DISCUSSION PAPER SERIES

How Do Peers Impact Learning? An Experimental Investigation of Peer-To-Peer Teaching and Ability Tracking

Erik O. Kimbrough
Andrew D. McGee
Hitoshi Shigeoka

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

How Do Peers Impact Learning?
An Experimental Investigation of
Peer-To-Peer Teaching and Ability Tracking

Erik O. Kimbrough
Simon Fraser University

Andrew D. McGee
University of Alberta and IZA

Hitoshi Shigeoka
Simon Fraser University and NBER

MAY 2017

ABSTRACT

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

Classroom peers are believed to influence learning by teaching each other, and the efficacy of this teaching likely depends on classroom composition in terms of peers' ability. Unfortunately, little is known about peer-to-peer teaching because it is never observed in field studies. Furthermore, identifying how peer-to-peer teaching is affected by ability tracking – grouping students of similar ability – is complicated by the fact that tracking is typically accompanied by changes in curriculum and the instructional behavior of teachers. To fill this gap, we conduct a laboratory experiment in which subjects learn to solve logic problems and examine both the importance of peer-to-peer teaching and the interaction between peer-to-peer teaching and ability tracking. While peer-to-peer teaching improves learning among low-ability subjects, the positive effects are substantially *offset* by tracking. Tracking reduces the frequency of peer-to-peer teaching, suggesting that low-ability subjects suffer from the absence of high-ability peers to teach them.

Corresponding author:
Hitoshi Shigeoka
Department of Economics
Simon Fraser University
8888 University Drive
Burnaby, BC V5A 1S6
Canada
E-mail: hitoshi_shigeoka@sfu.ca

NON-TECHNICAL SUMMARY

How important are classroom peers as teachers? In and out of the classroom, students may teach each other when working on assignments or simply chatting about class. In this study, we provide evidence from a laboratory experiment in which students learn to solve logic problems that this "peer-to-peer" teaching is potentially very important. Allowing peer-to-peer teaching for only 10 minutes in our experiment increases in learning by as much as 42 percent.

The use of a laboratory experiment is important because we are able to compare students who are offered the opportunity to teach each other with students who are not given this opportunity – a comparison which is impossible in real classrooms. Furthermore, the experiment allows us to vary the peers with which students interact. We find that grouping students of similar ability as tracking policies do in schools has a very adverse effect on peer-to-peer teaching for low ability students. Specifically, these students – who stand to gain the most from peer-to-peer teaching – lose out on interactions with higher ability peers who can teach them. In the end, grouping students of similar ability completely eliminates the benefits of peer-to-peer teaching for low-ability students.

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. Despite the self-evident preeminence of peer-to-peer teaching, there appear to have been no studies of this mechanism in economics.[1]

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 ability tracking is widely employed, it 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 hence learning, 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]).

Unfortunately, the effects of ability tracking, peer-to-peer teaching, 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 likewise may influence teacher behavior if reducing the heterogeneity in student ability in a classroom enables teachers to better tailor the pace and content of instruction. Finally, peer-to-peer teaching is rarely—if ever—observed in field data.

To fill the gap, 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 ideal for investigating

---

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

the effects of peer-to-peer teaching and tracking because we can exogenously vary subjects' ability to interact with each other—which is impossible in the real world—and peer group composition.[2] 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.

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). 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 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.[3]

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

---

[2] Many field studies such as those from the U.S. service academies (e.g., Lyle [2007], Carrell et al. [2009]) are able to exploit exogenous variation in peer group *composition* to identify the effects of peers, but we know of no field studies in which the possibility for student *interaction* varies exogenously.

[3] As we discuss below, we also varied the compensation scheme in the Evaluation block from *piece-rate* payments to *tournament*-style payments. Our incentive structure in the *tournament* treatment, however, may have been too linear to induce treatment effects relative to the *piece-rate* treatment, and this treatment manipulation failed to influence observed behavior. Therefore, our analysis of this treatment is relegated to Appendix A.

subjects to teach each other improves learning by 0.119 standard deviations (SD) of the average solving time in the Ability block (or a reduction in the raw average solving time by 11.6 seconds) relative to the *no-teaching* treatment. This represents a 42 percent increase in learning compared to the mean of 27.6 seconds 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 it alone, improves learning by as much as 0.119 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.

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. More importantly, *tracking* has a large detrimental effect on learning in the *teaching* treatment—very 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 the 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 study fills a major gap in the education production function literature by establishing the potentially outsized contribution of peer-to-peer teaching to learning. The importance of peer-to-peer teaching for learning, however, 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. Significantly, we know of no study in which students in the control treatments were expressly forbidden from communicating with other students. 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. We believe ours to be the first study to identify the total effect of peer-to-peer teaching relative to an environment in which peer-to-peer teaching is impossible.

Our findings concerning the importance of peer-to-peer teaching also help to reconcile seemingly conflicting findings in the tracking literature. Specifically, Duflo et al. [2011]—among others—find that tracking has positive effects on students of *all* ability levels in a randomized control trial in Kenya, 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.[4] Notably, the key institutional difference between two studies is that the tracking in Garlick [2016] did not involve modifications to the curriculum or other changes to the learning environment, whereas the tracking program studied by Duflo et al. [2011] did involve modifications to the curricula and resulted (as they show) in increased teacher effort in tracking schools. Because the treatment effects we identify are driven by changes in students' behaviors alone and not contaminated by either adaptive teacher behaviors or by changes in the curriculum, our findings suggest that the non-instructional (residential) tracking in Garlick [2016] harms low-ability students because they no longer benefit from interactions with high-ability peers.[5] On the other hand, our findings suggest that the benefits of ability tracking to low-ability students observed by Duflo et al. [2011] may stem from the positive effects of changes in instruction and curriculum associated with tracking that *overcome* any negative effects on low-ability students of losing out on interactions with higher-ability peers.

Our findings have important implications for policy decisions made in every school and classroom regarding the amount of peer-to-peer teaching to encourage and the composition of student groups.[6] Specifically, the findings suggest that grouping students by prior achievement disadvantages low-ability students to the extent that these students benefit from being taught by higher-ability peers. Unless ability tracking is accompanied by curricular customization and/or incentives for teachers to adapt their instruction to the student group composition, ability tracking alone may harm those students who stand to gain the most from peer-to-peer teaching.

---

[4] Most studies of tracking join Duflo et al. [2011] in finding 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.
[5] 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.
[6] There is a small literature on optimal design of group composition (e.g., Bhattacharya [2009], Carrell et al. [2013], Booij et al. [2016]).

In addition to the literature on tracking, this paper belongs to the broader literature on peer effects in the classroom. 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.[7] Even in a randomized control setting, however, classrooms may differ on multiple dimensions in addition to peer group composition if teachers and schools respond to changes in group composition by changing their effort or allocation of other resources. Thus, isolating a causal effect of peers on student outcomes via peer group composition net of these other factors requires strong assumptions. In this paper, we argue that because only peer-to-peer interactions are present (and exogenously varied) in our setting, we can directly examine how ability tracking and hence group composition affect learning through peer-to-peer interactions.

This study is also related to a small but recently growing literature in economics using laboratory 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.[8]

Finally, our findings are directly relevant for the optimal design of so-called Massive Online Open Courses (MOOCs). As technological improvements have made such courses more viable amid a push by universities (and students) to reduce costs over the last decade, enrollment in MOOCs has grown tremendously (Deming et al. [2015]). Students in MOOCs interact in virtual environments very 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 classroom environments when instruction is not being customized for particular students. In particular, our findings suggest that learning in MOOCs could be enhanced if these courses encouraged student interaction through assignments completed by groups that include students of varied ability.

---

[7] See, for example, Hoxby [2000], Zimmerman [2003], Angrist and Lang [2004], Ammermueller and Pischke [2009], Imberman et al. [2012], Lavy, Paserman, and Schlosser [2012], Burke and Sass [2013], Booij et al. [2016], and Feld and Zölitz [2017] to name a few.

[8] See also Koch et al. [2015] for a survey relating findings from lab experiments to issues in education.

The rest of the paper is organized as follows. Section 2 describes the experimental design and our hypotheses. Section 3 describes the data and summary statistics. Section 4 reports the findings. Section 5 discusses a replication exercise, and Section 6 concludes.

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 distinct 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 substantial variety of Sudoku puzzle 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 more detail.

1. Elicitations: Each session contained 8 subjects and began with collecting some basic demographic information as well as incentivized measures of risk attitudes and prosociality, which we include as controls in some of our analyses (see Appendix C for detailed instructions and Appendix D

for screenshots). Risk attitudes were elicited using a multiple price list design based on Holt and Laury [2002]. Subjects made nine binary choices between option A (a fixed lottery with an equal chance of paying $1 and $3) and option B (a lottery between $0 and $3, where the probability of the higher payoff is increasing over the choice sequence from 0.1 to 0.9). Subjects were paid based on their decision in one-randomly selected choice problem from the set of nine. Our risk measure is the number of times that a subject chose the risky option such that higher scores (ranging from 0 to 9) indicate more risk-loving subjects. 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 indicated they would give to their partners out of the $5, such that higher values suggest more prosociality. Payoffs for these incentivized 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 basic instructions explaining the rules of Sudoku and a common "lecture" seen by all subjects in all treatments. For the lecture, we selected an online video explaining some of the aforementioned puzzle solving strategies and required all subjects to watch the video prior to the start of the experiment.[9] This ensured that each "student" received the same "lecture" and thereby controlled for the influence of the instructor. This is an important feature of our design because it has often been difficult to disentangle the exogenous effect of tracking from the endogenous response of the teacher in a tracked classroom (see, for example, Duflo et al. [2011] and Booij et al. [2016] for discussions of this issue). We chose the particular video lecture because it highlights learnable strategies of varying difficulty that subjects might also teach to (or reinforce in) one another if allowed the opportunity.

4. Ability Block (T=0): After watching the video, subjects in all treatments were told that they would have 10 minutes to work on Sudoku puzzles on their own being paid $0.50 for each puzzle that they completed correctly in that time. We refer to this as the "Ability Block," and it provides us with a measure of individual performance (ability) that is free from any influence by peers 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

---

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

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.

After the Ability Block, subjects are told that they have 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).[10] 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. 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 provides subjects with information about the ability of others in the group. As such, we displayed information about the performance of all group members in both the *untracked* and *tracked* treatments to guarantee that subjects in both treatments had the same information about their peer group.

Next, subjects read instructions explaining that first they would have a chance to practice Sudoku for 10 minutes and then they would have 15 minutes to work on another block of puzzles in which they would again be paid based on their performance. Half of the sessions were told that they would be paid based on 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). Specifically, 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 the treatments based on pilot data measuring Sudoku performance.

Thus subjects were informed about both the matching scheme and the incentive system that they would face in the "Evaluation Block" *prior* to their participation in the "Practice Block." This allowed for the possibility that knowledge of the matching and incentive schemes would influence the decision 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 compensated). In the *no-teaching* sessions, this was time for individual practice, which allowed subjects to test out various strategies or refine their skills. The *no-teaching*

---

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

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 still 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 member of the group and updated in real time on each group member's screen (see Figure 1). The subjects were connected via audio chat, and they were each represented on the 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.[11] The purpose of the Practice Block was described in the instructions for the *teaching* treatment as follows (See Appendix C for details):

> 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, the *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 make it clear that this is an opportunity for group work—which we intended to encourage peer-to-peer teaching—but subjects in the *teaching* sessions were not provided with any additional incentives (or disincentives) to teach each other (i.e., payment is at the individual level rather than the group level). In this sense, observed teaching likely results from some combination of intrinsic motivation and an experimenter demand effect. We adopted this approach because actual classrooms are also usually devoid of explicit incentives for students to teach each other. Audio chats were conducted via the open-source software OpenTokRTC, which allowed us to record a complete audio record for each group.

---

[11] 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.

6. Evaluation Block (T=1): The final portion of the experiment was a 15-minute Evaluation Block in which subjects in all sessions again worked independently to solve Sudoku puzzles. Prior to the start of 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). The difference in performance in the Ability and Evaluation Blocks provides our main data on the extent of learning in our virtual classroom. After the Evaluation Block ended, subjects were informed about their earnings from the Elicitations as well as from 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. In total we report data from 448 subjects in our Sudoku experiments (56 experimental sessions).[12] 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).[13] Below we highlight the main hypotheses our design is intended to test.

2.2 Hypotheses

As we show in Appendix A, our tournament incentive scheme turned out to have a negligible effect on behaviour—we find neither main effects of the treatment nor interactions with the other

---

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

treatments.[14] 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), the possibility of peer-to-peer teaching will have a positive effect on subject performance as subjects help each other learn to solve puzzles. Such an effect should be largest among those who perform the worst in the Ability Block because they have the most to gain. We test this hypothesis by comparing the rate of learning across the *teaching* and *no-teaching* treatments.

Hypothesis 2 (Effect of *Tracked* under *No-teaching*): ambiguous. It is unclear how tracking will influence behavior without teaching given the evidence in the literature of psychological encouragement/discouragement effects from information about relative standing in performance (beyond those provided by incentives).[15] Subjects in the top (bottom) group may be encouraged (discouraged) when learning their relative ranking. The potential for offsetting effects makes the overall effect of tracking ambiguous, but our design provides evidence that helps resolve this theoretical ambiguity.

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 more similar ability may be more effective in teaching one another if they find it easier to express their difficulties to one another or to target their suggestions to address those difficulties. As such, the potential for offsetting effects makes the effect of ability tracking ambiguous, but our design facilitates an empirical resolution.

---

[14] 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). Indeed, 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.

[15] 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). 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.[16] Formally, our measure of learning for each individual 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. Because our learning measure is skewed to the right especially among subjects at the lower tail of the distribution as shown below, 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). We can interpret estimates using the change in logged learning as our dependent variable in terms of percentage changes in average solving time.

Another natural candidate for the outcome variable would be the change in the number of Sudoku puzzles solved from T=0 to T=1. Unfortunately, the change in the number of Sudoku puzzles solved has three limitations as a dependent variable. First, it is not a direct measure of *ability* because a high-ability subject who knows his/her place in the ability distribution might not have an incentive to work for the entire available time (especially under tournament incentives). If there is disutility from effort or fatigue, it may be optimal to solve a few puzzles quickly and then rest rather than working until the end of the experiment. Second, due to the limited 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,

---

[16] Using a standardized measure of learning also facilitates comparisons across environments in which there are differences in ex ante (baseline) proficiency. Indeed, we make such a comparison in a robustness check discussed in Section 5.

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 2.47 (=75.1/30.4) and 0.55 (=2.8/5.1) for changes in raw average solving time and changes in the number of Sudoku puzzles solved, respectively. Third, 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; using average solving time for correctly solved problems allows us to distinguish between these subjects.

3.2. Summary Statistics and Balance Checks

Table 2 presents summary statistics and the results of balance tests examining whether our randomization of subjects to treatments was successful. Column (1) 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 seconds) per puzzle solved, while in the Evaluation Block (AST_{1i}) it is 88 seconds, yielding average raw "learning" of 31 seconds. 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.[17] Finally, the mean of our main outcome—learning measured in standard deviations of AST_{0i}—is 0.32.

