The effect of monitoring on productivity: experimental evidence from elementary schools

by

Andrea Kiss

Submitted to

Central European University

Department of Department of Economics

In partial fulfillment of the requirements for the degree of Master of Arts

Supervisor: Professor Adam Szeidl

Budapest, Hungary

2014

I, the undersigned [Andrea Kiss], candidate for the degree of Master of Arts at the Central European University Department of Economics, declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of work of others, and no part the thesis infringes on any person's or institution's copyright. I also declare that no part the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Budapest, 5 June 2014

_____

Signature

Acknowledgments

I would like to express my appreciation to my supervisor, Adam Szeidl for sharing his time whenever it was needed. His insightful advice helped me along the way of the process and was essential to carry out this research. Special thanks to Thomas Rooney for reading and correcting this thesis, several times. I am also thankful to the CEU Research Grant Committee for financially supporting the data collection. In addition, I would like to acknowledge all the teachers and schools that provided me with the opportunity to carry out the experiment for their assistance during the process. Finally I wish to thank my family and friends for their understanding and caring during my studies.

Abstract

This thesis studies the effect of monitoring on students' outcomes in an experimental setting. During the experiment 270 primary school children were asked to solve mazes in different environments. Four types of treatments were examined, which differed in their intensity and the monitoring agent. The research has three main findings: 1) the presence of an authority, i.e. the teacher, has a significant positive effect on scores and reduces the number of mistakes; 2) girls are more mistake-prone and slower than boys in an intensive monitoring environment, perhaps due to the higher stress of this situation; 3) neighbors have a positive effect on each other's score when the teacher is not present.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Which is the best way to monitor a child? Or is monitoring necessary at all for them? These questions have not been answered with the tools of economics and the answers to them are not all evident.

Children are supervised at home by their parents and in school by their teachers as well. Although the first can be more important, the state can mostly affect the monitoring system in schools if it wishes to do so, via the educational system. Moreover, supervision in schools is more official: students receive grades for both their written and oral tests, which are in fact monitoring devices. For these reasons in this thesis I turn to monitoring in schools, which can provide the environment to test different types of monitoring techniques and compare them. The frequently used tool that I refer to as a baseline method, involves that students hand in their work, which bears their names, to the teacher, who then grades their submitted assignments. However, students sometimes work in groups, when they might be checked by their peers, while in other times they can control the others. Besides the person who supervises students, the intensity can be varied on the scope of very little control and intensive monitoring. These two dimensions give together the different types

of monitoring. Testing the effects of various kind of supervision can result in fine-tuned monitoring for children at least at schools.

Perfectly adjusted monitoring is beneficial and important because the type of monitoring can influence agents' outcomes (Frey, 1993). In schools this means that the way children are supervised affect their outcomes, which are mostly their grades. Later these grades usually have a role in secondary school admission decisions (just like in Hungary) which can shape their whole future. Thus the kind of monitoring at schools can have long lasting effects, for which reason it is the best interest of governments and parents to maximize the grades of children.

Earlier findings suggest that the effect of monitoring on outcomes heavily depends on the underlying dominant processes. There are two main motifs that can be associated with monitoring: 1) increased cost of shirking (Grossman & Hart, 1983) and 2) crowding-out of intrinsic motivation, which can be caused by distrust (Frey, 1993), violated reciprocity (Falk & Kosfeld, 2006) or expected difficulty (Bénabou & Tirole, 2003). The higher cost of shirking has a positive effect on output while the others go against it and depending on which ones dominate in the actual situation, people's productivity can rise or fall.

As the effect of monitoring very much depends on the different circumstances (Frey, 1993), for this thesis I ran an experiment and tested the effects of four treatments, all of them intended to mimic real life situations. The first represented the situation, when students are left alone and the controlling power is low. The second aimed to correspond to the baseline monitoring condition, where students' output is graded while the class is under the eye of the teacher. Then in the third and the fourth specification students' neighbors at the desk received monitoring power. These latter two treatments differ in the intensity of the monitoring; while one involves correcting the neighbor's work, the other is a continuous and probably annoying monitoring type.

The experiment was carried out with 270 Hungarian school-aged children in 2014.

They were asked to complete worksheets of mazes in a given amount of time, hence their productivity was easily measured in the four treatments. This research thus contributes to the experimental economics literature of children, which is still a relatively new area[1]. Furthermore under the assumption that children behave just as adults do under monitoring, it is possible to offer tentative conclusions for the labour market about the preferred type of monitoring at workplaces. Finally, the novelty of this research is that the method and the data set allowed me to look for possible gender differences and peer effects and also to check how they change in reaction to the treatments.

Comparing the treatments, I found that students' productivity significantly dropped when their teacher (i.e. the person who had power over them) was not present. On average children solved 2.25 (0.000) less mazes in the without teacher condition than in the baseline treatment, where the p-value is in the brackets. Once the teacher was inside the room, the other treatments did not have significant main effects on *scores*.

The second result is that in the intensive monitoring treatment, gender differences were present controlling for class fixed effects: girls turned out to be more mistake-prone (they erred more with around 0.2 mazes (0.028) than boys in this round); and they were also slower by 1.6 mazes (0.008) in this round than boys. These two effects together caused that girls' *score* became lower in the intensive monitoring condition with approximately 1.8 mazes (0.003) compared to boys. Otherwise, the two genders behaved similarly during the experimental sessions.

Third, I regressed the student's *score* on his/her neighbor's *score* for every treatment controlling for class fixed effects in search for peer effects. The estimated coefficients for the neighbor's *score* were usually insignificant, except in the without teacher condition. Because in general there are many difficulties in detecting peer effects (Manski, 1993) I

[1]For a short overview see (Krause & Harbaugh, 2001).

ran tests for stronger support. I assigned a new pseudo-neighbor for every student in two different ways and estimated the regression with the pseudo partner's *score* instead of the true partner's *score*. The estimated βs became zero which is more consistent with peer effects than other alternatives.

Based on these results the baseline monitoring procedure is preferable in class rooms and if possible intensive monitoring should be avoided for the sake of girls. Moreover, peer effects between neighbors were not detected in those conditions, where the teacher was in the class room; thus seating does not seem to matter for the productivity of a student.

