for the full instructions used in the experiment.

B. 2×2×2 Factorial Experimental Design (Sudoku)

		No-teaching		**Teaching**	
		Piece-rate	Tournament	Piece-rate	Tournament
Untracked	# Sessions	6	6	8	8
	# Subjects	48	48	64	64
Tracked	# Sessions	6	6	8	8
	# Subjects	48	48	64	64

Notes: The total sample size is 448 subjects in 56 sessions, divided into 112 groups after the Ability Block. See Appendix D for the full instructions used in the experiment. As we show in Appendix A, our tournament incentive scheme (*piece-rate* vs. *tournament*) turned out to have a negligible effect on behavior (either through main effects of the treatment or interactions with the other treatments). As a result, our primary analysis pools data across incentive schemes and focuses on the effects of *teaching*, *tracking*, and their interaction—essentially reducing the study to a 2×2 factorial experimental design (in boldface in the table).

Table 2: Descriptive Statistics (Sudoku)

A. Summary Statistics

	Overall	Heterogeneity		
Variable	Mean (1)	Bottom half (rank5-8) (2)	Top half (rank1-4) (3)	Dif (2)-(3) (4)
Male	0.47	0.52	0.42	0.10**
	[0.50]	[0.50]	[0.49]	(0.05)
Experienced	0.68	0.48	0.88	-0.39***
	[0.47]	[0.50]	[0.33]	(0.04)
Risk Attitude (0–9)	3.35	3.31	3.40	-0.09
	[1.64]	[1.71]	[1.58]	(0.15)
Prosociality (0–5)	1.58	1.70	1.47	0.23**
	[1.03]	[0.99]	[1.05]	(0.09)
Solved None at T=0	0.02	0.04	0.00	0.04***
	[0.13]	[0.19]	[0.00]	(0.01)
Solved None at T=1	0.00	0.00	0.00	0.00
	[0.05]	[0.07]	[0.00]	0.00
Raw Average solve time at T=0 (*sec*)	119.14	167.60	70.69	96.91***
	[97.39]	[117.79]	[20.54]	(8.09)
Raw Average solve time at T=1 (*sec*)	88.09	114.49	61.70	52.79***
	[51.16]	[58.69]	[20.08]	(3.93)
Raw Learning (=AST0-AST1) (*sec*)	31.05	53.11	8.99	44.12***
	[73.49]	[98.48]	[12.23]	(6.45)
Standardized average solve time at T=0	0.00	0.50	-0.50	1.00***
	[1.00]	[1.21]	[0.21]	(0.08)
Standardized average solve time at T=1	-0.32	-0.05	-0.59	0.54***
	[0.53]	[0.60]	[0.21]	(0.04)
Learning	0.32	0.55	0.09	0.45***
	[0.75]	[1.01]	[0.13]	(0.07)
# of Sessions	56	56	56	
# of Groups	112	112	112	
# of Subjects	448	224	224	

Notes: Column (1) reports means for the full sample with standard deviations in brackets. Columns (2) and (3) report the means by ranks in the Ability Block (T=0). The bottom half consists of those subjects ranked 5–8, and the top half consists of those subjects ranked 1–4. Column (4) reports the difference in means between subjects in the top half and subjects in the bottom half with standard errors clustered at the group level in parentheses. Experienced takes a value of one if a subject reports having prior experience with Sudoku. Risk attitudes take on the values from 0 to 9 with higher numbers indicating more risk-loving subjects. Prosociality takes on the values from 0 to 5 with higher numbers indicating higher prosociality. See Appendix D for details on the elicitation of risk attitudes and prosociality and Appendix E for screenshots. Learning is calculated by subtracting the standardized AST in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in solving time. Note that AST in both the Ability and Evaluation Blocks is standardized by the mean and standard deviation of raw AST at T=0 so that standardized AST at T=0 has a mean of zero and standard deviation of 1. There were 56 sessions with 448 subjects (8 subjects per session). Each session consisted of two groups (4 subjects per group). Significance levels: *** p<0.01, ** p<0.05, * p<0.10