Column (2) reports p-values for F-tests of the null hypothesis that the means across the eight treatments (2×2×2) are equal in order to check whether our random assignment of subjects to treatments was successful. Each cell reports the p-value for a separate test for each variable in the far-left column.

---

[17] 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. Also, 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 (results available upon request).

For all variables, we fail to reject the null hypothesis of no differences across treatments.[18] While this exercise is limited to these observable variables that we collect in the experiment, it is reassuring that the randomization seems to succeed on these dimensions. As mentioned earlier, we pool data across incentive schemes in our main analysis and focus on teaching (*no-teaching/teaching*) and tracking (*tracked/untracked*) and their interaction, which essentially reduces our design to four treatments (2×2). Therefore, Column (3) reports the p-value for 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 used in that robustness exercise (Section 5). As such, in much of the analysis to follow we control for these subject characteristics.

As we are especially interested in the heterogeneity of learning by initial performance, Columns (4) and (5) 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 (6) 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 seconds vs. 71 seconds), 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 seconds vs. 9 seconds). Also, the minimal raw learning (9 seconds) by subjects in the top half suggests that high-ability subjects already achieve near-peak performance solving Sudoku puzzles 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

---

[18] 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 that the variables are measured on an ordinal or continuous scale, an assumption which does not apply for our binary variables.

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 are already performing at a very high level at T=0, while subjects in the bottom half have far more room for improvement.

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.[19] 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). Standard errors are clustered at the group level, where each session consists of two groups, whether in the *untracked* or *tracked* treatment.

Next, to evaluate whether learning is heterogeneous across the ability distribution, we estimate the following equation:

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

where $Bottom$ is a dummy equal to one for subjects in 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),

---

[19] 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.

while the top half consists of the subjects ranked 1–4. Because our interest is in the heterogeneous treatment effects by initial ability (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.[20] 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.[21] This improvement by 0.12 SD corresponds to a reduction in raw average solving time of 11.6 seconds. Given that the mean learning in the no-teaching treatment is 27.6 seconds, this translates into a 42 percent increase in learning.

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

---

[20] 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.

[21] 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.

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 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 facilitate learning.[22] As such, we fill a significant gap in our knowledge of the education production function as our findings suggest that peer-to-peer teaching may be a critical component of successful learning environments.

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.

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 the subjects in top half and bottom half, 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).

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*

---

[22] 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.

treatment. Appendix Figure B2 displays the empirical CDFs of learning by treatment among subjects in the bottom half and points to the same conclusion.

Robustness Checks: Because our learning measure is skewed to the right especially among subjects in the bottom half, we estimate alternative models that allow us to assess the robustness of our estimates after attempting to mitigate the influence of skewness. Specifically, 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.

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, it is important to understand the direct effect of tracking in the *absence* of teaching, which is theoretically ambiguous. Subjects in the top group may be encouraged when learning their relative ranking while subjects in the bottom group may be discouraged. In addition, learning about the ability of other 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.[23]

We estimate

$$Learning_i = \alpha_0 + \beta_1 Tracked_i + \beta_2 Teaching_i + \beta_3 Tracked_i \times Teaching_i + \gamma X_i' + \varepsilon_i -[4]$$

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

---

[23] 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, however, are unlikely to be important in our short experiment.

teaching treatment (Hypothesis 3 as shown later). To examine heterogeneous effects by initial ability, we estimate

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

which adds to estimating 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. The standard errors are derived by the *lincom* command in Stata. We report the coefficient estimates from equations [4] and [5] in the Appendix Table B2 for reference.

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 two-sample 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*)

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 seconds) in the *untracked* and *tracked* sessions, respectively.[24] 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

---

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

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] (again, the coefficient estimates are reported in Appendix Table B2).

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. 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 0.145 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 B3). 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.[25] For subjects in the bottom half, the estimates in Column (4) indicate that as *tracking* reduces learning by as much as 0.28 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.[26]

Robustness Checks: Table 5 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.[27]

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—

---

[25] 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. Again, the standard errors are derived using the *lincom* command in Stata.

[26] 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.

[27] The coefficient estimates from equations [4] and [5] are reported in Appendix Table B4.

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. That low-ability students benefit from being in untracked groups is consistent with Lyle's [2009] finding that West Point cadets benefit from greater heterogeneity in their peers and that this benefit—though present throughout the distribution—is largest for low-ability cadets. It is also consistent with Jain and Kapoor's [2015] finding that low-ability MBA students benefit from greater heterogeneity in their roommates' ability. Importantly, in these studies as in our own, the heterogeneity at the group level (West Point companies, roommates, and our experimental groupings) does not affect the instruction received.

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' verbal conversations with members of their group during the Practice Block. Two research assistants (who were unaware of both 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 met and cross-checked their counts, resolving the few disagreements.[28]

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 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

---

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

22

same distribution for groups composed of the top half of Ability Block performers in the *tracked* treatment (N=16).

Table 6 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 actual instances of peer-to-peer teaching.

Evidence: Table 6 presents the results of our analysis of teaching frequency. The analysis 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 for each session there are two groups each. 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 *tracked* groups containing the bottom half of performers in the Ability Block (ranks 5–8) and the top half (ranks 1–4), respectively.

Columns (4)–(6) of Table 6 report the differences between Columns (1) and (2) estimated via different econometric models. Column (4) presents the estimates from OLS, Column (5) estimates from a Poisson model in order to account for the discrete nature of teaching frequencies, and Column (6) estimates from a zero-inflated Poisson model in order to further account for the fact that there are a number of groups without any teaching. Note that the Vuong test of the zero-inflated Poisson model versus the standard Poisson reported in the table indicates that the zero-inflated models are preferred in all specifications in Table 6. Columns (7)–(9) 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, and all the estimates except for Column (4) are statistically significant at conventional levels. The reduction in teaching frequency due to ability tracking is much more substantial among subjects in the top half (Columns (7)–(9) of Table 6) than the reduction among subjects in the bottom half (Columns (4)–(6) of Table 6). 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.

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 (ability difference) is necessary to generate a meaningful exchange of information in the form of questions and (correct) answers.

To explore the mechanism behind the reduction in teaching frequency due to tracking, Appendix Table B4 correlates the number of teaching statements with the group mean and group standard deviation (SD) of standardized average solving time at T=0. Here, note that 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).[29]

Throughout Columns (2)–(4), the most robust result is that the 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. These results are again consistent with Lyle [2009] and Jain and Kapoor [2015], both of which show that larger group heterogeneity benefits students in settings where instruction was unlikely to be influenced by group heterogeneity.[30]

Column (4) shows that, while imprecisely estimated, the coefficient 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. However, given that the standard deviation of group mean is 0.56, the estimate in Column (4) implies that even the group with the worst subjects in our data with group mean of 2.00 would still benefit from larger group SD (= $0.40 - 0.18 \times 0.56 \times 2 > 0$).

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

---

[29] The Vuong test of the zero-inflated Poisson model versus the standard Poisson reported in the table indicates that the zero-inflated model is preferred in all specifications in Appendix Table B4.

[30] Most studies of group heterogeneity in the classroom—wherein the instruction presumably responds to group heterogeneity—find that greater heterogeneity leads to worse academic outcomes (e.g., Booij et al. [2016]). One exception is Vigdor and Nechyba's [2007] finding that greater classroom heterogeneity had positive effects on math and reading test scores for elementary and middle school students in North Carolina.

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 (based on our estimates) ineffective.[31]

5. Replication Exercise Using a Different Game: Nonograms

So far, we have documented that most of the 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." That is, high-ability subjects may already achieve near-peak performance in solving Sudoku puzzles even during the baseline (T=0) and thus have no scope for improvement no matter what the treatment. Indeed, mean raw learning by subjects in the top half is only 9 seconds per puzzle (as compared to 53 seconds 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 C). 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.[32] 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 7 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*

---

[31] 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.

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

treatment (16 sessions) with a total of 256 subjects. 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 two-sample Kolmogorov-Smirnov tests for equality of the distributions shown below are statistically significant even though it is apparent that the distributions of learning visually look 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. In total, only 2 percent of subjects had prior experience with Nonograms. In addition, 20 percent of subjects could not solve any Nonograms at T=0. 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. Thus unlike in the Sudoku experiments, even subjects in the top half may have some scope for improvement. Also, Column (2) shows that for all variables, we fail to reject the null hypothesis of no difference across treatments, and thus randomization seems successful. Because we pool sessions from different incentive schemes again, however, we also report the p-values for tests of the null hypothesis that the covariates are balanced across four treatments (2×2).[33] Here, the p-values for several variables are just above conventional significance levels suggesting that some variables are not perfectly balanced across treatments due to the small sample. 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 as well. 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 the subjects in the bottom half. These figures clearly show that the learning among subjects in the *teaching* treatment stochastically dominates the learning among subjects in the *no-teaching* treatment—especially among subjects in the bottom half.

Table 8, which corresponds to Table 3 for Sudoku, confirms this visual inspection.[34] 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. This

---

[33] Again, our tournament incentive scheme turned out to have a negligible effect on behavior (either through main effects of the treatment or interactions). For example, 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 Tables A1).

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

average effect, however, may not capture differences in the distribution of learning among subjects. 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 have a positive impact on learning for subjects in the top half despite their lack of familiarity with Nonograms (0.02 with p-value of 0.29). Given that students in the top half improve by almost 27 seconds on average as reported in Appendix Table B6 in the Nonogram experiments compared to 9 seconds for top half subjects in the Sudoku experiments, "ceiling effects" are less likely to be an issue in the Nonogram experiments. 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.[35]

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 have any impact on 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 9 summarizes the estimates based on the outputs of equations [4] and [5] to formalize the inference from the visual inspection of Figure 9.[36] 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

---

[35] 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 (0.26) 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 on the top and the bottom halves improve by a similar amount. 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.

[36] The coefficient estimates from equations [4] and [5] are reported in Appendix Table B8.

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 9 shows that ability tracking reduces learning by 0.063 SD on average in the *teaching* treatment, but this estimated treatment 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.

To further investigate this possibility, Appendix Figure B3 reproduces Figure 9-B separately for subjects in the top half and subjects in the bottom half. Despite the small sample, the figure clearly shows that both groups of subjects are negatively affected by *tracking* in the *teaching* environment. These results are in contrast to the corresponding figures for Sudoku in Appendix Figure B2, which show that only subjects in the bottom half are negatively affected by *tracking* in the *teaching* treatment.

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 (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.[37] 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.[38]

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).[39] Subjects in the bottom half—regardless of whether they were in the *tracked* or *untracked*

---

[37] 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.

[38] An alternative analysis pooling the Sudoku and Nonogram estimates (facilitated by the use of a common, standardized outcome measure) yields estimates broadly similar to those in the Nonogram experiment (where ceiling effects are likely absent). The results of this meta-analysis are available from the authors upon request.

[39] For completeness, Columns (5)–(8) in Appendix Table B4 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 very imprecise given the small sample size (N=32).

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 performing peers available with the answers to these questions.

6. Conclusion

Despite being an obvious and pre-eminent mechanism for peer effects in the classroom (Sacerdote [2011]), peer-to-peer teaching and its effect have gone completely unstudied by economists. The reason for this apparent oversight is not altogether surprising: peer-to-peer teaching is difficult to observe and measure in the field and even more difficult (if not impossible) to exogenously vary.

Far more attention has been paid to the effects of ability tracking. Grouping students of similar abilities, however, undoubtedly has implications for the efficacy of peer-to-peer teaching insofar as it changes the distribution of potential peer teachers across classrooms. On the one hand, it has been suggested that ability tracking by prior achievement might disadvantage low-ability students to the extent that students benefit from having higher-ability peers to teach them (e.g. Epple et al. [2002]). On the other hand, ability tracking may encourage learning at all ability levels if students of more similar ability are more effective in teaching one another (Schunk [1991]).

Given these ambiguities, understanding the importance of peer-to-peer teaching and how ability tracking affects this teaching is crucial to attempts to improve learning by altering the composition of classrooms and the degree of interaction among students—interventions which are by themselves costless to schools. Existing evidence from field experiments on tracking is mixed—perhaps due to the fact that distinguishing between the various (and possibly counteracting) mechanisms through which tracking affects learning has proven very difficult even in randomized field settings.

To establish the importance of peer-to-peer teaching and identify the interaction between this teaching and ability tracking, we conduct a laboratory experiment to mimic a classroom environment. 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

29

down potential competing channels through which tracking may influence learning such as the responses of teachers to group composition.

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 this finding highlights the potentially sizable effect of peer-to-peer teaching among students, our study also suggests that the effects of these interactions are shaped by the composition of peer groups. Specifically, we show that the positive effect of peer-to-peer teaching on low-ability subjects is substantially offset when subjects are tracked by ability. This result implies that ability tracking based on prior achievement has the potential to disadvantage low-ability students who may miss out on interactions with high-ability peers who can teach them.

Our findings have important implications for school policy concerning the optimal composition of classrooms by ability levels. Specifically, tracking without appropriately tailored instruction may harm lower-ability students. Unless teachers also tailor the curriculum and instruction for low-ability students in tracked classrooms, schools should be cautious about implementing ability tracking.

Using a laboratory experiment allows us to estimate in a credible fashion the effect of peer-to-peer teaching—typically unmeasured in other settings. Establishing the potential for peer-to-peer teaching to enhance learning is a fundamentally important contribution to our knowledge of the education production function, but we acknowledge that our use of a laboratory has its limitations. Learning may not be best measured by our short-term measure of performance improvement, and the effects of peers may themselves depend on the context (e.g., math or language skills). These shortcomings, however, are shared in one way or another by most field studies.

Nonetheless, we are confident that laboratory experiments such as ours 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 the findings to differences in context, measures, 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 generating positive peer effects 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 for peer effects of having students who can serve as "bridges" between groups of students with different abilities. Experiments like these could additionally shed more 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.

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.

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.

Bhattacharya, Debopam. 2009. "Inferring Optimal Peer Assignment from Experimental Data." *Journal of the American Statistical Association*, 104: 486–500.

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.

Elsner, Benjamin, and Ingo Isphording. 2017. "A Big Fish in a Small Pond: Ability Rank and Human Capital Investment." *Journal of Labor Economics*, forthcoming.

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.

Koch, Alexander, Julia Nafziger, and Helena Skyt Nielsen. 2015. "Behavioral Economics of Education." *Journal of Economic Behavior and Organization*, 115: 3–17.

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.

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.

Lyle, David. 2009. "The Effects of Peer Group Heterogeneity on the Production of Human Capital at West Point." *American Economic Journal: Applied Economics*, 1: 69–84.

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.*

Vigdor, Jacob, and Thomas Nechyba. 2007. "Peer Effects in North Carolina Public Schools," in *Schools and the Equal Opportunity Problem*, eds. Woessmann, Ludger and Paul. E. Peterson, 73–101, MIT Press.