The thesis is organized as follows. I introduce the relevant related literature to summarize what is already known about monitoring in Chapter 2. Then in Chapter 3 the experimental design is presented in more detail and in Chapter 4 I state my hypotheses. The data set is discussed in Chapter 5 which is followed by the the results in Chapter 6. In the Conclusion I summarize and discuss which monitoring practice might be applicable to education and to the workplaces.

Chapter 2

Background

Based on existing literature, the effect of monitoring on workers' productivity is pretty vague, both theoretically and empirically. On the one hand, principal-agent models predict that monitoring should affect the workers' outcome positively, since it will increase the cost of shirking. The increased cost will in turn decrease the level of shirking as the rational agent sets his effort level according to the marginal cost and marginal benefit of his misbehavior (Grossman & Hart, 1983).

The other possibility is given by the crowding-out literature, which suggests that monitoring can reduce the effort level of the workers. This is possible – according to Frey (1993) – through the channels of

- decreasing intrinsic motivation and/or
- reciprocity.

In the case of the first channel, if extrinsic incentives are introduced the intrinsic motivation might be crowded out. As monitoring is an external incentive, it can reduce the intrinsic motivation of workers and thus diminish their productivity. In the case of the reciprocity

channel, since monitoring signals distrust in the agent, the principal who checks his agent can be judged as a person who violates the unwritten laws of trust between the two actors. Based on Frey (1993) the agent will reply to monitoring by reducing his effort level to punish the principal for this violation.

In addition to these two effects, Bénabou and Tirole (2003) add another channel: information. According to them the agent can infer the high level of difficulty of the task from the very fact that monitoring exists. The possible complexity of the task can discourage agents, which can result in lower outcome. Sliwka (2007) gives additional evidence for the informational effect: agents can conclude earlier misbehavior of their peers from the intensity of supervision.

Like the theoretical approaches, the experimental results are ambiguous. However, this is not surprising as the crowding out literature and the classical principal-agent paradigm are not necessarily contradictory (Frey, 1993). For instance, Dickinson and Villeval (2008) and Nagin, Rebitzer, Sanders, and Taylor (2002) find a positive significant monitoring effect. On the other hand Dickinson and Villeval (2008) report that the intensity of monitoring has an overall positive significant effect in interpersonal principal-agent experiments, however, their findings are also consistent with the reciprocity model of the crowding-out hypothesis.

In a field experiment Nagin et al. (2002) examined a call-center firm while they exogenously varied the probability of supervision. The rational cheater model, which says that agents shirk if the benefits outweigh the expected costs of getting caught, seems to fit the collected data well. Yet the paper shows that there is a great heterogeneity of workers. The same kind of heterogeneity is detected by Falk and Kosfeld (2006). In their experiment they find that the principals may restrict the agents, which, however, decreases the agents' effort level. The reason behind this is that control-averse agents, who are in majority, respond negatively to any kind of control, including monitoring.

The negative effect of supervision is replicated in a recent study in Malawi as well (Guiteras & Jack, 2014). Half of the participants were randomly assigned to the monitoring treatment, which involved quality checks, and in case of errors subjects were sent back to correct their mistakes. People in the monitoring arm had lower output and, interestingly, this decrease was higher for females than males.

Belot and Schröder (2013) examined the spill-over effect of increased monitoring in more detail. They find that although the output of the agents increased thanks to the incentives, in line with the reciprocity model agents punished the supervisors by not completing the task on time.

Based on the research shown, it seems that both principal-agent and crowding-out models can represent the reality of classrooms, but most probably they both play a role at the same time. Still, these studies do not focus on peer monitoring where the principals and the agents know each other well, which would depict the situation at classrooms better.

In contrast to monitoring, there are not many publications about experiments with children. The few that exist are mostly interested in whether children are rational or not; according to Harbaugh, Krause, and Berry they most likely are (2001). Evidence also suggest that children have hyperbolic time preferences (Bettinger & Slonim, 2007) and their social preferences are different from that of adults (Martinsson, Nordblom, Rützler, & Sutter, 2011). Some researchers found gender differences at such young ages for example in competitiveness (Gneezy & Rustichini, 2004) and in cooperation (Cárdenas, Dreber, Von Essen, & Ranehill, 2012). So far the effect of monitoring on children was not examined and thus this thesis aims to fill this gap with an experiment, described in the following chapters.

Chapter 3

Experimental design

This experiment took place in elementary schools in April 2014 with students being the subjects. One of the advantages to experiment with children was that a great number of students could be involved in the study without huge expenses for show-up fees and other payoffs, since their opportunity cost was low if not zero[1]. This helped to have a relatively large sample size in a short period of time.

The procedure of the experiment was as follows. Students solved worksheets containing mazes for four minutes in every round. Their measured outcome was the number of completed mazes in the given round. This task was easy for them for the mazes were intended for pre-school aged children. This means that they were able to solve about fifteen mazes in one round and possible variation between them could occur. Since the mazes could differ in their difficulty, the order of the mazes was randomized.

At the very beginning, students were seated randomly as they had to work in pairs

[1]Participants are usually paid their opportunity cost, the amount they give up participating in the experiment.

in some treatments. At the same time children were given a code which represented them during the whole experiment. The codes contained letters and numbers as well and provided information about the school, the class and the exact seat of a student. As the identifier of a student was a code, the name of the student did not need to be known; hence anonymity could be preserved.

A short introduction followed the seating procedure. The introduction contained information about the study, the nature of the task and some useful tips for solving a maze. Then a familiarization round took place, during which the pupils solved one separate maze. When any struggled with this maze, they got help at this point, but not later. The experiment continued when the sample maze was solved by everyone. In addition to practicing the task, the familiarization helped the students to get used to writing their code on the sheets. Moreover, this round reduced or possibly eliminated the learning-by-doing effect in the latter rounds.

After the familiarization round there were four treatments; three treatments consisted of one round, while one included two. Although I describe them here in a logical order, the treatments followed a random order during every session. The difference between the treatments is the type of the monitoring participants had to bear. The type refers both to the person who monitors the agent and the intensity.