B. Balance tests

	Bivariate regression		Equality test	
	Teaching	Tracking	2×2×2	2×2
			(*p-value*)	
Variable	(1)	(2)	(3)	(4)
Male	-0.02	0.01	*0.77*	*0.96*
	(0.05)	(0.05)		
Experienced	0.04	0.03	*0.65*	*0.70*
	(0.05)	(0.04)		
Risk Attitudes (0-9)	0.26*	-0.01	*0.35*	*0.17*
	(0.16)	(0.16)		
Prosociality (0-5)	0.10	-0.13	*0.39*	*0.27*
	(0.10)	(0.10)		
Solved none at T=0	-0.01	0.00	*0.23*	*0.11*
	(0.01)	(0.01)		
[Raw] Average solve time at T=0 (sec)	6.32	-9.29	*0.31*	*0.15*
	(9.24)	(9.20)		
[Standardized] Average solve time at T=0	0.07	-0.10	*0.31*	*0.15*
	(0.10)	(0.09)		
# of Sessions	56	56	56	56
# of Group	112	112	112	112
# of Subjects	448	448	448	448

Notes: Columns (1) and (2) report a set of bivariate regressions that test how each variable in the far-left column is related to the teaching treatment (Column 1) and to the tracking treatment (Column 2). Standard errors are reported in parenthesis. Columns (3) and (4) report the p-values for each variable in the far-left column of the null hypotheses that the means are equal across 8 treatment combinations (Column 3) and 4 treatment combinations pooling across the incentive treatments (Column 4). Experienced takes a value of one if a subject reports having prior experience with Sudoku. Risk attitudes take on the values from 0 to 9 with higher numbers indicating more risk-loving subjects. Prosociality takes on the values from 0 to 5 with higher numbers indicating higher prosociality. See Appendix D for details on the elicitation of risk attitudes and prosociality and Appendix E for screenshots. There were 56 sessions with 448 subjects (8 subjects per session). Each session consisted of two groups (4 subjects per group). Significance levels: *** p<0.01, ** p<0.05, * p<0.10

Table 3: Effect of *Teaching* on Learning (Sudoku)

Outcome: Learning

	(1)	(2)	(3)	(4)
A. Overall				
Teaching	0.113**	0.119**		
	(0.052)	(0.052)		
B. Heterogeneity				
Teaching for Top half			-0.018	-0.016
			(0.017)	(0.020)
Teaching for Bottom half			0.240***	0.240***
			(0.096)	(0.096)
Controls	No	Yes	No	Yes

Notes: Each column reports the results from a different OLS regression. Columns (1) and (2) come from equation [2] with and without controls using the full sample. Columns (3) and (4) come from equation [3] with and without controls using full sample. Here, the control group is the *no-teaching* treatment. The estimated treatment effects and their standard errors reported in the table were computed using the *lincom* command in STATA. The coefficient estimates from equations [2] and [3] are reported in Appendix Table B1 for reference. Standard errors clustered at the group level are reported in parentheses. The outcome is learning, which is calculated by subtracting the average solving time (AST) in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in solving time. Note that AST in the Evaluation Block (T=1) and in the Ability Block (T=0) is standardized by the mean and standard deviation of raw AST at T=0 before taking the difference so that standardized AST at T=0 has a mean of zero and standard deviation of 1. The bottom half consists of those subjects ranked 5–8 and the top half those subjects ranked 1–4 at T=0. All regressions—even those labeled as including "no" controls—include a dummy for the eight subjects who could not solve any Sudoku puzzles in T=0. The controls further include a dummy for being male, a dummy for being experienced with Sudoku, risk attitudes (0–9), and prosociality (0–5). See Table 2 for definitions of each control variable. See also Appendix Table B1 for coefficient estimates for all of the control variables. There were 56 sessions with 448 subjects (8 subjects per session). Each session consisted of two groups (4 subjects per group), and thus there were 112 groups. Significance levels: *** p<0.01, ** p<0.05, * p<0.10

Table 4: Effects of *Tracking* on Learning in the *No-teaching* and *Teaching* Treatments (Sudoku)

Outcome: Learning

	In the *No-teaching* Treatment		In the *Teaching* Treatment	
	(1)	(2)	(3)	(4)
A. Overall				
Tracked	-0.039		-0.145*	
	(0.061)		(0.082)	
B. Heterogeneity				
Tracked for Top half		0.024		-0.032
		(0.026)		(0.032)
Tracked for Bottom half		-0.095		-0.283*
		(0.108)		(0.154)
Controls	Yes	Yes	Yes	Yes