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)
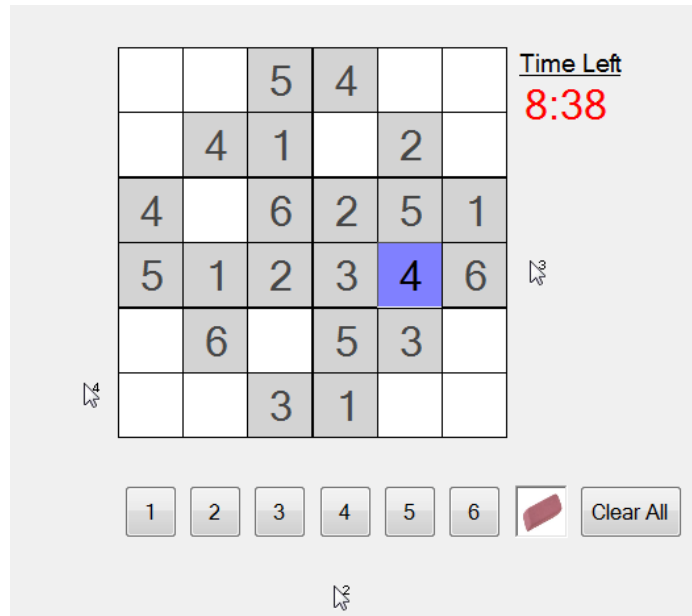
Notes: Shown from the perspective of the subject ranked 1 out of 4 in his group. Subjects are able to simultaneously edit a common 6×6 Sudoku puzzle during the Practice Block. Each other mouse arrow is labeled with the within-group performance rank of the person in the Ability Block (T=0). Performance 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
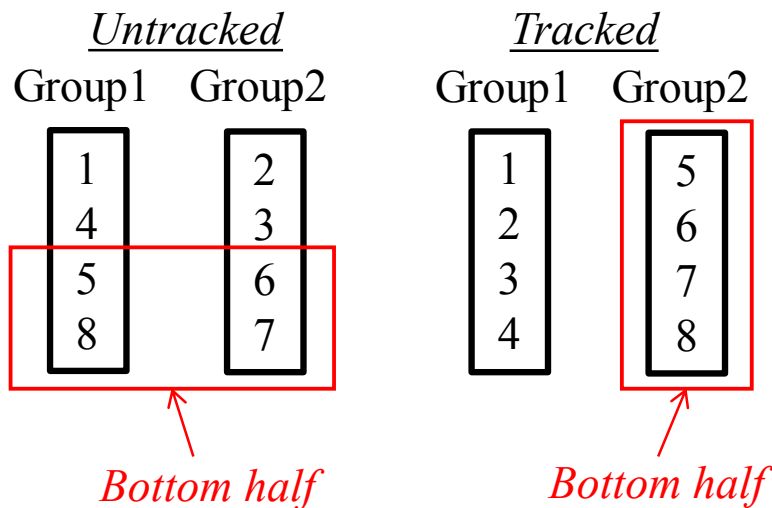
Notes: This figure describes the procedure for assigning subjects to groups in the *untracked* and *tracked* treatments. Rank is based on performance in the Ability Block (T=0). Performance is measured by the number of Sudoku puzzles solved with the average solving time serving as a tie-breaker. We define 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)**
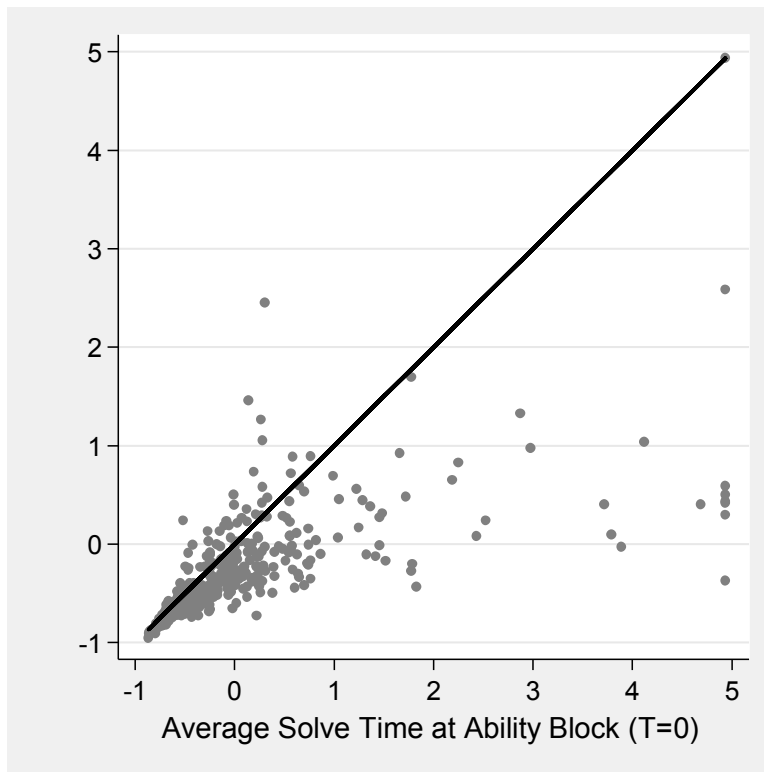
Notes: Scatter plots of the standardized average solving times (AST) in the Ability Block (T=0) and the Evaluation Block (T=1) are displayed. Raw average solving times (in seconds) for both blocks are standardized by the mean and standard deviation of average solving time at T=0 so that standardized 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 seconds, 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
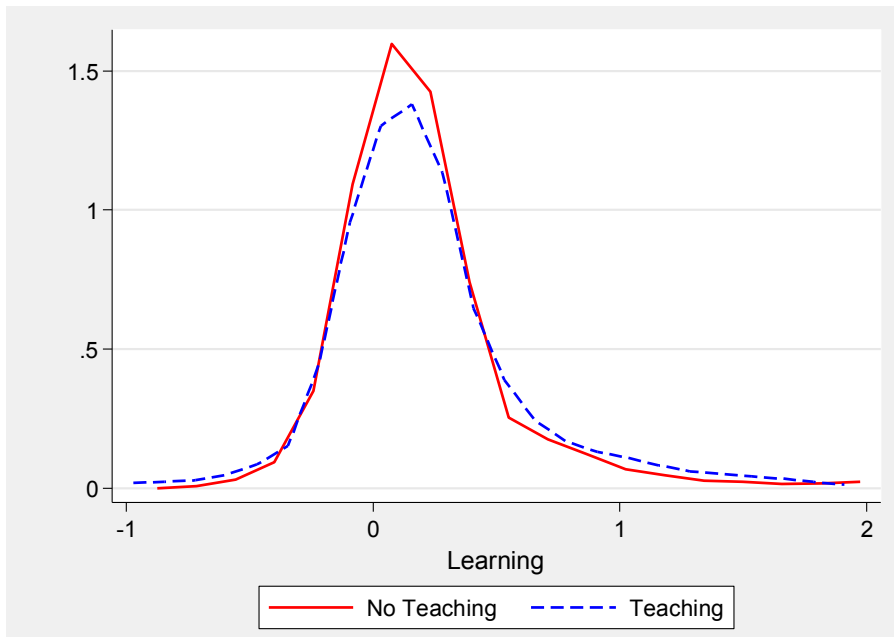
B. Among Subjects in the Bottom Half
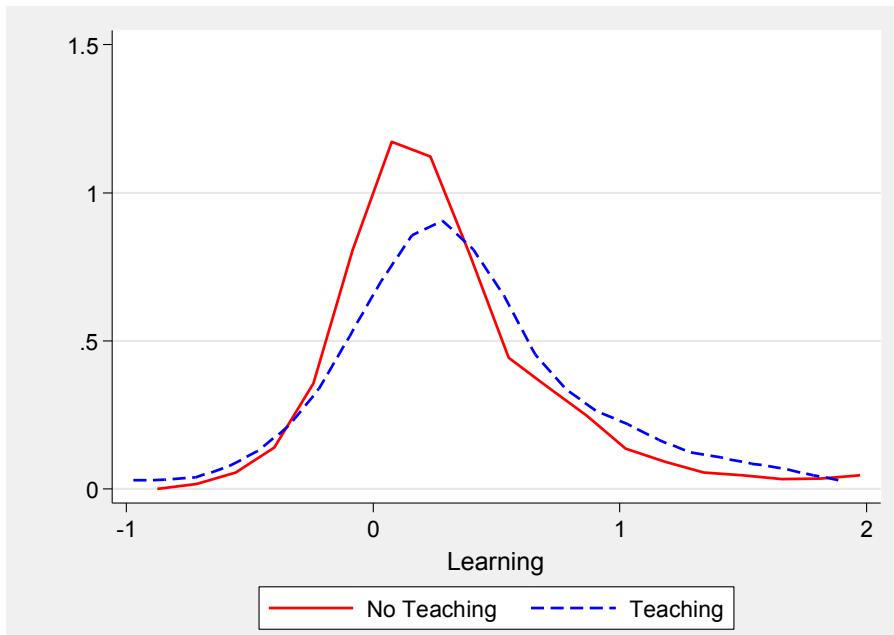
Notes: Kernel density plots of learning in the *no-teaching* and *teaching* treatments are displayed. Panel A is for the full sample while Panel B is restricted to subjects in the bottom half who ranked 5–8 in the Ability Block (T=0). Learning is calculated by subtracting the standardized average solving 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
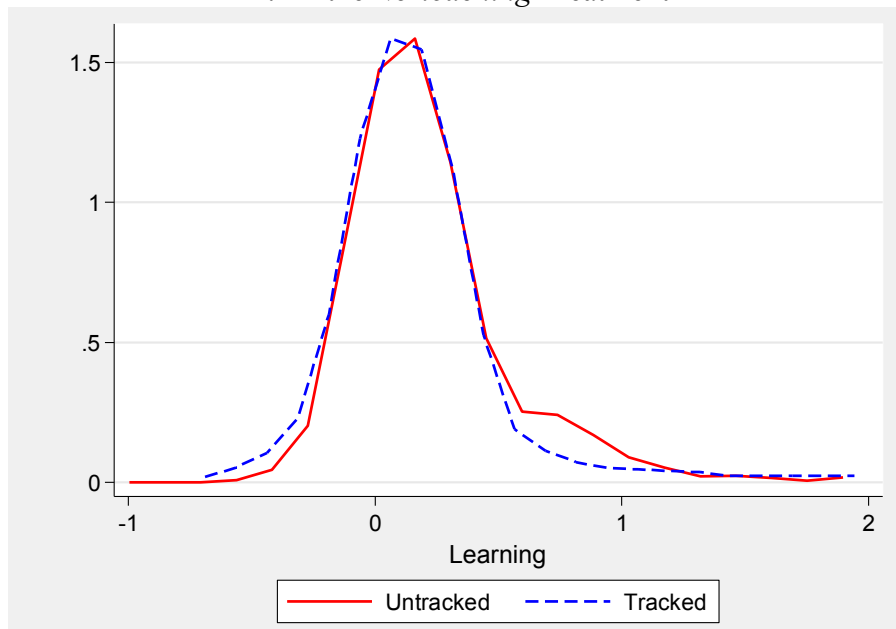
B. In the *Teaching* Treatment
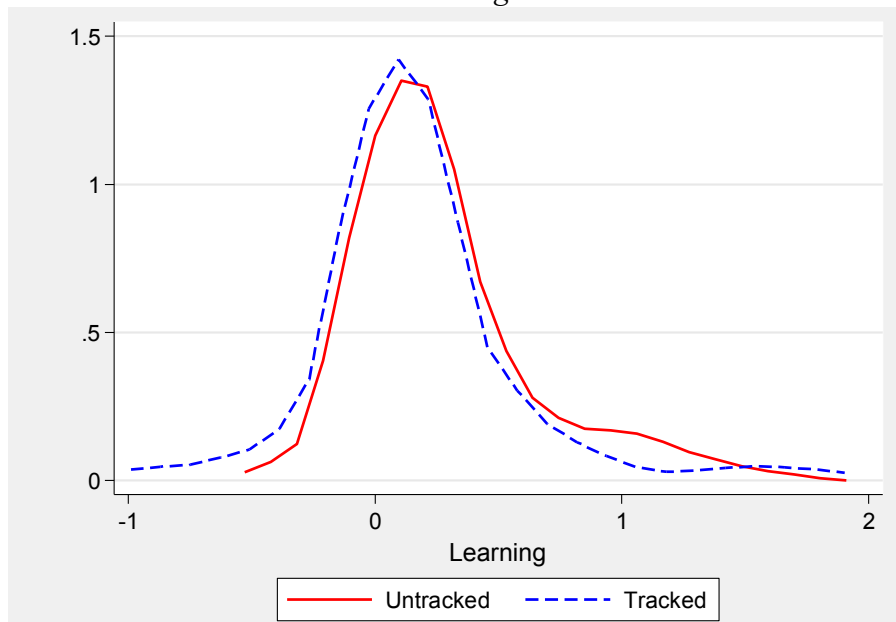
Notes: Kernel density plots of learning for the *untracked* and *tracked* treatments are displayed. Learning is calculated by subtracting the standardized average solving time in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in average solving 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)
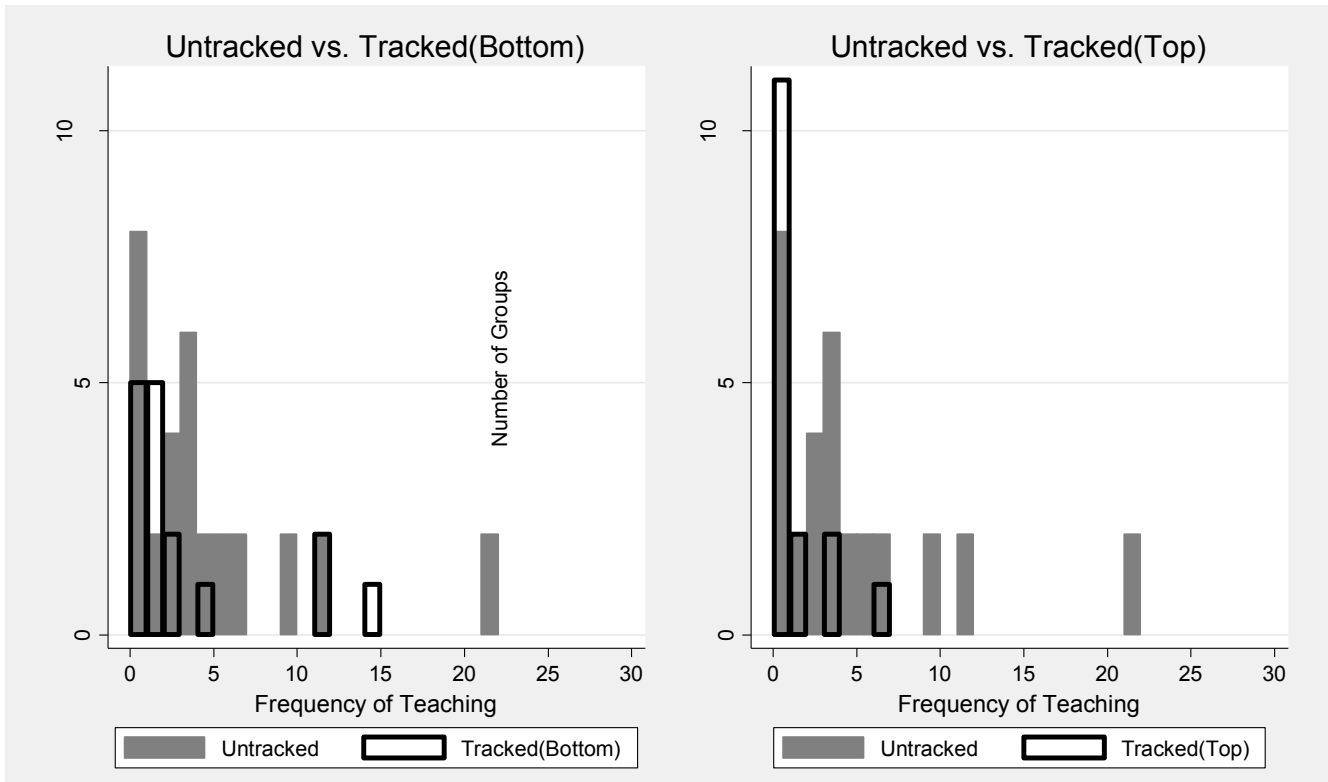
Notes: The unit of observation is a group. The sample is limited to the 32 *teaching* treatment sessions with 16 sessions each for the *untracked* and *tracked* treatments. In the *untracked* treatment, there are total of 32 groups (two groups for each session), while for the *tracked* treatment there 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
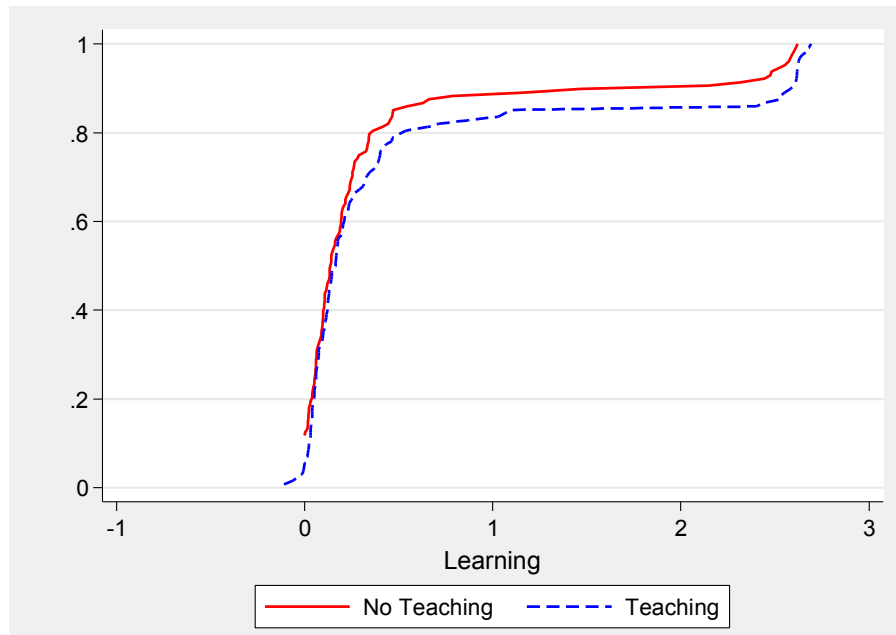
B. Among Subjects in the Bottom Half
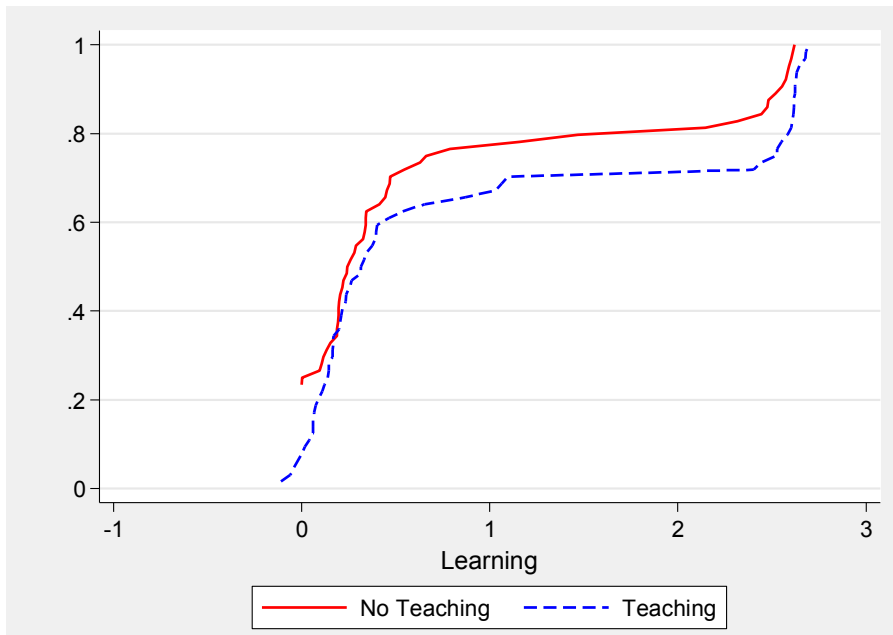
Notes: Cumulative distributions of learning in the *no-teaching* and *teaching* treatments are displayed. Panel A is for the full sample while Panel B is restricted to subjects in the bottom half who ranked 5–8 in the Ability Block (T=0). Learning is calculated by subtracting the standardized average 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.