In the first treatment the purpose was to replicate a situation where the authorities are not present. This would have required that there had been no other people in the classroom, just the students. However, students were not allowed to be left unsupervised; thus the closest I could get to this setting was that students worked without the teacher being present in the room during this treatment. I arranged this with the teachers at the first meeting, when they kindly agreed and except on one occasion they always stayed nearby so that they could come back at the end of this round. In the meantime worksheets were distributed to children face-down. When everyone got a worksheet they were asked to

flip the worksheet, copy their code to a designated area and start solving the mazes. This procedure was applied in all the rounds and ensured that every child started working at the same time. When the restricted time was over, the sheets were collected, the teacher came back and the next round started.

To measure the effect of the teacher at the classroom, a baseline round was carried out, during which participants solved mazes again for four minutes while both the teacher and I stayed in the room. This round serves to measure the effect of the presence of the teacher.

Later, when I introduced the peer monitoring round, neighbors started to play additional supervisory roles. At the beginning of this treatment students were informed that their neighbor would check their work at the end of the round. After the four minutes passed, neighbors exchanged their worksheets and corrected each other's work. The neighbors then indicated the *score* and their own code on the cover page. In this round the outcome might be lower than in the previous one, as based on Frey (1993) social connectedness increases the crowding out effect. As the principal and the agent were classmates, they were definitely in a social relationship.

Finally, the last treatment took place. It was in some sense similar to the previous treatment because neighbors worked together, and just as before one of them was the agent, the other the principal. However, the type of the monitoring was exceptionally intensive – so intensive that it could be very disturbing. The student in the agent role solved mazes for four minutes as in any previous treatment. But while she was working, the principal (i.e. her neighbor) constantly observed her and marked on a separate scoring sheet if the agent solved a maze. Both the scoring sheets and the worksheets included the pupils' codes and were collected. When done, the roles were flipped – the agents became the supervisors and vice versa – and the treatment was repeated with these new roles in a new

round[2]. I did this to check whether students' productivity decreased compared to the other rounds, especially to the baseline treatment. If so, then it can be concluded that intensity has a negative effect on outcomes. This unfavorable effect can stem from reciprocity and the increased stress level that this round exposed on participants. At this point, gender differences might evolve, because it was found that genders deal stress differently (Jick & Mitz, 1985).

The treatments, although meaningful on their own, can be combined into a bigger picture. Table 3.1 summarizes the different treatments and their features, which help identifying the important effects. At first, in the without teacher treatment students' performance was measured by an unknown principal, i.e. the researcher, and shirking had no consequences at all. Then the control got more intensive in the baseline treatment, where the presence of the teacher exposed additional control and then in the peer monitoring treatment, where the peers, whom they meet day to day with, could actually detect misbehavior (though they might not report it). The difference between the without teacher treatment and the baseline treatment shows the effect of the teacher being present, while the difference between the baseline and the peer monitoring treatment uncovers how much personal relationships matter in supervision. Finally, the intensive monitoring treatment should point out the effect of constant and probably annoying monitoring by a peer once the effect of baseline monitoring (calculated from the earlier round) is deduced.

At the end of the sessions, each participating student – independently whether they entered the final sample or not – received a chocolate to compensate for their efforts. In addition to this, they received a small letter to their parents to inform them about the nature of the research. In the letter, they could also find contact information in case they had any questions or wanted to exclude their children from the sample. To make sure the

[2]The *scores* of these two rounds were later combined to one variable, to *score_i*.

Table 3.1: Treatments at a glance

	no teacher	baseline	peer monitoring	intensive
researcher was present	x	x	x	x
teacher was present		x	x	x
partner involved			x	x
continous check				x

parents eventually received the message, students were asked to glue it into their school prospectus that is checked by parents regularly. After this step, teachers immediately received a 5 000 HUF gift voucher (about €16.5). A voucher was given to organizers as well (this meant maximum one person per school); they helped to schedule the sessions and provided administrative help besides being the contact person on behalf of the schools. Importantly, those, who made the decision about whether to let students participate or not, that is directors or their deputy, were not affected by this payment at the moment of their decision: they either did not get a voucher, or if they did because they turned out to be the organizers they had not known that organizers would get a voucher as well. Therefore, conflict of interest was not an issue during the school selection procedure.

Chapter 4

Hypotheses

Based on earlier findings in the literature and a pilot that was carried out in March 2014, I have four hypotheses:

1. *The average outcome will depend on the treatment and will form an inverse U shape once depicted in the logical order of treatments.*

This means that I expect to see differences between the number of solved mazes in the treatments, as the treatments are substantially different. Moreover, an inverse U shape would indicate that the number of solved mazes is the lowest in the without teacher condition, then the outcomes increase gradually through the baseline and the peer monitoring case until the effect of monitoring backfires – possibly in the intensive monitoring treatment – resulting in lower outcomes. The inverse U shape also reflects the idea that introducing some monitoring to a non-monitoring environment increases the productivity much more than additional incentives; that is monitoring has a decreasing marginal productivity.

2. *Genders will have significantly different outcomes in some of the treatments, which has the highest probability in the intensive monitoring treatment.*

Females are more error-prone in a stressful environment(Jick & Mitz, 1985); hence especially in the intensive monitoring treatment I expect girls to underperform boys. This round was designed to be annoying and stressful at the same time, which might cause girls to loose their confidence more than boys.

3. *The outcomes of students in the fourth grade will be higher than that of the third graders.*

Fourth graders are about one year older than the third graders. This extra year can result in higher productivity thanks to more developed attention and abilities.

4. *Neighbors at the same desk will have correlated outcomes due to peer effects.*

Although the members of the pairs were randomized, the outcome of a given participant's neighbor might be correlated with his own outcome. That is, if one student in a pair works hard, then his partner at the desk will probably work harder compared to the counterfactual, when he had a less hard-working pair. This finding would support the existence peer effects.[1] In addition, positive effects might be measured because of correlated shocks even in a randomized assignment. This possibility should be checked and discarded to support peer effects.

[1]Note that theoretically peeking can be misidentified as a peer effect; however, this is not a serious issue due to the strict time limit, the different speed they solved the mazes and the presence of two adults constantly controlling them.