Notes: The estimated treatment effects in Columns (1) and (3) come from equation [4], while the estimated treatment effects in Columns (2) and (4) come from equation [5]. The control treatment is the *untracked* treatment. The estimated treatment effects and their standard errors were computed using the *lincom* command in STATA. The coefficient estimates from equations [4] and [5] are reported in Appendix Table B2 for reference. Standard errors clustered at the group level are reported in parentheses. The outcome is learning, which is calculated by subtracting the average solving time (AST) in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in solving time. Note that AST in the Evaluation Block (T=1) and in the Ability Block (T=0) is standardized by the mean and standard deviation of raw AST at T=0 before taking the difference so that standardized AST at T=0 has a mean of zero and standard deviation of 1. The bottom half consists of those subjects ranked 5–8 in the Ability Block (T=0). See Figure 2 for details of the procedure used to assign subjects to groups in the *untracked* and *tracked* treatments. The controls include a dummy for being male, a dummy for being experienced with Sudoku, risk attitudes (0–9), prosociality (0–5), and a dummy for the eight subjects who could not solve any Sudoku puzzles in the Ability Block (T=0). There were 24 *no-teaching* sessions with 192 subjects, and 32 *teaching* sessions with 256 subjects (8 subjects per session). Each session consisted of two groups (4 subjects per group). Significance levels: *** p<0.01, ** p<0.05, * p<0.10

Table 5: Frequency of Teaching in the *Tracked* vs. *Untracked* Treatments (Sudoku)

Outcome: Frequency of Teaching

	Untracked	*Tracked*		Difference (2)–(1)		Difference (3)–(1)	
		Bottom Half	Top Half	OLS	Zero-Inflated Poisson	OLS	Zero-Inflated Poisson
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	4.78	3.13	0.94	-1.66	-1.66	-3.84**	-3.84***
	[6.60]	[4.76]	[1.91]	(1.86)	(1.02)	(1.69)	(0.86)
Vuong test of Zero-Inflated model vs. Standard Poisson					z= 3.13 p= 0.0009		z= 3.15 p= 0.0008
# of Groups	32	16	16	48	48	48	48
# of Sessions	16	16		32	32	32	32

Notes: The unit of observation is a group. The sample is limited to the 32 *teaching* treatment sessions with 16 sessions each for the *untracked* and *tracked* treatments. In the *untracked* treatment, there are 32 groups (two groups per session), while in the *tracked* treatment there are 16 groups each for subjects in the bottom half (Group 2 in the *tracked* treatment in Figure 2) and for those in the top half (Group 1 in the *tracked* treatment in Figure 2). Column (1) reports the mean number of teaching statements exhibited by groups in the *untracked* treatment, and Columns (2) and (3) report them for the *tracked* treatment for the bottom half group and the top half group, respectively. Standard deviations are reported in brackets. Columns (4) and (5) report the estimated difference between Columns (1) and (2) from OLS and zero-inflated Poisson (where the inflation equation includes just a dummy for *tracked* sessions) models, respectively, with standard errors in parentheses. Columns (6) and (7) report the corresponding estimated differences between Columns (1) and (3). A teaching statements is defined to be any utterance in which subjects are engaged in trying to teach each other how to do Sudoku such as "You can't have a five there; there is already one in that column." The Vuong tests of the zero-inflated Poisson models against the standard Poisson models are reported with the z-scores and corresponding p-values; these tests support the use of the Zero-Inflated model. Significance levels: *** p<0.01, ** p<0.05, * p<0.10

Table 6: Structure of the Experiment (Nonograms)

2×2×2 Factorial Experimental Design

		No-teaching		**Teaching**	
		Piece-rate	Tournament	Piece-rate	Tournament
Untracked	# Sessions	4	4	4	4
	# Subjects	32	32	32	32
Tracked	# Sessions	4	4	4	4
	# Subjects	32	32	32	32

Notes: See Appendix D for the full instructions used in the experiment, which are identical to the instructions for Sudoku. As we show in Appendix A, our tournament incentive scheme (*piece-rate* vs. *tournament*) also turned out to have a negligible effect on behavior (either through main effects of the treatment or interactions with the other treatments) in the Nonogram experiment as well. As a result, our primary analysis pools data across incentive schemes and focuses on the effects of *teaching*, *tracking*, and their interaction—essentially reducing the study to a 2×2 factorial experimental design (in boldface in the table).