41

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

A. In the *No-teaching* Treatment
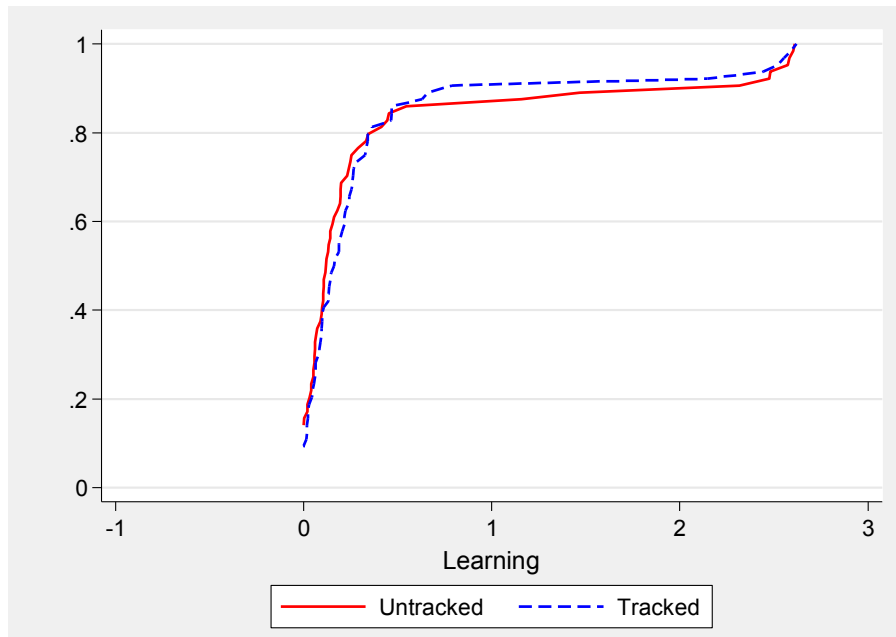
B. In the *Teaching* Treatment
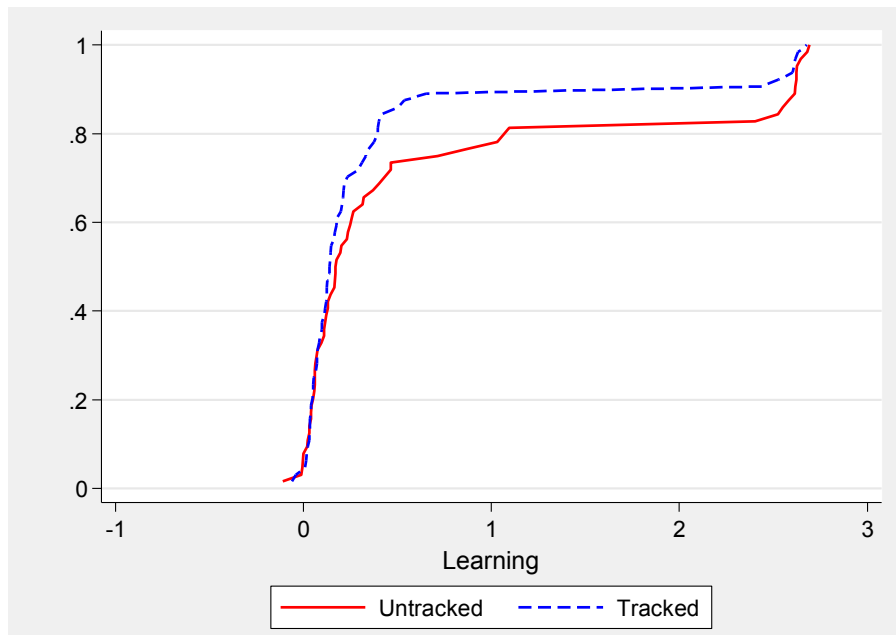
Notes: Cumulative distributions of learning in the *untracked* and *tracked* treatments are displayed. Learning is calculated by subtracting the standardized average solving time in the Evaluation Block (T=1) from that in the Ability Block (T=0), so that higher values indicate *improvement* in solving 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 C 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 C 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: Summary Statistics and Balance Tests (Sudoku)

	A. Overall			B. Heterogeneity		
	Mean	*p-value* of equality test		Bottom half (rank5-8)	Top half (rank1-4)	Dif (5)-(6)
		2×2×2	2×2			
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Male	0.47	*0.77*	*0.96*	0.52	0.42	0.10**
	[0.50]			[0.50]	[0.49]	(0.05)
Experienced	0.68	*0.65*	*0.70*	0.48	0.88	-0.39***
	[0.47]			[0.50]	[0.33]	(0.04)
Risk Attitude (0–9)	3.35	*0.35*	*0.17*	3.31	3.40	-0.09
	[1.64]			[1.71]	[1.58]	(0.15)
Prosociality (0–5)	1.58	*0.39*	*0.27*	1.70	1.47	0.23**
	[1.03]			[0.99]	[1.05]	(0.09)
Solved None at T=0	0.02	*0.23*	*0.11*	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	*0.31*	*0.15*	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 (=AST$_0$-AST$_1$) (*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	*0.31*	*0.15*	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 p-values for each variable in the far-left column of the null hypotheses that the means are equal across 8 treatment combinations (Column (2)) and 4 treatment combinations pooling across the incentive treatments (Column (3)). Columns (4) and (5) 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 (6) 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 C for details on the elicitation of risk attitudes and prosociality. 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

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 B3 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: Effect of *Tracking* on Logged Learning, Robustness Checks (Sudoku)

Outcome: Learning (logged)

	In the *No-teaching* Treatment		In the *Teaching* Treatment	
	(1)	(2)	(3)	(4)
A. Overall				
Tracked	-0.025		-0.081**	
	(0.035)		(0.040)	
B. Heterogeneity				
Tracked for Top half		0.022		-0.021
		(0.026)		(0.034)
Tracked for Bottom half		-0.068		-0.152**
		(0.060)		(0.068)
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 B4 for reference. Standard errors clustered at the group level are reported in parentheses. The outcome is logged learning, which is defined as the difference between logged average solving time (AST) in the Ability Block (T=0) and logged AST in the Evaluation Block (T=1). Note that learning is calculated by subtracting logged AST at T=1 from that at T=0, so that higher values indicate *improvement* in solving time. 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). The bottom half consists of those subjects ranked 5–8 in the Ability Block (T=0). There were 24 *no-teaching* sessions with 192 subjects, and 32 *teaching* sessions with 256 subjects. Each session consisted of two groups (4 subjects per group). Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$

Table 6: Frequency of Teaching in the *Tracked* vs. *Untracked* Treatments (Sudoku)

| | Untracked | Tracked | | Difference (2)–(1) | | | Difference (3)–(1) | | |
		Bottom Half	Top Half	OLS	Poisson	Zero-Inflated Poisson	OLS	Poisson	Zero-Inflated Poisson
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Teaching	4.78	3.13	0.94	-1.66	-0.43***	-0.53***	-3.84**	-1.63***	-1.00***
	[6.60]	[4.76]	[1.91]	(1.86)	(0.16)	(0.17)	(1.69)	(0.27)	(0.30)
Vuong test of Zero-Inflated model vs. Standard Poisson					z-score = 1.94			z-score = 2.06	
					p-value = 0.0263			p-value = 0.0198	
# of Groups	32	16	16	48	48	48	48	48	48
# of Sessions	16	16		32	32	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)–(6) report the estimated difference between Columns (1) and (2) from OLS, Poisson and Zero-inflated Poisson (where the inflation equation includes just a dummy for *tracked* sessions) models, respectively, with standard errors in parentheses. Columns (7)–(9) 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." Reported below Columns (5)–(6) and (8)–(9) are z-scores and p-values for the Vuong test of the Zero-Inflated Poisson model against a standard Poisson; these tests support the use of the Zero-Inflated model. Significance levels: *** p<0.01, ** p<0.05, * p<0.10

Table 7: 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 C 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 8: 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 9: 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

Appendix

(For Online Publication)

Appendix A: *Piece-rate* vs. *Tournament* Treatments

In this Appendix, we describe the results from our incentive scheme treatments and explain why we have not included tests of the hypotheses related to the incentive schemes in the main text. In short, our incentive structure in the *tournament* treatment seems not to have been salient enough to induce changes in effort compared to the *piece-rate* treatment.

Hypothesis A1 (Main Effect of *Tournament*): ambiguous

The effect of competition in the *tournament* treatment relative to the *piece-rate* treatment is theoretically ambiguous. In order to derive a comparative static prediction, we would need to know the optimal effort in the *piece-rate* and *tournament* treatment, but the optimal effort levels depend on the parameters (payoffs) as well as the unobserved costs of effort. As such, we cannot predict the effect of *tournament* treatment on learning relative to the *piece-rate* treatment.

Columns (1) and (2) in Pane A of Table A1 present the estimated treatment effects of competition on learning with the *piece-rate* treatment serving as the control treatment. There is no difference in learning between the two treatments even after controlling the individual covariates in Column (2).[44] Columns (3) and (4) present the estimated treatment effects by initial performance (rank). For both subjects in the top half and bottom half, these estimates are far from statistically significant at conventional levels and small in magnitude. Figure A1 confirms that the kernel densities of learning for the full sample, subjects in the bottom half, and subjects in the top half are almost identical in both treatments. The p-value for the two-sample Kolmogorov-Smirnov test for the equality of the distributions for the full sample (Panel A) is 0.618, while the p-value for subjects in bottom half (Panel B) is 0.982.

Hypothesis A2 (Interaction of *Tournament* and *Teaching*): negative

We also hypothesized that there would be interactions between the *tournament* and *teaching* treatments. Specifically, we expected that subjects would be teach each other less in the Practice Block when they anticipated competing against the other group members subsequently in the Evaluation Block—an effect which would be strongest when grouped with subjects of similar ability (i.e., in the *tracked* treatment). We found, however, no evidence of any such interaction effects. These and all of our estimates concerning the (non-)effects of the *tournament* treatment are available on request.