Chapter 5

Data and analysis

5.1 Data set

Before the experimentation, I contacted schools in Budapest, Hungary to see if they were willing to participate in the research and if yes, which classes could be approached. I managed to find four schools and arranged a suitable time slot in April 2014 for fourteen classes[1]. These schools were located inside the Grand Boulevard of Budapest; one was located on the Buda side, the other three in Pest. In three schools, I organized four sessions per school, while in one school I got a chance to meet only two classes – one from grade three and one from grade four.

The data set contains 270 subjects. Although the experiment was carried out with over 300 participants, some were excluded from the sample. To start with, children who sat alone at a desk (i.e. who did not have a neighbor) were eliminated, because they could not get the proper peer treatments. Secondly, an entire class was excluded due to an external

[1]The timeline of the sample size is illustrated by Figure 1 in the Appendix.

failure of the experiment. And for one class I eliminated the last round due to teacher incompliance[2].

5.2 Variables

In order to preserve students' anonymity I did not collect background information about them. If I had wanted to, I would have asked the permission of their parents for revealing such information, which would necessarily have required more time and might have resulted in a smaller sample size. After considering these possibilities, I decided to keep administration as simple as possible; thus the data set does not contain students' characteristics, except their *gender*. Essentially, this data was not collected directly from students, but rather I myself noted down the gender of every student on a separate sheet linked to their code while they were working on mazes. This assured that students were not aware of that their gender mattered in any way. In fact, no clue was given about the investigation of possible gender differences during the whole procedure. As no permission is needed to collect visible data, this practice met research ethics obligations.

Besides other basic background variables[3] such as *grade* and *class*, the data set contains measured experimental outcomes of the participants. Two kinds of outcomes were recorded for each round: 1) the number of correctly solved mazes, also referred to as *score*; and 2) the number of *mistakes* a student made while solving a particular worksheet. There were three sources of mistakes, when the student:

[2]The teacher did not come back after the without teacher treatment and hence caused a breakdown in the subsequent round, which was one of the intensive monitoring round. In that round, children received two treatments at the same time – namely the without teacher and the intensive monitoring treatment – and it was pointless trying to disentangle the two effects.

[3]For the full list of variables see Table 1 on page 38 in the Appendix.

- went through the walls of the maze;
- did not fully solved the maze; or
- jumped forward in the worksheet and left intermittent mazes blank.

These mistakes could be intentional or unintentional, yet because it is not possible to differentiate cheating and sloppiness in this set up I treat them in the same way. At the same time, whenever the last maze in a worksheet was not finished because the time was up it was not counted as a mistake.

Since the codes represented the exact location where the student was seated, neighbor related data were available in the analysis. Therefore for all students, every data of their neighbor were connected to theirs, including but not limited to the *gender of the neighbor*, their *scores* and *mistakes* in each rounds.

5.3 Randomization check

Randomization was applied during the experiment to

- distribute mazes into worksheets in order;
- determine the order of treatments for each session; and
- seat children in the class room.

One caveat of the study is that schools were not selected randomly, so there might be a selection bias concerning which schools participated and which not. Moreover, since it is complicated to measure the difficulty of mazes, I cannot verify randomization there; thus I only provide the randomization check of the second and third randomization procedure.

Table 5.1: Treatment randomization check

	Mean	Std. Err.	95% Conf.	Interval
no teacher	3.000	0.453	2.013	3.987
baseline	3.231	0.411	2.336	4.126
peer	2.615	0.350	1.853	3.377
intensive even	2.769	0.411	1.874	3.664
intensive odd	3.385	0.368	2.584	4.185

The order of the treatments was randomized on class level. In sum, there were four treatments with five rounds[4] within the included thirteen classes. Taking the average of the thirteen different orders of the five rounds should give the mean, that is around three. However, since the sample of different classes is rather small (it is only thirteen), it is highly possible that there are slight swings in either direction. These swings are evident in Table 5.1 but the average orders are around three and the 95% confidence intervals of the mean estimates always contain this number, which is a support for randomization.

The other randomization check considers the seating procedure. As mentioned above, the codes of the participants included the exact place, the student was seated. Moreover, this code ended with a two digit number, which was odd for those sitting on the left and even for those sitting on the right side of the desk. The purpose of dividing the students into two groups (odd and even) was to provide a natural test for randomization. The characteristics of these groups should be the same on average in case the randomization was done properly. Table 5.2 shows the comparisons of the means of outcomes and characteristics for the two groups[5].

Based on the statistics, students sitting on the left versus sitting on the right side do not differ significantly from each other in most of the cases. The exceptions involve both

[4]Remember that one treatment, intensive monitoring, consists of two rounds; hence the difference.

[5]The descriptions of the variables are in Table 1 in the Appendix.

the *scores* and the number of *mistakes* in the intensive monitoring treatment. At the first sight it might seem to contradict proper randomization, yet this treatment actually brings together two distinguished rounds, which occurred at different stage of a session and involved different worksheets. For this reason, I do not consider them severe problems.

Table 5.2: Randomization check

	Code	Mean	Std. Err.	95% Conf.	Interval
female					
	even	0.496	0.043	0.411	0.581
	odd	0.504	0.045	0.415	0.593
grade					
	even	0.467	0.043	0.382	0.552
	odd	0.415	0.045	0.327	0.502
score_nt					
	even	14.356	0.386	13.595	15.116
	odd	14.496	0.401	13.705	15.287
score_b					
	even	16.785	0.394	16.009	17.561
	odd	17.089	0.422	16.259	17.920
score_p					
	even	17.000	0.380	16.253	17.747
	odd	16.699	0.443	15.828	17.571
score_i					
	even	15.941	0.413	15.127	16.754
	odd	17.870	0.484	16.917	18.823
mistake_nt					
	even	0.681	0.086	0.511	0.852
	odd	0.797	0.114	0.573	1.021
mistake_b					
	even	0.341	0.077	0.188	0.493
	odd	0.382	0.100	0.186	0.578
mistake_p					
	even	0.370	0.070	0.233	0.507
	odd	0.496	0.100	0.299	0.693
mistake_i					
	even	0.281	0.066	0.151	0.412
	odd	0.187	0.042	0.104	0.270