Table 7: Effect of *Teaching* on Learning (Nonograms)

Outcome: Learning

	(1)	(2)	(3)	(4)
A. Overall				
Teaching	0.203***	0.195***		
	(0.064)	(0.064)		
B. Heterogeneity				
Teaching for Top half			0.021	0.022
			(0.022)	(0.021)
Teaching for Bottom half			0.378***	0.359***
			(0.121)	(0.121)
Controls	No	Yes	No	Yes

Notes: Each column reports the estimated treatment effects from a different OLS regression. Columns (1) and (2) come from equation [2] with and without controls using the full sample. Columns (3) and (4) come from equation [3] with and without controls using the full sample. The control treatment is the *no-teaching* treatment. The estimated treatment effects and their standard errors were computed using the *lincom* command in STATA. The coefficient estimates from equations [2] and [3] are reported in Appendix Table B7 for reference. Standard errors clustered at the group level are reported in parentheses. The outcome is learning, which is calculated by subtracting the average solving time (AST) in the Evaluation Block (T=1) from that in the Ability Block (T=0) so that higher values indicate *improvement* in solving time. Note that AST in the Evaluation Block (T=1) and in the Ability Block (T=0) is standardized by the mean and standard deviation of raw AST at T=0 before taking the difference so that standardized AST at T=0 has a mean of zero and standard deviation of 1. The bottom half consists of those subjects ranked 5–8 in T=0 and the top half consists of those subjects ranked 1–4. All regressions—including those labeled as including no controls—control for a dummy for the fifty subjects (19.4 percent) who could not solve any Nonogram puzzles in T=0. The controls further include a dummy for being male, a dummy for being experienced with Nonograms, risk attitudes (0–9), and prosociality (0–5). See Table 2 for definitions of each control variable. There were 32 sessions with 256 subjects (8 subjects per session). Each session consisted of two groups (4 subjects per group). Significance levels: *** p<0.01, ** p<0.05, * p<0.10

Table 8: Effect of *Tracking* on Learning in the *No-teaching* vs. *Teaching* Treatment (Nonograms)

Outcome: Learning

	In the *No-teaching* Treatment		In the *Teaching* Treatment	
	(1)	(2)	(3)	(4)
A. Overall				
Tracked	0.067		-0.063	
	(0.095)		(0.093)	
B. Heterogeneity				
Tracked for Top half		0.024		-0.063**
		(0.044)		(0.031)
Tracked for Bottom half		0.095		-0.081
		(0.181)		(0.167)
Controls	Yes	Yes	Yes	Yes

Notes: The estimated treatment effects in Columns (1) and (3) come from estimating equation [4] on the Nonograms sample, while the estimated treatment effects in Columns (2) and (4) come from equation [5]. The control treatment is the *untracked* treatment. The estimated treatment effects and their standard errors were computed using the *lincom* command in STATA. The coefficient estimates from equations [4] and [5] are reported in Appendix Table B8 for reference. Standard errors clustered at the group level are reported in parentheses. The outcome is learning, which is calculated by subtracting the average solving time (AST) in the Evaluation Block (T=1) from that in the Ability Block (T=0) so that higher values indicate *improvement* in solving time. Note that AST in the Evaluation Block (T=1) and in the Ability Block (T=0) is standardized by the mean and standard deviation of raw AST at T=0 before taking the difference so that standardized AST at T=0 has a mean of zero and standard deviation of 1. The bottom half consists of those subjects ranked 5–8 in the Ability Block (T=0). See Figure 2 for details of the procedure used to assign subjects to groups in the *untracked* and *tracked* treatments. The controls include a dummy for being male, a dummy for being experienced with Nonograms, risk attitudes (0–9), prosociality (0–5), and a dummy for the fifty subjects (19.4 percent) who could not solve any Nonogram puzzles in the Ability Block (T=0). There were 16 sessions with 128 subjects for both the *no-teaching and teaching* treatments (8 subjects per session). Each session consisted of two groups (4 subjects per group). Significance levels: *** p<0.01, ** p<0.05, * p<0.10