Overall, the *tournament* treatment does not seem to have affected learning. One possibility is that the payoff structure by rank in the *tournament* treatment was very nearly linear, and the payoffs were chosen so as to lead to similar earnings in the *piece-rate* treatment and the *tournament* treatment. As such, the tournament incentives may have been insufficient to induce significant changes in behavior. We find similar results in the Nonogram experiments presented in Panel B of Table A1.

---

[44] The raw means of learning in the *piece-rate* and *tournament* treatments are very similar (24.4 vs. 25.5 seconds).

Figure A1: Kernel Density of Learning by *Piece-rate* vs. *Tournament* (Sudoku)

A. Full Sample

B. Among Subjects in Bottom Half
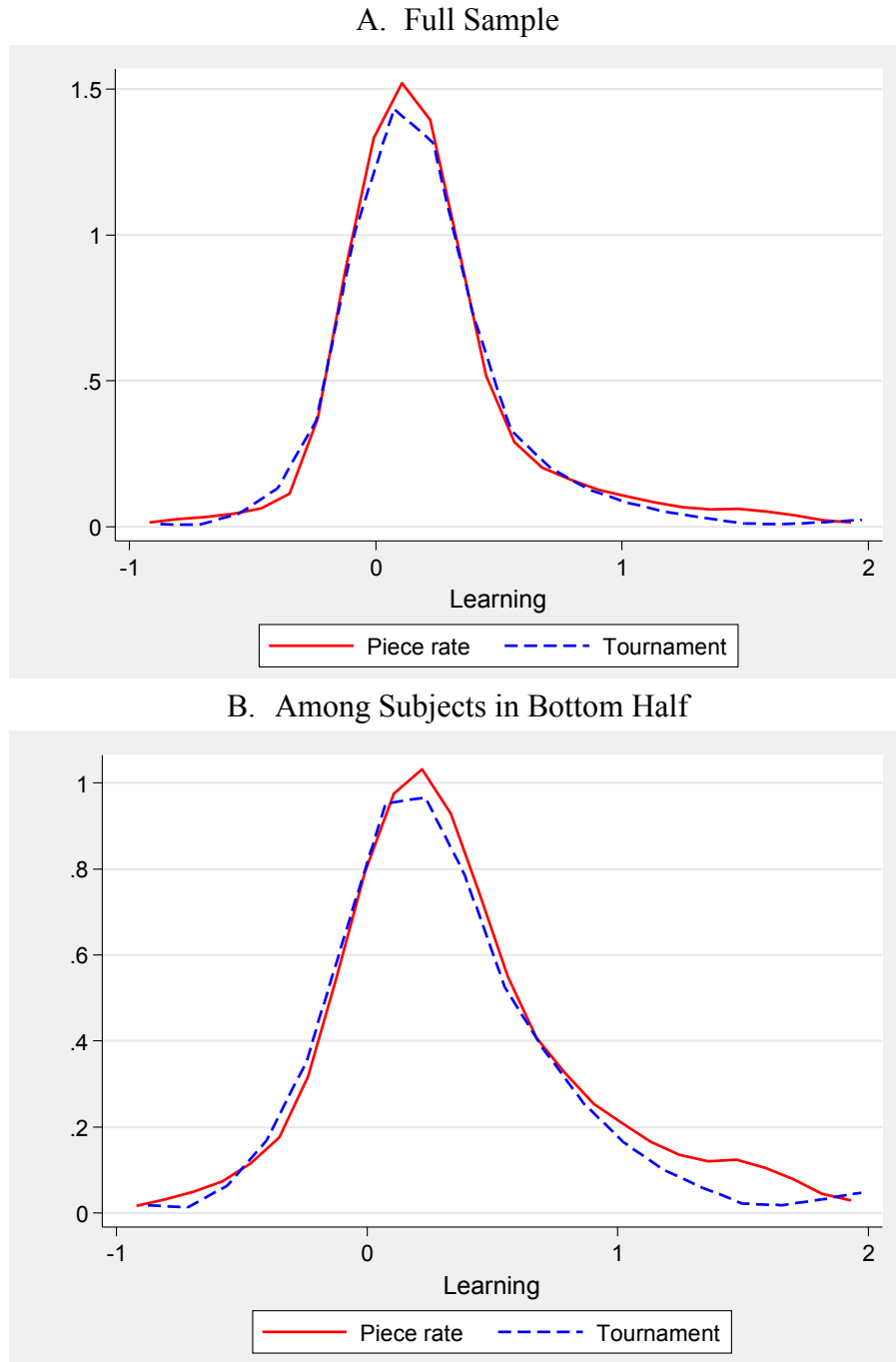
Notes: Kernel density plots of learning for the *piece-rate* and *tournament* treatments are displayed. Panel A uses the full sample, while Panel B is limited to subjects in bottom half who ranked 5–8 in the Ability Block (T=0). Learning is calculated by subtracting the average solving time (AST) in 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. For Panel A, the p-value for the two-sample Kolmogorov-Smirnov test for the equality of the distributions is 0.618, while that for Panel B is 0.982. There are 448 subjects.

Table A1: Effect of *Tournament* on Learning

A. Sudoku

	(1)	(2)	(3)	(4)
A. Overall				
Tournament	-0.006	-0.005		
	(0.056)	(0.054)		
B. Heterogeneity				
Tournament for Top half			-0.006	-0.005
			(0.056)	(0.054)
Tournament for Bottom half			-0.006	-0.005
			(0.056)	(0.054)
Controls	No	Yes	No	Yes

B. Nonograms

	(1)	(2)	(3)	(4)
A. Overall				
Tournament	0.006	0.003		
	(0.070)	(0.068)		
B. Heterogeneity				
Tournament for Top half			0.023	0.010
			(0.021)	(0.024)
Tournament for Bottom half			-0.012	-0.007
			(0.134)	(0.131)
Controls	No	Yes	No	Yes

Notes: Each column reports the estimated treatment effects from a different OLS regression. The estimates in Columns (1) and (2) are obtained by estimating equation [2] with the *teaching* dummy being replaced by a *tournament* dummy. The estimates in Columns (3) and (4) are obtained by estimating equation [3] with the *teaching* dummy being replaced by a *tournament* dummy (the original estimates available upon request). The control group is the *piece rate* treatment. The estimated treatment effects and their standard errors were computed using the *lincom* command in STATA. 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 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 subjects in bottom half are those subjects ranked 5–8 in the Ability Block (T=0) and the top half those ranked 1–4. 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). See Table 2 for definitions of each control variable. For Panel A (Sudoku), there were 28 sessions with 224 subjects (8 subjects per session) each for the *piece rate* and *tournament* treatments, respectively. For Panel B (Nonograms), there were 16 sessions with 128 subjects each for the *piece rate* and *tournament* treatments, respectively. Each session consisted of two groups (4 subjects per group). Significance levels: *** p<0.01, ** p<0.05, * p<0.10
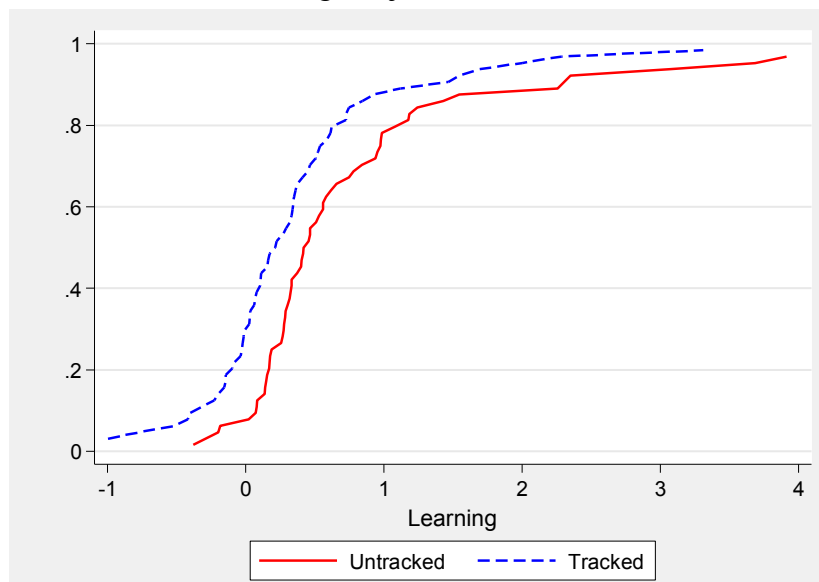
Appendix B: Additional Figures and Tables

Figure B1: Cumulative Distribution of Learning by *Tracking* under *Teaching* separately for Top and Bottom half (Sudoku)

A. Among Subjects in Top Half
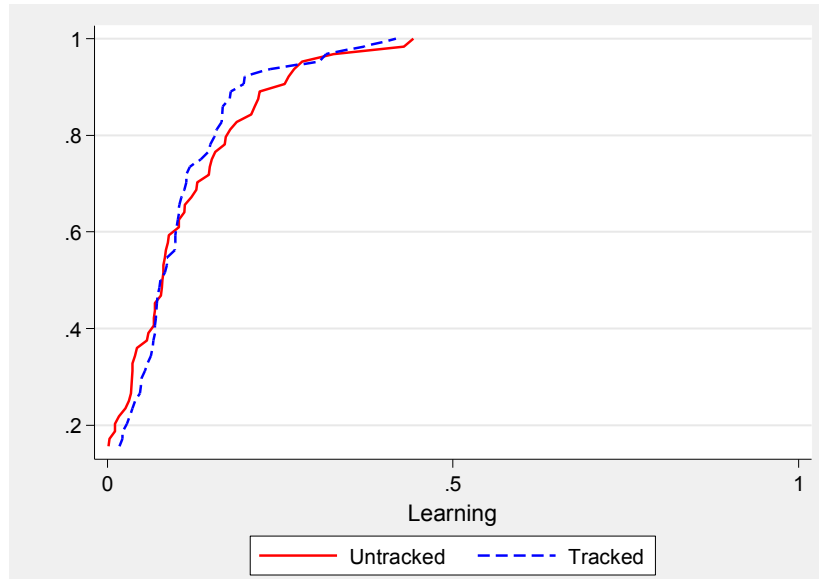
B. Among Subjects in Bottom Half

Notes: The cumulative distributions of learning reported in Figure 6-B are separately displayed for subjects in the top half (Panel A) and subjects in the bottom half (Panel B). Learning 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. Also note that the scales on the *x*-axes are different for the two graphs. The sample is limited to 32 *teaching* sessions with 256 subjects. For each session, there is one group for the top half and the bottom half; thus there are 32 groups each for the top half and bottom half. For Panel A, the p-value for the two-sample Kolmogorov-Smirnov test for the equality of the distributions in the *untracked* and *tracked* treatments is 0.843, while that for Panel B is 0.004.

Figure B2: Screenshot from the *Teaching* Treatment during the Practice Block (Nonograms)

Notes: Subjects are able to simultaneously edit a common 5×5 Nonogram puzzle during the Practice Block. Each 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 Nonogram puzzles solved with the average solving time serving as a tie-breaker. In the *no-teaching* treatment, the three arrows showing the within-group rank of the other subjects would not have been visible as each subject worked independently.

Figure B3: Cumulative Distributions of Learning by *Tracking* in the *Teaching* Treatment for Subjects in the Top and Bottom Half (Nonograms)
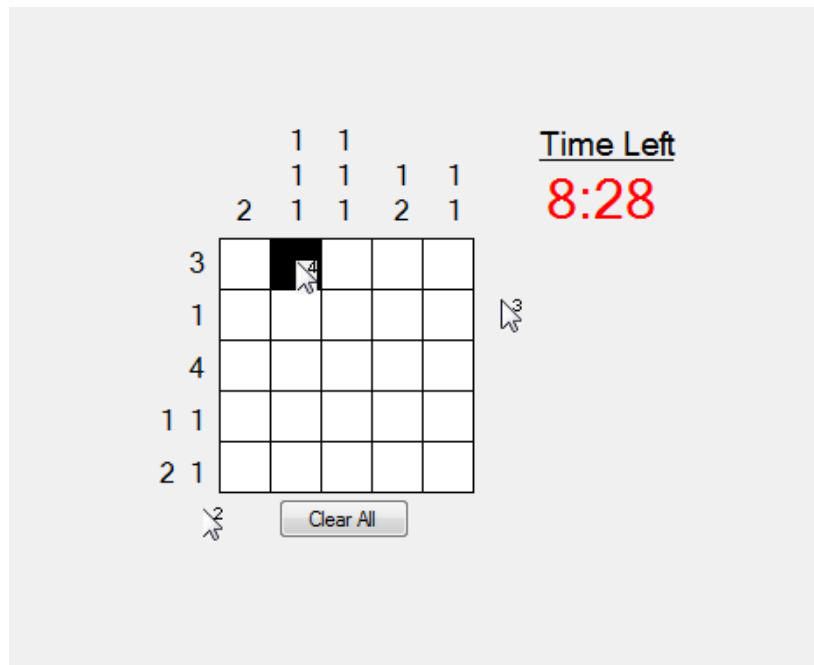
A. Among Subjects in the Top Half
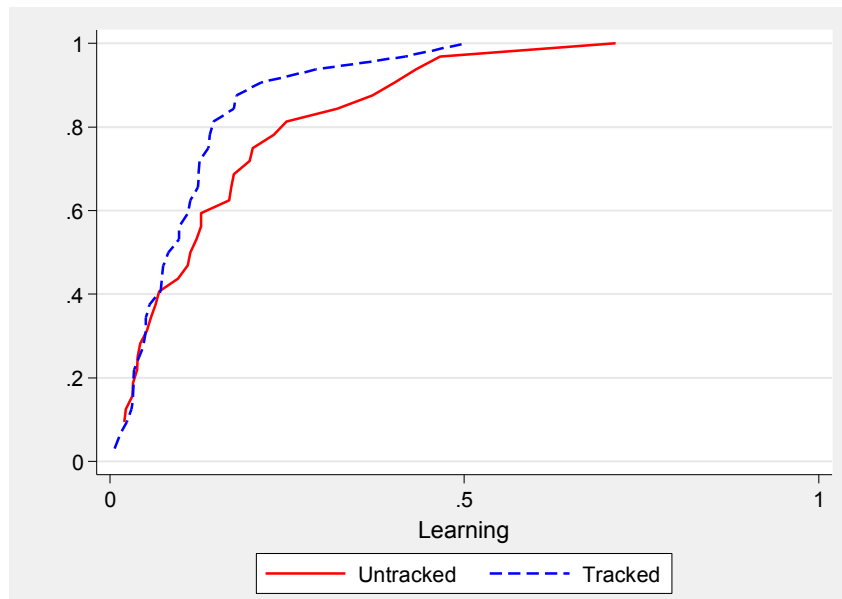
B. Among Subjects in the Bottom Half
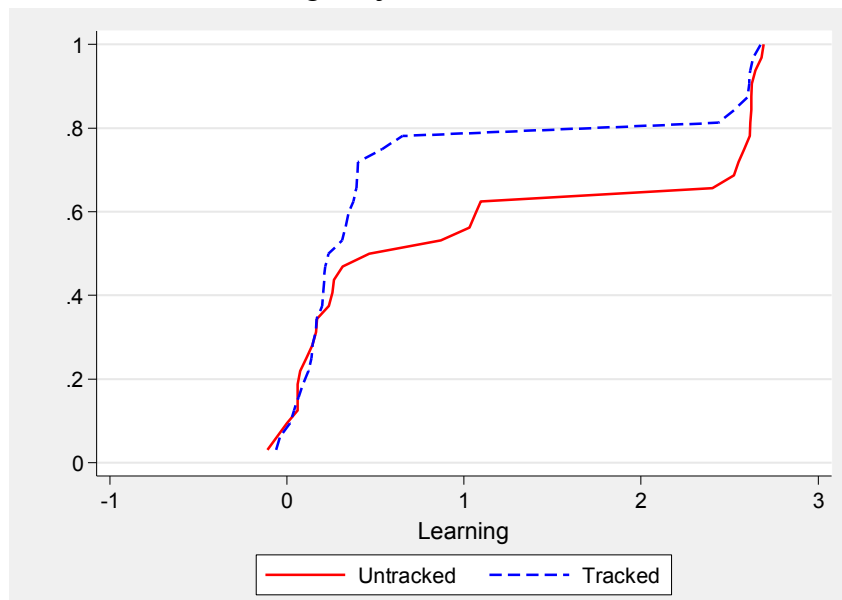
Notes: The cumulative distributions of learning reported in Figure 9-B are separately displayed for subjects in the top half (Panel A) and subjects in the bottom half (Panel B). Learning is calculated by subtracting the average solving time (AST) in Evaluation Block (T=1) from that in the Ability Block (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. Also, note that the scales on the *x*-axes are different between the two graphs. The sample is limited to 16 *teaching* sessions with 144 subjects. For each session, there is one group of subjects in the top half and one group of subjects in the bottom half; thus there are 16 groups each for the top half and bottom half. For Panel A, the p-value for the two-sample Kolmogorov-Smirnov test for the equality of the distributions in the *untracked* and *tracked* treatments is 0.434, while that for Panel B is 0.160.