Chapter 6

Results

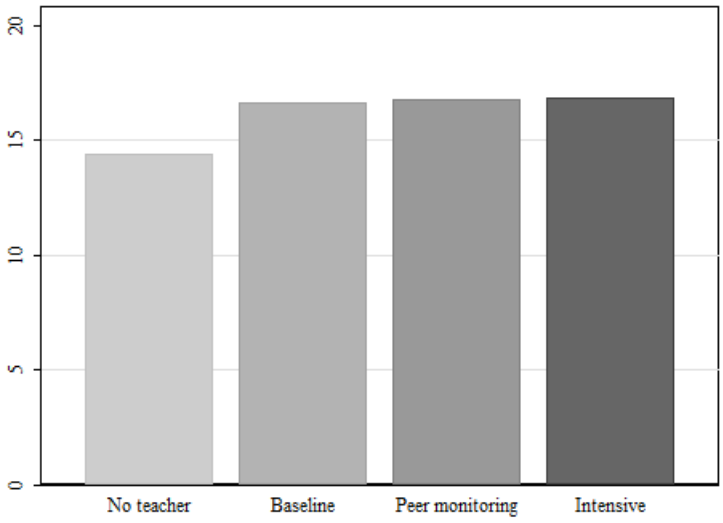
6.1 Treatment effects

Since the order of the rounds were randomized for every session, treatment effects can be calculated as taking differences of the outcomes' sample averages. The effects of the external shocks are examined for both of the two measured outcomes – the *scores* and the number of *mistakes* in a given round.

Figure 6.1 depicts the bar chart of the mean *scores* for the four treatments. Students on average solved more than 16 mazes when the teacher was in the class room, while in the without teacher condition they scored about 14.5. The differences between the without teacher and the rest of the treatments are significant at the 5% level[1]. Yet, the further treatments (the baseline, the peer monitoring and the intensive one) do not differ from each other significantly. This suggests that once the teacher was present, the alternative ways of monitoring did not have an effect on *scores*. Moreover, the presence of the teacher increased

[1]The precise numbers and confidence intervals are shown in the Appendix in Table 2, page 38.

Figure 6.1: Treatment effect on number of solved mazes



the average *scores* by 2.25, which is the increment between the baseline treatment and the without teacher condition. This estimate suggests that productivity can grow roughly by 15.68% when a person in charge looks after the students while they have a task to work on.

The increase in *scores* can have two reasons: 1) children became faster; or 2) they turned out to be more accurate when their teacher was also in the room. The first option means that as students worked more rapidly, they could solve more mazes correctly while they also made proportionally the same rate of mistakes. On the other hand, in case of the second possibility they did not work on more mazes, but rather they were more precise and erred less proportionally. For the sake of learning what causes the improvement, I define the *mistake ratio* as the fraction of mistakes over total number of mazes that the student

worked on. That is,

$$\text{mistake ratio} = \frac{\text{number of mistakes}}{\text{number of correctly solved mazes} + \text{number of mistakes}} = \\ = \frac{\text{number of mistakes}}{\text{score} + \text{number of mistakes}}. \tag{6.1}$$

As explained earlier, studying how the percentage of mistakes changes during the treatments helps to identify which of the two arguments is more sound. Based on Figure 6.2, which shows the *average mistake ratio* by treatments, students worked more precisely when the teacher was in the class room compared to the without treatment effect[2]. The difference is significant and thus presents support for increased accuracy as opposed to improved speed. In addition, the *mistake ratio* also drops between the peer monitoring and the intensive monitoring treatments, which means that students made significantly less errors[3] in the latter condition.

Combining the information of Figure 6.1 and Figure 6.2, the findings are the following:

- The presence of the teachers had a positive significant effect on *scores* by making students more accurate. Teachers on average increased the *scores* of the student by approximately 2.5 mazes. This suggests that people's productivity increases due to baseline monitoring.

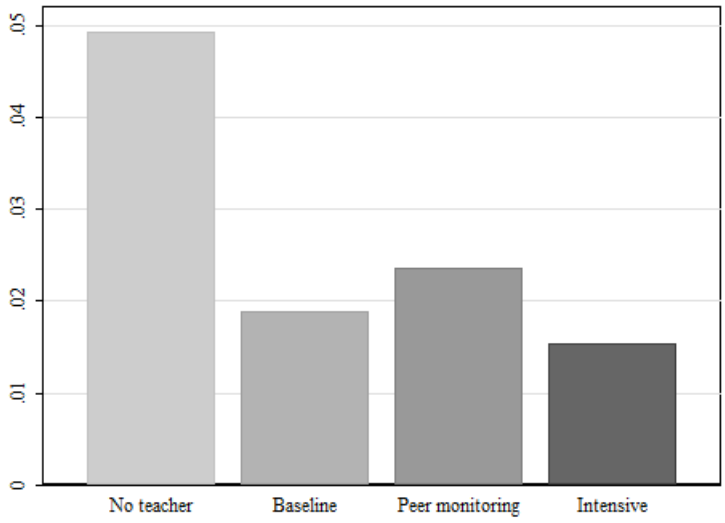
- When teachers were present, there was no other significant treatment effect on the average *scores*, which means that peer monitoring and intensity did not affect *scores*.

- Students' accuracy increased in the intensive monitoring condition compared to the

[2]The basis of this bar chart is Table 3 on page 38 in the Appendix.

[3]Figure 2 shows the average number of mistakes by treatments on page 37 in the Appendix.

Figure 6.2: Average mistake ratio by treatments



peer monitoring treatment, while their *scores* did not differ significantly in these rounds. This suggests that they worked at a slower pace in the intensive monitoring treatment.

6.2 Gender differences

So far average treatment effects were assessed within subjects. Now I turn to between subject comparisons to look for possible gender differences, which can be measured due to the randomization process. Therefore instead of comparing the outcomes by *gender*, I estimated OLS regressions with fixed effects on classes. The rationale behind this is that schools were not randomized in the study; hence, schools cannot be expected to have the same characteristics on average, for which, one way or another, they should appear in the

regressions as fixed effects[4]. As a substitute for school fixed effects I use class fixed effects to control for variations between classes besides the kind of schools, such as: teacher and class room differences, the day of the experiment, the order of the treatments and the number of the students in the class.