Table B1: Coefficient Estimates for Equations [2] and [3]
to Estimate the Effect of *Teaching* on Learning (Sudoku)

Outcome: Learning

	A. Overall		B. Heterogeneity	
	(1)	(2)	(3)	(4)
Teaching	0.113**	0.119**	-0.018	-0.016
	(0.052)	(0.052)	(0.017)	(0.020)
Bottom Half			0.258***	0.256***
			(0.096)	(0.095)
Teaching × Bottom Half			0.185***	0.167***
			(0.053)	(0.052)
Male		0.041		0.007
		(0.059)		(0.057)
Experienced		-0.203***		-0.066
		(0.076)		(0.069)
Risk Attitudes (0–9)		0.010		0.015
		(0.020)		(0.019)
Prosociality (0–5)		-0.007		-0.024
		(0.030)		(0.031)
None Correct at T=0	3.528***	3.416***	3.385***	3.356***
	(0.386)	(0.395)	(0.386)	(0.387)
Constant	0.191***	0.286**	0.102***	0.142
	(0.028)	(0.128)	(0.010)	(0.112)
Controls	No	Yes	No	Yes
R-squared	0.38	0.40	0.44	0.44
# of Sessions	56	56	56	56
# of Groups	112	112	112	112
# of Subjects	448	448	448	448

Notes: Each column reports the results from a different OLS regression. 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 Evaluation Block (T=1) from that of 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. Experienced takes a value of one if a subject indicates experience with Sudoku puzzles before the experiment. Risk attitudes take on the values from 0 to 9 with higher number indicating higher risk-aversion. Prosociality takes on the values from 0 to 5 with higher number indicating higher prosociality. See Appendix C for details on the elicitation of risk attitude and prosociality. We also include a dummy for the eight subjects who could not solve any Sudoku puzzles in the Ability Block (T=0). There were 56 sessions with 448 subjects (8 subjects per session) in the experiment. The estimates were used to produce Table 3 in the main text. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$

Table B2: Coefficient Estimates for Equations [4] and [5] (Sudoku)

Outcome: Learning

	A. Overall	B. Heterogeneity
	(1)	(2)
Tracked	-0.039	0.024
	(0.061)	(0.026)
Teaching	0.171**	0.012
	(0.073)	(0.024)
Tracked × *Teaching*	-0.105	-0.056
	(0.101)	(0.044)
Bottom half		0.231***
		(0.085)
Tracked × Bottom half		-0.119
		(0.107)
Teaching × Bottom half		0.322**
		(0.127)
Tracked × *Teaching* × Bottom half		-0.132
		(0.183)
Controls	Yes	Yes
R-squared	0.41	0.45
# of Sessions	56	56
# of Groups	112	112
# of Subjects	448	448

Notes: Each column reports the results from a different OLS regression. The outcome is learning, which is calculated by subtracting the average solving time (AST) in 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 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 subjects who could not solve any Sudoku puzzles at T=0. The bottom half consists of those subjects ranked 5–8. 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). The estimates were used to produce Table 4 in the main text. Significance levels: *** p<0.01, ** p<0.05, * p<0.10

Table B3: Effect of *Teaching* on Logged Learning, Robustness Checks (Sudoku)

Outcome: Learning (logged)

	(1)	(2)	(3)	(4)
A. Overall				
Teaching	0.030	0.031		
	(0.026)	(0.027)		
B. Heterogeneity				
Teaching for Top half			-0.021	-0.024
			(0.021)	(0.022)
Teaching for Bottom half			0.080*	0.079*
			(0.046)	(0.046)
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 the equation [3] with and without controls using the 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 not reported to save space (results available upon request). Standard errors clustered at the group level are reported in parentheses. The outcome is logged learning, which is defined as the difference between logged average solving time (AST) in the Ability Block (T=0) and the Evaluation Block (T=1). The bottom half consists of those subjects ranked 5–8 and the top half those subjects ranked 1–4 in 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 at 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. 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 B4: Group Mean, Group Standard Deviation, and Teaching Frequencies (Sudoku)

	A. Sudoku				B. Nonograms			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Group mean	0.36***		0.02	0.28	0.20		0.04	1.54
	(0.10)		(0.21)	(0.32)	(0.16)		(0.35)	(0.98)
Group SD		0.36***	0.34**	0.40**		0.22	0.18	-0.42
		(0.09)	(0.17)	(0.18)		(0.16)	(0.35)	(0.50)
Group mean × Group SD				-0.18				-1.34
				(0.17)				(0.84)
Vuong test of Zero-Inflated model vs. Standard Poisson	z=3.61 p=0.0002	3.57 0.0002	3.53 0.0002	3.31 0.0005	2.91 0.0018	2.83 0.0023	2.74 0.0031	2.83 0.0023
# of Groups	64	64	64	64	30	30	30	30

Notes: Each column reports the results from a different zero-inflated Poisson regression in which the number of teaching statements recorded for a given group is the dependent variable. Group mean is the group average of *raw* average solving time (AST) in the Ability Block (T=0) so that higher values indicate that the group consisted of lower-ability subjects while lower values indicate that the group consisted of higher-ability subjects. Because this is a group-level analysis, the sample size for Sudoku in Panel A is 64, while that for Nonograms in Panel B is 30 as there was one Nonogram session (two groups) in which the audio recording did not work. The Vuong tests of the zero-inflated model against standard Poisson models are reported in each column with the z-scores and corresponding p-values. All of these tests support the use of the zero-inflated model.

Table B5: Coefficient Estimates for Equations [4] and [5] Using Logged Learning, Robustness Checks (Sudoku)

Outcome: Learning (Logged)

	A. Overall	B. Heterogeneity
	(1)	(2)
Tracked	-0.025	0.022
	(0.035)	(0.026)
Teaching	0.059	-0.002
	(0.039)	(0.026)
Tracked × *Teaching*	-0.056	-0.044
	(0.053)	(0.044)
Bottom half		0.137**
		(0.055)
Tracked × Bottom half		-0.091
		(0.062)
Teaching × Bottom half		0.123*
		(0.066)
Tracked × *Teaching* × Bottom half		-0.040
		(0.095)
Controls	Yes	Yes
R-squared	0.17	0.23
# of Sessions	56	56
# of Groups	112	112
# of Subjects	448	448

Notes: Each column reports the results from a different OLS regression. The outcome is logged learning, which is defined as the difference between the logged average solving time (AST) in the Ability Block (T=0) and the Evaluation Block (T=1). 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 subjects who could not solve any Sudoku puzzles at T=0. The bottom half consists of those subjects ranked 5–8. 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). The estimates were used to produce Table 5 in the main text. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.10$

Table B6: Summary Statistics and Balance Tests (Nonograms)

	A. Overall			B. Heterogeneity		
	Mean	*p-value* of equality test		Bottom half	Top half	Dif
		2×2×2	2×2	(rank5-8)	(rank1-4)	(5)-(6)
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Male	0.45	*0.60*	*0.43*	0.41	0.49	-0.09
	[0.52]			[0.49]	[0.55]	(0.07)
Experienced	0.02	*0.33*	*0.57*	0.00	0.03	-0.03**
	[0.12]			[0.00]	[0.17]	(0.02)
Risk Attitude (0–9)	3.63	*0.12*	*0.14*	3.63	3.63	0.00
	[1.60]			[1.60]	[1.60]	(0.18)
Prosociality (0–5)	1.79	*0.28*	*0.13*	1.85	1.72	0.13
	[0.97]			[0.99]	[0.94]	(0.11)
Solved None at T=0	0.20	*0.47*	*0.18*	0.39	0.00	0.39***
	[0.40]			[0.49]	[0.00]	(0.04)
Solved None at T=1	0.07	-	-	0.13	0.00	0.13***
	[0.25]			[0.34]	[0.00]	(0.03)
Raw Average solve time at T=0 (*sec*)	186.47	*0.26*	*0.14*	307.31	65.64	241.66***
	[209.92]			[240.90]	[31.91]	(21.47)
Raw Average solve time at T=1 (*sec*)	85.38	-	-	132.37	38.40	93.97***
	[140.16]			[186.77]	[10.46]	(16.23)
Raw Learning (=AST0-AST1) (*sec*)	101.09	-	-	174.94	27.25	147.69***
	[170.62]			[216.42]	[25.00]	(19.02)
Standardized average solve time at T=0	0.00	*0.26*	*0.14*	0.58	-0.58	1.15***
	[1.00]			[1.15]	[0.15]	(0.10)
Standardized average solve time at T=1	-0.48	-	-	-0.26	-0.71	0.45***
	[0.67]			[0.89]	[0.05]	(0.08)
Learning	0.48	-	-	0.83	0.13	0.70***
	[0.81]			[1.03]	[0.12]	(0.09)
# of Sessions	32			32	32	
# of Groups	64			32	32	
# of Subjects	256			128	128	

Notes: Column (1) reports means for the full sample with standard deviations in brackets. Columns (2) and (3) 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 (2)) and 4 treatment combinations pooling across the incentive treatments (Column (3)). Columns (4) and (5) 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 (6) 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. See the notes for Table 2 for descriptions of the variables. The experienced variable equals one if a subject reported prior experience with Nonograms and zero otherwise. There were total of 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 B7: Coefficient Estimates for Equations [2] and [3]
for the Effect of *Teaching* on Learning (Nonograms)

Outcome: Learning

	A. Overall		B. Heterogeneity	
	(1)	(2)	(3)	(4)
Teaching	0.203***	0.195***	0.021	0.022
	(0.064)	(0.064)	(0.022)	(0.021)
Bottom Half			-0.018	-0.006
			(0.077)	(0.081)
Teaching × Bottom Half			0.358***	0.337***
			(0.121)	(0.121)
Male		0.040		0.039
		(0.071)		(0.072)
Experienced		-0.279***		-0.166***
		(0.062)		(0.039)
Risk Attitudes (0–9)		0.003		0.002
		(0.025)		(0.026)
Prosociality (0–5)		-0.042		-0.038
		(0.041)		(0.041)
None Correct at T=0	0.091**	0.145	0.119***	0.163
	(0.036)	(0.152)	(0.011)	(0.148)
Constant	1.477***	1.481***	1.390***	1.394***
	(0.172)	(0.171)	(0.180)	(0.179)
Controls	No	Yes	No	Yes
R-squared	0.53	0.53	0.55	0.55
# of Sessions	32	32	32	32
# of Groups	64	64	64	64
# of Subjects	256	256	256	256

Notes: Each column reports the results from a different OLS regression. 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 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 subjects who could not solve any Nonogram puzzles at T=0. The bottom half consists of those subjects ranked 5–8 in the Ability Block (T=0). There were 16 *no-teaching* and *teaching* sessions with 128 subjects, respectively (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 B8: Coefficient Estimates for Equations [4] and [5] (Nonograms)

Outcome: Learning

	A. Overall	B. Heterogeneity
	(1)	(2)
Tracked	0.067	0.024
	(0.095)	(0.044)
Teaching	0.260***	0.066**
	(0.093)	(0.029)
Tracked × *Teaching*	-0.131	-0.087*
	(0.128)	(0.050)
Bottom half		-0.040
		(0.129)
Tracked × Bottom half		0.071
		(0.186)
Teaching × Bottom half		0.381**
		(0.187)
Tracked × *Teaching* × Bottom half		-0.089
		(0.243)
Controls	Yes	Yes
R-squared	0.53	0.55
# of Sessions	32	32
# of Groups	64	64
# of Subjects	256	256

Notes: Each column reports the results from a different OLS regression. 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 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 subjects who could not solve any Nonogram puzzles at T=0. The bottom half consists of those subjects ranked 5–8 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 B9: Frequency of Teaching in the *Tracked* vs. *Untracked* Treatments (Nonograms)

	Untracked	*Tracked* Bottom Half	Top Half	Difference (2)–(1) OLS	Poisson	Zero-Inflated Poisson	Difference (3)–(1) OLS	Poisson	Zero-Inflated Poisson
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Teaching	4.14	5.75	4.00	1.61	0.33*	0.06	-0.14	-0.04	-0.31
	[5.26]	[7.13]	[2.98]	(2.65)	(0.20)	(0.20)	(2.04)	(0.22)	(0.22)
Vuong test of Zero-Inflated model vs. Standard Poisson				z-score = 2.64 p-value = 0.0041			z-score = 2.66 p-value = 0.0039		
# of Groups	14	8	8	22	22	22	22	22	22
# of Sessions	7	8		15	15	15	15	15	15

Notes: The unit of observation is a group. The sample is limited to the 15 *teaching* treatment sessions with 7 sessions for the *untracked* and 8 sessions for *tracked* treatments as there was one *untracked* session in which the audio recording did not work. For the *untracked* treatment, there are total of 14 groups (two groups for each session), while for the *tracked* treatment there are 8 groups each for subjects in bottom half (Group 2 in the *tracked* treatment in Figure 2) and for those in top half (Group 1 in the *tracked* treatment in Figure 2). Column (1) reports the mean number of teaching statements in the *untracked* treatment, and Columns (2) and (3) report the means for the *tracked* treatment for the bottom half group and the top half group, respectively. Standard deviations are reported in brackets. Columns (4)–(6) report the estimated difference between Columns (1) and (2) from OLS, Poisson and zero-inflated Poisson (where the inflation equation includes just a indicator for whether the group was *tracked*) models, respectively, with standard errors in parentheses. Columns (7)–(9) report the corresponding estimated differences between Columns (1) and (3). The bottom half consists of those subjects ranked 5–8 in the Ability Block (T=0) and the top half consists of those subjects ranked 1–4. A teaching statement is defined to be any utterance in which subjects are engaged in trying to teach each other how to do Nonograms 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. All of the tests support the use of the zero-inflated model. Significance levels: *** p<0.01, ** p<0.05, * p<0.10

Appendix C: Experiment Instructions

Below we include the instructions for the experiment. Notes for the reader (not visible to subjects) indicate page breaks. These notes are contained in [brackets] and explain which instructions were viewed by subjects in each treatment. The instructions below are for the Sudoku experiment. The instructions for the Nonogram experiment were identical with "Nonogram" appearing everywhere "Sudoku" appears below; the only exception is that the portion of the instructions explaining the rules of the game varied accordingly, as indicated below.