The estimated equation is the following:

$$\text{score}_k = \alpha_k + \beta_k * \text{female} + \sum_{j=1}^{13} \gamma_{kj} * \text{class}_j + e_k \tag{6.2}$$

where k stands for the four treatments; and j represents the thirteen classes in the sample. Table 6.1 presents the corresponding regression outputs. In all estimated versions of Equation 6.2 the coefficient of the *female* dummy variable has a negative sign, though in most cases it is statistically indifferent from zero. In the intensive monitoring treatment, however, girls on average gained significantly less *scores* than boys by approximately 1.8 mazes. This result is significant at the 1% level and is relatively high. Also, the estimate is quite robust for adding more variables to the right hand side of Equation 6.2 including *neighbor's gender* or *scores*.

Table 6.1: Gender differences in scores by treatments

	no teacher	baseline	peer monitoring	intensive
female	−0.520	−0.534	−0.473	−1.790**
r2	0.391	0.404	0.415	0.237
r2_a	0.360	0.374	0.385	0.196
N	270	270	270	258

legend: * p<0.05; ** p<0.01; *** p<0.001

The negative female effect in the intensive monitoring condition can stem from the

[4]Note that school fixed effects were not needed in the earlier estimations, because treatment effects were examined within subjects, not between subjects.

slower speed and/or from the increasing *mistake ratio*. When the *mistake ratio* was regressed on the female dummy and class fixed effects, the only treatment for which the gender difference was significant happened to be the intensive monitoring treatment[5]. There the coefficient of the female dummy is highly positive, which implies that girls made more mistakes in relative terms in this treatment compared to boys. At the same time, girls became slower as the average number of mazes that they worked on is significantly lower compared to the baseline condition[6]. The fact that girls both erred more and were more sluggish in the intensive monitoring treatment resulted in their higher *mistake ratio*.

Yet it is still a question why girls are more mistake-prone than boys in this treatment. The other variables in the data set do not have main effects or interaction effects on *scores* and do not affect significantly the coefficient of the *female* dummy. Therefore at this point, relying on the literature, I can only turn to plausible hypotheses. One implies that females makes more mistakes since they bear stress less well than boys (Jick & Mitz, 1985). Although the intensive monitoring treatment was indeed designed to be more annoying, this idea cannot explain why both genders have a lower *mistake ratio* than in any other treatment. If the stress-dealing explanation were true, the *mistake ratio* for girls should be higher at least compared to the peer monitoring condition; in fact, it was lower.

An alternative explanation can be that for girls the unwritten contract between them and the monitoring agent is more important; hence when they were exposed to such a strict control, they reduced their overall productivity by reducing their speed more than boys did. As a punishment they could also make intentionally more mistakes, but I do not have the tools to differentiate between intentional and unintentional mistakes. For this reason,

[5]The results are not reported.

[6]In other words, in Equation 6.1 the numerator (i.e. the number of *mistakes*) increased, while the denominator, which is the total number of mazes on which someone work, decreased. Both of these resulted in higher *mistake ratio* for girls.

both possible explanations are appealing; the *mistake ratio* is higher for females because they became slower and made more mistakes either intentionally as a way of punishment, or unintentionally due to being under pressure.

The last potential explanation for gender differences is that girls might believed in the difficulty of the task more than boys as the monitoring increases. Hence they might shied away from seemingly complicated mazes. This possibility cannot be ruled out perfectly as students were not asked to reveal their guesses about the difficulty of the tasks; yet, because everyone learnt how to do the mazes in the familiarization round and it was emphasised that the mazes in the worksheets would be similar both in difficulty and style, this argument is the least likely.

Due to the significant gender difference in the intensive monitoring treatment, the preferred treatment which gave the highest *score* for students varies for boys and girls[7]. For boys, the order of the conditions is the same as for the whole sample: their *score* is the smallest in the without teacher treatment, while the other three conditions do not differ statistically from each other. On the other hand, the results for girls show a slightly different picture. The without teacher procedure is still the worst in their case as well, but then girls have a significantly higher average *score* in the baseline condition than in the intensive monitoring round. Besides this girls' *scores* in the peer monitoring case are non-differentiable from the baseline and the intensive monitoring treatment. For these reasons, the favored treatments for boys are any of the three, where the teacher was in the classroom, if the only aim is to maximize their *scores*. Girls had the highest *score* in the baseline condition, which is the preferred treatment for them.

[7]The corresponding graph (Figure 3) is in the Appendix on page 39.

6.3 Age differences

In the data set both third and fourth graders were included. Because of this, their comparison can provide insights about how much it counts to be older by approximately one year. In Section 4 I formed a hypothesis that fourth grader students might actually achieved higher *scores* than third graders. For this to be tested, I regressed *scores* on the *grade* dummy variable and school fixed effects.

$$\text{score}_k = \alpha_k + \beta_k * \text{grade} + \sum_{l=1}^{4} \gamma_{kl} * \text{school}_l + e_k \tag{6.3}$$

where, k is the running variable for the four treatments and l stands for the four schools in the sample.

According to Table 6.2, which reports the estimated coefficients, fourth graders in fact performed significantly better in three out of four treatments. (Note that the *grade* dummy, as Table 1 shows in the Appendix, equals to one if the child is from the third grade.) There is a difference between the two age groups only in those conditions, when the teacher was in the class room. Strikingly the age differences in those treatments are huge; fourth graders can gain even about four mazes more in case of the intensive monitoring.

Table 6.2: Age differences in scores by treatments

	no teacher	baseline	peer monitoring	intensive
grade	0.753	−1.835**	−3.797***	−3.900***
r2	0.161	0.059	0.262	0.148
r2_a	0.148	0.044	0.251	0.135
N	270	270	270	258

legend: * p<0.05; ** p<0.01; *** p<0.001

6.4 Peer effects

During the experimental process I randomized students into pairs via changing their place in the class room. This allowed me to test whether a student's neighbor affects his/her own productivity, which is measured as his/her *score*. A positive significant correlation would be more consistent with peer effect.

The existence of peer effects is quite controversial in the literature for the difficulty of identifying it due to the reflection problem (Manski, 1993). Sacerdote (2001) carried out a randomized experiment and managed to identify peer effects for college students living in the same dorm room in some of their outcomes. In a sense, the same analysis would be possible using the data set of this monitoring experiment. However, the circumstances are not that ideal for this estimation compared to the dorm experiment, because although students were randomized into pairs, they were not randomized into schools, and schools themselves were not randomized in the sample either. To mitigate these issues as much as possible I regress the *score* of a child on his neighbor's *score* while adding class fixed effects. The regression was run for students with even and odd codes separately.

The estimated regressions are the following:

$$\text{score}_k = \alpha_k + \beta_k * \text{neighbor's score} + \sum_{j=1}^{13} \gamma_{kj} * \text{class}_j + e_k \tag{6.4}$$

where, as before, k denotes the treatments and j the thirteen classes in the sample. The results are summarized in Table 6.3 for students with an odd code (the ones sitting on the left side of the desk), but since the results for the even numbered students are statistically the same, I report only one table.

According to the estimates the social effect is not present in three treatments, as the point estimates are insignificant; while in the without teacher condition neighbor's *scores*

Table 6.3: Peer effects on children with an odd code

		odd coded's score			
		no teacher	baseline	peer monitoring	intensive
neighbor's score	no teacher	0.224*			
	baseline		0.097		
	peer mon			0.066	
	intensive				0.078

legend: * p<0.05; ** p<0.01; *** p<0.001

have significant effect on own *score*, which is true for children both with odd and even codes. These are possible because when the teacher was not in the class room and hence the control was less tight, students could occasionally talk with their neighbors. Talking with the neighbor, on the other hand, meant that they jointly did not work, which makes their *scores* correlated. When the teacher was in the room, they did not have a possibility for little talks as teachers immediately disciplined them.

Even though the estimated coefficient of the neighbor's *score* on own *score* is significant in the without teacher condition, it cannot be interpreted as a causal effect due to the reflection problem (Sacerdote, 2001). The only thing that can be known at this point is that the results suggest positive peer effects in the without teacher round. Zero peer effects would require insignificant β coefficient in Equation 6.4 (Sacerdote, 2001), which is in fact quite the opposite.

It is not just peer effect that can result in positively correlated *scores*, but common shocks, – which affected both students of a pair – can also account for these estimates. To test if in the without teacher condition peer effects or common shocks were measured, I regress own *score* on someone else's *score* in the class (who is other than the neighbor), which should give zero as a coefficient[8]. In the data set I assigned a new pseudo partner

[8]Of course these estimations include class fixed effects as well.

to every student following two strategies: in the first case I picked the new partner for a student from his/her neighborhood, while in the second case, I assigned a partner randomly from the entire class. Suppose originally A and B sat at one desk during the experiment, hence they formed a group. In line with the first method, the pseudo neighbor of A is another student, say C, who sat in the neighborhood of A and had a code which ended like B's code; that is, if B had an even code then the pseudo neighbor had too and the other way round. With this procedure the pseudo pairs consist of one even and one odd coded child, just as in the original case. The general rule of forming pseudo pairs was that the pseudo partner (C) sat across the aisle to student A. If A sat on the side seats of the room, then his pseudo partner was the one sitting in front of him diagonally, or if there was no such seat then backwards diagonally. Occasionally, the spacial arrangement of the desks was not the usual one, where the desks were in clear rows, but rather they formed small circles to enhance group activity. In that instances the assignment of the partners was made individually by each cases in line with the earlier defined criteria.

After this assignment procedure, which was possible because the exact place of students were known, I re-estimated Equation 6.4 separately for even and odd coded students using the pseudo neighbor's details instead of the true neighbor's. As before, the results do not differ whether the odd or the even coded students were examined; thus I only report the results for the odd students in Table 6.4, which shows the results of Equation 6.4 for all treatments. The point estimates are basically zero, not just for the last three treatments but also for the without teacher condition. This is also supported by the second case of pseudo neighbor assignment, where again all estimates became zero[9]. In other words, since the positive relationship vanished when the own *score* was regressed on the pseudo neighbor's *score* in the without teacher treatment, the results are more consistent with

[9]These estimations are not reported.

Table 6.4: Pseudo peer effects on children with an odd code

		odd coded's score			
		no teacher	baseline	peer monitoring	intensive
pseudo peer's score	no teacher	0.140			
	baseline		0.068		
	peer mon			-0.064	
	intensive				0.123

legend: * p<0.05; ** p<0.01; *** p<0.001

peer effects between true neighbors than correlated shocks.

Based on these results neighbor's *scores* do not affect own *scores* in a controlled environment (i.e when the teacher is in the class room). In schools there is hardly a situation when students' outcomes are measured in an uncontrolled way; they are almost never left unsupervised during tests. For this reason, it is not possible to increase social gains by seating children in a specific pattern.

Chapter 7

Summary and conclusions

In this thesis I investigated the effect of monitoring on children's outcomes and what consequences supervision has on gender differences, age differences and peer effects. I conducted an experiment to collect data from four schools in this research. Due to the experimental design, many hypotheses were able to be tested and thus the thesis has many results. Among them the most important are that 1) teachers had a positive main effect on students' *scores* and negative on their *mistakes*; 2) girls lagged behind boys in the intensive monitoring treatment for being slower and making more mistakes; and 3) peer effects were strongly suggested only when the teacher was not present. That is, all of the hypotheses turned out to be adequate, except the inverse U shape of *scores*.

These findings have consequences for education. In classrooms control has a key role: children significantly performed better with than without supervision. The usual technique for class work, which is to submit home works and tests to the teacher, while the assignment only holds the name of the student, maximizes the productivity of girls compared to the other three tested environments. Therefore I suggest teachers use this method. Note that even though this treatment does not give the highest average *score* for boys, for them

the treatments do not differ significantly once the teacher was inside; hence the baseline method can maximize their *scores* as well. The experiment also suggests that peer effects between neighbors disappear with the presence of the teacher; thus in the school it does not matter who sits next to someone for his/her performance. For this reason it is not possible to increase social gains by following a carefully planned seating procedure.

Although it is possible to form suggestions for the labor market as well based on the findings for students, the connection between children and adults behavior is not well known yet. Therefore any suggestion should be taken with caution and great suspicion. For the labor market the results imply that monitoring has a positive effect on outcomes in general, but very intensive monitoring does not increase average performance, instead it significantly hurts females' *scores*. Therefore I suggest companies adjust the intensity of supervision according to the baseline treatment. However, in jobs where monitoring is difficult or not very feasible, building on peer effects can result in possible positive impacts. This means that selecting carefully a worker's teammates or colleges in the same office room has a potential in productivity growth.