[Introduction – All treatments]

You are now participating in a decision-making experiment. If you follow the instructions carefully, you can earn a considerable amount of money depending on your decisions and the decisions of the other participants. Your earnings will be paid to you in CASH at the end of the experiment.

Today's experiment will involve multiple tasks, and your total payment will be the sum of your payments from each task. Please do not talk to other subjects during the experiment. This set of instructions is for your private use only. In case of questions, please raise your hand. Then we will come to you and answer your questions privately.

[Dictator Game – All treatments]

In Part I of this experiment, you are matched with one other person in this room.

Each person begins with 5 Dollars. Each person chooses how to allocate this money between him/herself and the person he/she is paired with. To specify an allocation, please type the amount you want to allocate to yourself and the amount you want to allocate to the other person and then click "Next." The two amounts must sum up to 5 Dollars.

After everyone has chosen their allocation, the computer will randomly choose one person from each pair whose decision is implemented. Your payment for Part I will be based on the randomly chosen person's decisions.

For example, if your decision is randomly chosen to be implemented, then you will be paid according to your allocation. If instead, the other person's decision is chosen to be implemented, you will be paid according to their allocation.

This is the end of the instructions for Part I. If you have any questions, please raise your hand and an experimenter will answer them privately.

[Risk Preference – All treatments]

In this task, you will make a series of choices between two uncertain options. For each decision, all you have to do is indicate whether you prefer Option A or Option B and click the appropriate button with your mouse.

At the end, the computer will randomly pick ONE decision and then randomly draw a whole number between 1 and 100 to determine your payoff. Your payoff then depends on your choice of A or B in the randomly chosen decision and the number drawn by the computer.

If you choose option A, then numbers between 1 and 50 will always pay $1 and numbers between 51 and 100 will always pay $3.

If you choose option B, then the relationship between the number drawn and the payment is different in each decision, so make sure you carefully compare options A and B to make your choices.

In the first decision, Option B will pay $0 on numbers between 1 and 90 and $3 on numbers between 91 and 100. As you proceed through the decisions, Option B will change. The chance of receiving $3 will increase, while the chance of receiving $0 will decrease.

Remember, after you make all of your choices, the computer will randomly pick one of them to count and will draw a random number to determine your payment.

This means that each decision could be the decision-that-counts so it is in your interest to treat each decision as if it could be the one that determines your payment.

We will reveal the outcome of this task at the end of today's experiment, just before you are paid.

[Risk Preference Decision – All treatments]

Choose whether you prefer Option A or Option B.

At the end of the experiment, the computer will choose one of these decisions for actual payment and will then draw a random number between 1 and 100, which together with your choices, determines how much you get paid.

Remember, this could be the decision that counts, so consider your choice carefully!

[Sudoku General – All treatments]

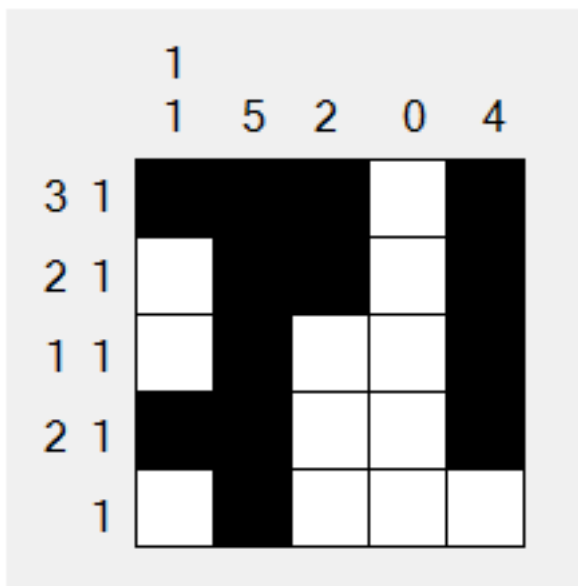The next task will involve solving puzzles known as Sudokus.

A 6×6 sudoku puzzle is a grid 6 squares wide and 6 squares deep. The lines of squares running horizontally are called rows, and the lines running vertically are called columns. The grid is further divided by the darker lines into six 2×3 rectangular 'boxes'.

Some of the squares already have numbers in them. Your task is to fill in the blank squares.

There's only one rule: Each row, column and box must end up containing all of the numbers from 1 to 6. This rule has an important side-effect, which is the basis of all solving techniques: Each number can only appear once in a row, column or box.

[Nonograms General – All treatments]

The next task will involve solving puzzles known as Nonograms.

A 5×5 Nonogram puzzle is a grid 5 squares wide and 5 squares deep. The lines of squares running horizontally are called rows, and the lines running vertically are called columns.

Each square must be either filled in black or left blank (or, equivalently, marked with an X).

Beside each row of the grid are listed the lengths of the runs of black squares on that row.

For example, the top row of the grid says 3,1. This means the row must contain a run of 3 consecutive black squares and a single black square.

Above each column are listed the lengths of the runs of black squares in that column.

For example, the leftmost column says 1, 1. This means the column must contain two (and only two) single black squares that are not consecutive.

Your aim is to find all black squares. Left click on a square to make it black. Left click a second time to mark it with an X. Left click a third time to turn it white again.

The puzzle is correct when all the rows and columns have the correct number of black squares in the proper positions.

[Sudoku Video – All treatments]

Before the experiment begins, we will watch the following video. Please pay attention to the video as you may find its content useful for the next portion of the experiment for which you will be paid.

This video describes strategies for solving 9×9 Sudokus, but they apply equally to 6×6 Sudokus.

[Ability Block – All treatments]

You will now be given 20 Sudoku puzzles to solve. You will have 10:00 minutes. You will be paid $0.5 for each puzzle that you correctly complete and $0 for each puzzle you do not complete or complete incorrectly. There is a timer in the upper right hand portion of the screen for your reference. Use the mouse to select cells. Numbers may be entered by clicking the buttons on screen or with the numbers on the keyboard. If you finish before the time limit, please wait quietly until the next puzzle begins.

[Review - *tracked* **and** *piece-rate***]**

Now we will begin task 4. You will participate in two periods where you will attempt to solve more sudoku puzzles. The first of these periods will be for practice and the second will be for payment which will be added to your earnings at the end of the experiment.

Based on your performance in Part I, you have been sorted into groups. The top half performers are in one group and the bottom half performers are in the other group. The highest performers are those who correctly completed the most puzzles, with ties broken in favor of those who completed them the fastest

In the second period, you will be paid $0.5 for each puzzle that you correctly complete and $0 for each puzzle you do not complete or complete incorrectly.

[Review - *tracked* **and** *tournament***]**

Now we will begin task 4. You will participate in two periods where you will attempt to solve more sudoku puzzles. The first of these periods will be for practice and the second will be for payment which will be added to your earnings at the end of the experiment.

Based on your performance in Part I, you have been sorted into groups. The top half performers are in one group and the bottom half performers are in the other group. The highest performers are those who correctly completed the most puzzles, with ties broken in favor of those who completed them the fastest

In the second period, your payment will be determined by your RELATIVE performance in your group. In other words, at the end of the second period we will count the number of puzzles completed correctly by each person, and pay will be based on your rank.

The participant in rank 1 will be paid $20
The participant in rank 2 will be paid $10
The participant in rank 3 will be paid $5
The participant in rank 4 will be paid $0

Ties will be broken in favor of those who completed the puzzles the fastest

[Review - *untracked* and *piece-rate*]

Now we will begin task 4. You will participate in two periods where you will attempt to solve more sudoku puzzles. The first of these periods will be for practice and the second will be for payment which will be added to your earnings at the end of the experiment.

Based on your performance in Part I, you have been sorted into groups. These groups have been balanced so that they contain both high and low performers. The highest performers are those who correctly completed the most puzzles, with ties broken in favor of those who completed them the fastest.

In the second period, you will be paid $0.5 for each puzzle that you correctly complete and $0 for each puzzle you do not complete or complete incorrectly.

[Review - *untracked* and *tournament*]

Now we will begin task 4. You will participate in two periods where you will attempt to solve more sudoku puzzles. The first of these periods will be for practice and the second will be for payment which will be added to your earnings at the end of the experiment.

Based on your performance in Part I, you have been sorted into groups. These groups have been balanced so that they contain both high and low performers. The highest performers are those who correctly completed the most puzzles, with ties broken in favor of those who completed them the fastest.

In the second period, your payment will be determined by your RELATIVE performance in your group. In other words, at the end of the second period we will count the number of puzzles completed correctly by each person, and pay will be based on your rank.

The participant in rank 1 will be paid $20
The participant in rank 2 will be paid $10
The participant in rank 3 will be paid $5
The participant in rank 4 will be paid $0

Ties will be broken in favor of those who completed the puzzles the fastest.

[Practice Block - *no-teaching*]

You will now be given a single Sudoku puzzle to solve. You will not be paid for this puzzle. You will have 10:00 minutes to solve the puzzle.

[Practice Block - *teaching*]

You will now be given a single Sudoku puzzle to solve. You will not be paid for this puzzle. You will have 10:00 minutes to solve the puzzle.

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.

[Evaluation Block - *piece-rate*]

You will now be given 30 Sudoku puzzles to solve. You will have 15:00 minutes.

Remember, you will be paid $0.5 for each puzzle that you correctly complete, and you will be paid $0 for each puzzle you do not complete or complete incorrectly. At the end of the 15:00 minutes, you will learn the number of puzzles you correctly completed.

You will also learn your rank among the other subjects in terms of the number of correctly completed items in the previous 15:00 minutes.

[Evaluation Block - *tournament*]

You will now be given 30 Sudoku puzzles to solve. You will have 15:00 minutes.

At the end of this task, you will learn your rank among the other subjects in terms of the number of correctly completed items. Remember, in this task, your payment will be determined by your RELATIVE performance in your group. In other words, at the end of this task we will count the number of puzzles completed correctly by each person, and pay will be based on your rank

The participant in rank 1 will be paid $20
The participant in rank 2 will be paid $10
The participant in rank 3 will be paid $5
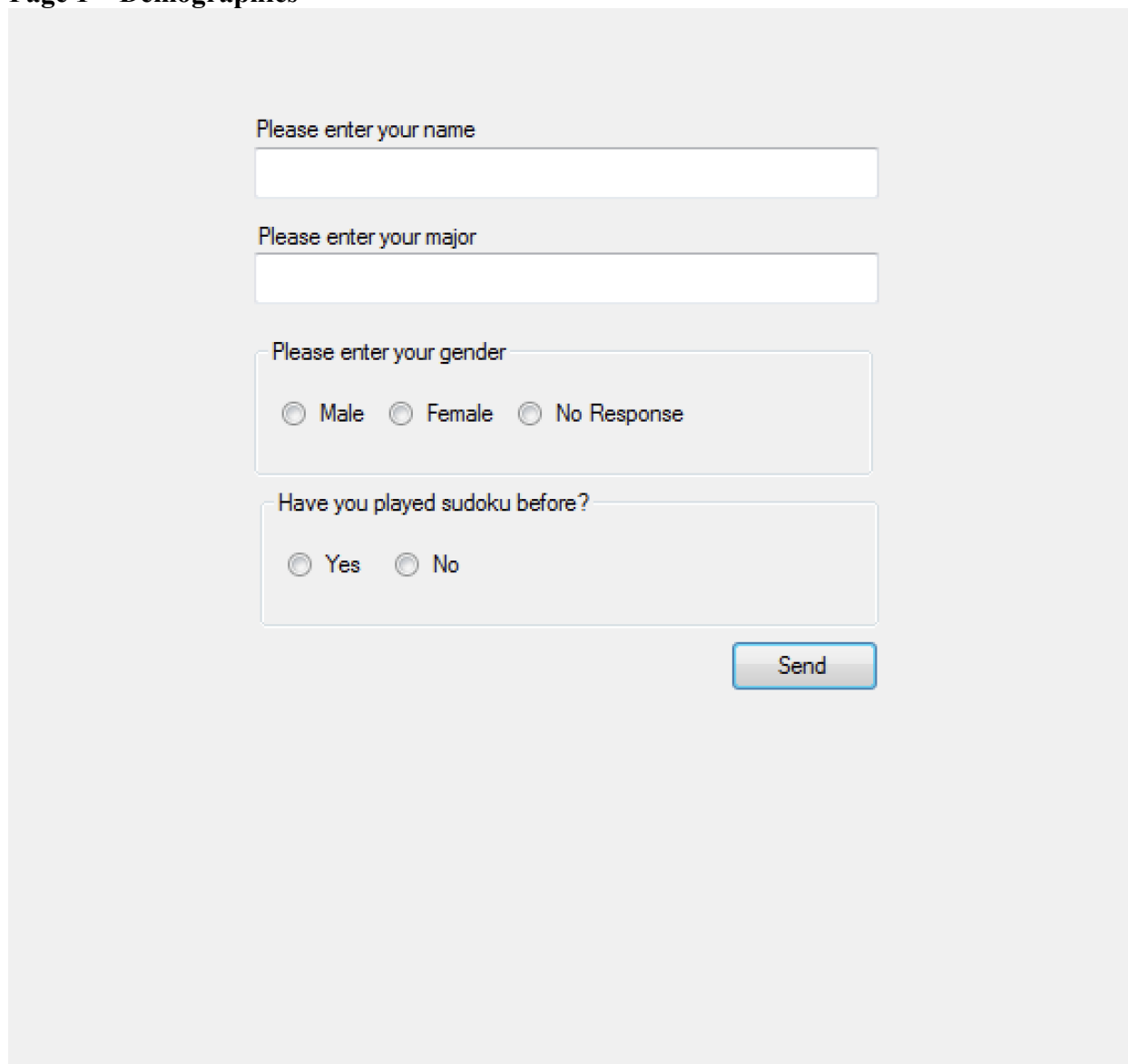The participant in rank 4 will be paid $0

Ties will be broken in favor of those who completed the puzzles the fastest

[At the conclusion, subjects observe their payment from each task and are asked to wait quietly to be called over for payment]

Appendix D: Screenshots

Screenshots from each stage of the Experiment. Nonograms screenshots included only where they differ from Sudoku.

Page 1 – Demographics

Page 2 – General Instructions

You are now participating in a decision-making experiment. If you follow the instructions carefully, you can earn a considerable amount of money depending on your decisions and the decisions of the other participants. Your earnings will be paid to you in CASH at the end of the experiment.

Today's experiment will involve multiple tasks, and your total payment will be the sum of your payments from each task. Please do not talk to other subjects during the experiment. This set of instructions is for your private use only. In case of questions, please raise your hand. Then we will come to you and answer your questions privately.

Next

Page 3 – Dictator Allocation

In Part I of this experiment, you are matched with one other person in this room.

Each person begins with 5 Dollars. Each person chooses how to allocate this money between him/herself and the person he/she is paired with. To specify an allocation, please type the amount you want to allocate to yourself and the amount you want to allocate to the other person and then click "Next." The two amounts must sum up to 5 Dollars.

After everyone has chosen their allocation, the computer will randomly choose one person from each pair whose decision is implemented. Your payment for Part I will be based on the randomly chosen person's decisions.

For example, if your decision is randomly chosen to be implemented, then you will be paid according to your allocation. If instead, the other person's decision is chosen to be implemented, you will be paid according to their allocation.

This is the end of the instructions for Part I. If you have any questions, please raise your hand and an experimenter will answer them privately.

You Receive   They Receive

$ _____ $ _____

Next

Page 4 – Risk Elicitation Outline

In this task, you will make a series of choices between two uncertain options. For each decision, all you have to do is indicate whether you prefer Option A or Option B and click the appropriate button with your mouse.

At the end, the computer will randomly pick ONE decision and then randomly draw a whole number between 1 and 100 to determine your payoff. Your payoff then depends on your choice of A or B in the randomly chosen decision and the number drawn by the computer.

If you choose option A, then numbers between 1 and 50 will always pay $1 and numbers between 51 and 100 will always pay $3.

If you choose option B, then the relationship between the number drawn and the payment is different in each decision, so make sure you carefully compare options A and B to make your choices.

In the first decision, Option B will pay $0 on numbers between 1 and 90 and $3 on numbers between 91 and 100. As you proceed through the decisions, Option B will change. The chance of receiving $3 will increase, while the chance of receiving $0 will decrease.

Remember, after you make all of your choices, the computer will randomly pick one of them to count and will draw a random number to determine your payment.

This means that each decision could be the decision-that-counts so it is in your interest to treat each decision as if it could be the one that determines your payment.

We will reveal the outcome of this task at the end of today's experiment, just before you are paid.

Next

Page 5 – Risk Elicitation Decision Screen (9 choices)

Choose whether you prefer Option A or Option B.

At the end of the experiment, the computer will choose one of these decisions for actual payment and will then draw a random number between 1 and 100, which together with your choices, determines how much you get paid.

Remember, this could be the decision that counts, so consider your choice carefully!

Option A	Option B
Numbers between 1 and 50 pay $1.	Numbers between 1 and 90 pay $0.
Numbers between 51 and 100 pay $3.	Numbers between 91 and 100 pay $3.
Select Option A	Select Option B

Next Decision

Current Decision: 1 / 9

75

Page 6a – Basic Sudoku Instructions

The next task will involve solving puzzles known as Sudokus.

	6	4	3	2	
5	3			6	4
4					1
3					2
6	4			1	3
	1	3	4	5	

A 6x6 sudoku puzzle is a grid 6 squares wide and 6 squares deep. The lines of squares running horizontally are called rows, and the lines running vertically are called columns. The grid is further divided by the darker lines into six 2 X 3 rectangluar 'boxes'.

Some of the squares already have numbers in them. Your task is to fill in the blank squares. There's only one rule:

·   Each row, column and box must end up containing all of the numbers from 1 to 6.
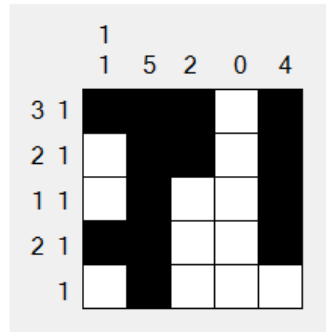
This rule has an important side-effect, which is the basis of all solving techniques:

·   Each number can only appear once in a row, column or box.

Next

Page 6b – Basic Nonograms Instructions

The next task will involve solving puzzles known as Nonograms.

A 5x5 Nonogram puzzle is a grid 5 squares wide and 5 squares deep. The lines of squares running horizontally are called rows, and the lines running vertically are called columns.

Each square must be either filled in black or left blank (or, equivalently, marked with an X).

Beside each row of the grid are listed the lengths of the runs of black squares on that row.

- For example, the top row of the grid says 3,1. This means the row must contain a run of 3 consecutive black squares and a single black square.

Above each column are listed the lengths of the runs of black squares in that column.

- For example, the leftmost column says 1, 1. This means the column must contain two (and only two) single black squares that are not consecutive.

Your aim is to find all black squares. Left click on a square to make it black. Left click a second time to mark it with an X. Left click a third time to turn it white again.
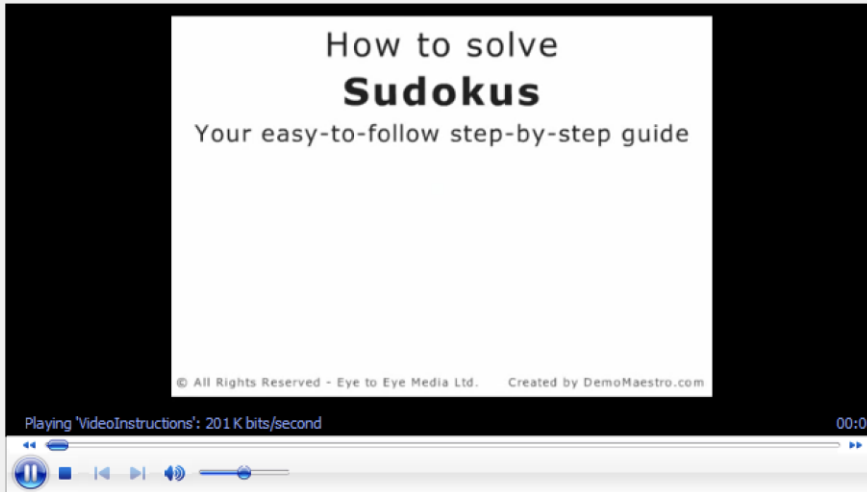
The puzzle is correct when all the rows and columns have the correct number of black squares in the proper positions.

Next

Page 7a – Sudoku Instructional Video

Before the experiment begins, we will watch the following video. Please pay attention to the video as you may find its content useful for the next portion of the experiment for which you will be paid .

This video describes strategies for solving 9x9 sudokus, but they apply equally to 6x6 sudokus.

How to solve
Sudokus
Your easy-to-follow step-by-step guide

© All Rights Reserved - Eye to Eye Media Ltd.   Created by DemoMaestro.com

Playing 'VideoInstructions': 201 K bits/second   00:00

Please Watch The Video

Page 7b – Nonogram Instructional Video

Before the experiment begins, we will watch the following video. Please pay attention to the video as you may find its content useful for the next portion of the experiment for which you will be paid .
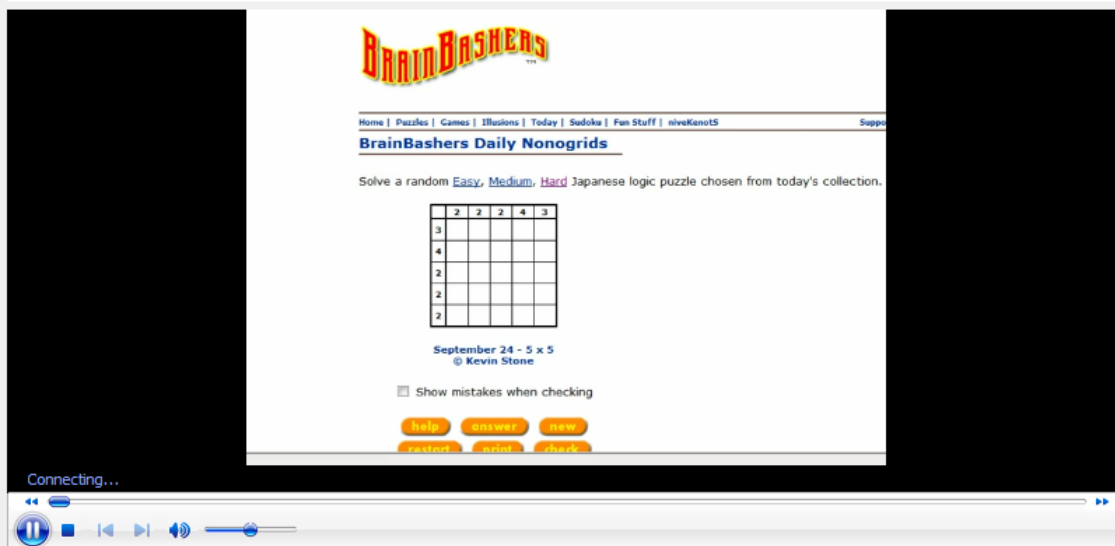

BrainBashers Daily Nonogrids

Home | Puzzles | Games | Illusions | Today | Sudoku | Fun Stuff | niveKenotS   Suppo

Solve a random Easy, Medium, Hard Japanese logic puzzle chosen from today's collection.

September 24 - 5 x 5
© Kevin Stone

☐ Show mistakes when checking

help   answer   new
restart   print   check

Connecting...

Please Watch The Video

Page 8 – Ability Block Instructions

You will now be given 20 Sudoku puzzles to solve. You will have 10:00 minutes. You will be paid $0.5 for each puzzle that you correctly complete and $0 for each puzzle you do not complete or complete incorrectly. There is a timer in the upper right hand portion of the screen for your reference. Use the mouse to select cells. Numbers may be entered by clicking the buttons on screen or with the numbers on the keyboard. If you finish before the time limit, please wait quietly until the next puzzle begins.

Click when ready to continue

Page 9 – Ability Block Screen

				3	
6	4	3	1	5	
	1	4	2	6	
	5	6	3	1	
	3	5	6	2	1
	6				

Time Left
4:57

[1] [2] [3] [4] [5] [6] [eraser] [Clear All]

Page 10 – Ability Block Results and Tracking Information

<div>

Your Results   Your Group's Results

Number of correct sudokus: 2   First place solved: 2   (You)

Your rank: 1   Second place solved: 1

You have earned: $ 1   Third place solved: 1

Fourth place solved: 1

</div>

Now we will begin task 4. You will participate in two periods where you will attempt to solve more sudoku puzzles. The first of these periods will be for practice and the second will be for payment which will be added to your earnings at the end of the experiment.

Based on your performance in Part I, you have been sorted into groups. The top half performers are in one group and the bottom half performers are in the other group. The highest performers are those who correctly completed the most puzzles, with ties broken in favor of those who completed them the fastest

In the second period, you will be paid $0.5 for each puzzle that you correctly complete and $0 for each puzzle you do not complete or complete incorrectly.

Click when ready to continue

Page 11 – Practice Block Instructions (Chat Treatments)

You will now be given a single Sudoku puzzle to solve. You will not be paid for this puzzle. You will have 10:00 minutes to solve the puzzle.

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.

Click when ready to continue

Page 12 – Practice Block Screen (Chat Treatments)

		5	4		
	4	1		2	
4		6	2	5	1
5	1	2	3	4	6
	6		5	3	
		3	1		

Time Left

8:38

1	2	3	4	5	6		Clear All

Page 13 – Practice Block Results

Your Results

Number of correct sudokus: 0

Click when ready to
continue

Page 14 – Evaluation Block Instructions (Piece Rate Treatment)

You will now be given 30 Sudoku puzzles to solve. You will have 15:00 minutes.

Remember, you will be paid $0.5 for each puzzle that you correctly complete, and you will be paid $0 for each puzzle you do not complete or complete incorrectly. At the end of the 15:00 minutes, you will learn the number of puzzles you correctly completed.

You will also learn your rank among the other subjects in terms of the number of correctly completed items in the previous 15:00 minutes.

> Click when ready to continue

Page 15 – Payment Summary Screen

Thank you for participating in this research study

In task 1, your choice was selected

From task 1 you have earned $: 5

In task 2, decision number 1 was randomly selected.
You selected choice A and recieved a roll of: 33

From task 2 you have earned $: 1

In task 3 you solved 2 sudokus correctly

From task 3 you have earned $: 1

For task 4, you solved 1 sudokus correctly
You are ranked number 1 out of 4 participants

From task 4 you have earned $: 0.5

Total Earnings $: 7.5

Please Wait Quietly

Appendix E: Methods Used to Count Instances of Teaching in Audio Data

Two research assistants were given audio files for each group in each session of the *teaching* treatments. We explained that subjects were working together on a single Sudoku puzzle and that the chat transcript provided a record of this interaction. We instructed them to come up with a count of the number of instances of peer-to-peer teaching in the text. To define teaching, they adapted a scheme developed by Kline (2016) that identifies teaching with behavior related to the relevant task in which the following events occur:

- **+/- verbal feedback**- Actor gives positive or negative verbal appraisal of F's behavior with respect to, in F's presence.
- **+/- consequences**- Actor creates positive or negative consequences for F. Includes physical punishment or reward, or verbal description of promised punishment or reward.
- **teasing**- Actor threatens or punishes but in a joking manner, as indicated by smiling and/or laughing. Verbal or gestural.
- **warning (of danger)**- Verbal warning of danger, a separately coded subset of negative verbal appraisal.
- **command to stop**- Verbal command to focal to stop, a separately coded subset of negative verbal appraisal.
- **command to say/do [x]**- Verbal command to say a given phrase, or repeat a particular phrase or gesture, a separately coded subset of positive verbal appraisal.
- **direct attention to object-** Actor directs F's attention toward an object or location, verbally or through gesture
- **direct attention to person-** Actor directs F's attention toward another person, verbally or through gesture.
- **command to watch**- Actor directs F's attention toward watching the actor.
- **other-prompted behavior-** F undertaking an action or behavior after being commanded to do so by another actor.
- **other-assisted behavior**- F undertaking an action or behavior made possible by another actor's help.
- **abstract communication-** Actor gives a verbal explanation or states abstract information to F – including the statement of rules and what is "taboo" behavior.
- **demonstration**- Actor performs behavior or action conditioned on F's attention; may follow a request from F, or the Actor may first manipulate F's attention.

After coding each session independently, our research assistants met to resolve any inconsistencies and produced a count of teaching instances for each group in each session.

A few other notes:

1) A single instance of teaching would include both a question asked by a subject and the answer received.

2) If no answer was received to a question, this was *not* counted as an instance of teaching.

3) We had hoped to get individual level data on teaching, but unfortunately it was not consistently possible to identify specific individuals by their voices and match them to the data.

References:

Kline, M.A. 2016. "TEACH: An Ethogram-based Method to Observe and Record Teaching Behavior." *Field Methods.* doi:10.1177/1525822X16669282