There are several ways in which this research can be extended. For instance, collecting more detailed background data about students can shed light on the underlying reasons of the results. Another direction is to alternate the treatments and focus on other types of monitoring (e.g. creating an environment without any control). In addition, experimenting on different age groups would show if children are indeed similar to adults regarding this trait or not. Finally, there might be empirical differences between cultures; hence it is important to carry out this research in other countries as well.

Bibliography

Belot, M., & Schröder, M. (2013). *Does monitoring work? a field experiment with multiple forms of counterproductive behaviour* (Tech. Rep.). Otto-von-Guericke University Magdeburg, Faculty of Economics and Management.

Bénabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies*, *70*(3), pp. 489-520.

Bettinger, E., & Slonim, R. (2007). Patience among children. *Journal of Public Economics*, *91*(1), 343–363.

Cárdenas, J.-C., Dreber, A., Von Essen, E., & Ranehill, E. (2012). Gender differences in competitiveness and risk taking: Comparing children in colombia and sweden. *Journal of Economic Behavior & Organization*, *83*(1), 11–23.

Dickinson, D., & Villeval, M.-C. (2008). Does monitoring decrease work effort?: The complementarity between agency and crowding-out theories. *Games and Economic Behavior*, *63*(1), 56–76.

Falk, A., & Kosfeld, M. (2006). The hidden costs of control. *The American economic review*, 1611–1630.

Frey, B. S. (1993). Does monitoring increase work effort? the rivalry with trust and loyalty. *Economic Inquiry*, *31*(4), 663–670.

Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. *American*

Economic Review, 377–381.

Grossman, S. J., & Hart, O. D. (1983). An analysis of the principal-agent problem. *Econometrica*, *51*(1), pp. 7-45.

Guiteras, R. P., & Jack, B. K. (2014). *Incentives, selection and productivity in labor markets: Evidence from rural malawi* (Tech. Rep.). National Bureau of Economic Research.

Harbaugh, W. T., Krause, K., & Berry, T. R. (2001). Garp for kids: On the development of rational choice behavior. *American Economic Review*, 1539–1545.

Jick, T. D., & Mitz, L. F. (1985). Sex differences in work stress. *The Academy of Management Review*, *10*(3), pp. 408-420.

Krause, K., & Harbaugh, W. (2001). Economic experiments that you can perform at home on your children. *University of Oregon Economics Working Paper*.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, *60*(3), pp. 531-542.

Martinsson, P., Nordblom, K., Rützler, D., & Sutter, M. (2011). Social preferences during childhood and the role of gender and agean experiment in austria and sweden. *Economics Letters*, *110*(3), 248–251.

Nagin, D., Rebitzer, J., Sanders, S., & Taylor, L. (2002). *Monitoring, motivation and management: The determinants of opportunistic behavior in a field experiment* (Tech. Rep.). National Bureau of Economic Research.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics*, *116*(2), pp. 681-704.

Sliwka, D. (2007). Trust as a signal of a social norm and the hidden costs of incentive schemes. *The American Economic Review*, *97*(3), pp. 999-1012.

Appendix

Figure 1: Sample size in time

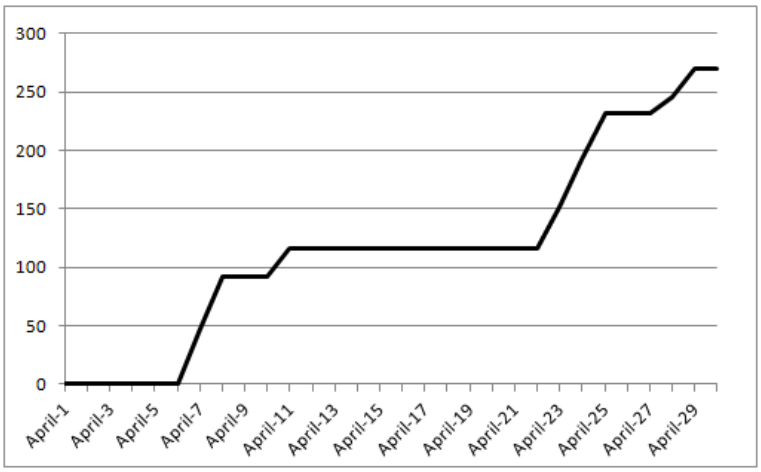


Figure 2: Average mistake ratio by treatments

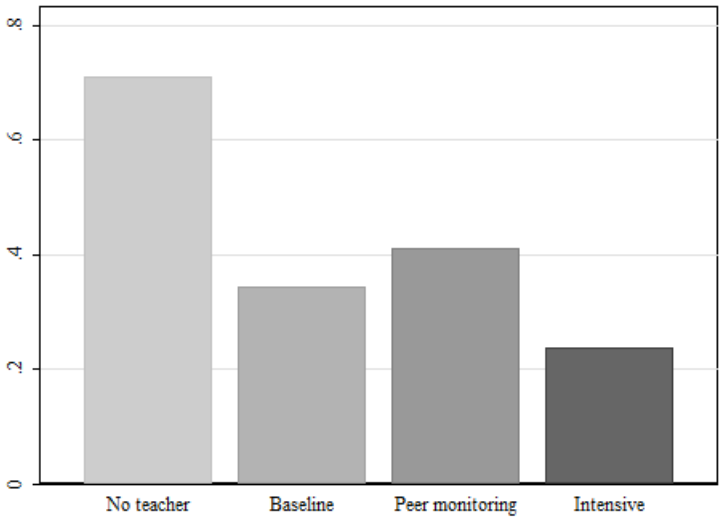


Table 1: List of variables

Variable[a]	Type	Description and values/range
code	string	the individual code for every participant
grade	dummy	0 - fourth grade; 1 - third grade
female	dummy	0 - male; 1 - female
school	integer	the school of the student (1 - 4)
class	integer	the class of the student (0 - 13)
odd	dummy	0 - even, right side; 1 - odd, left side
score	integer	the number of correctly solved mazes in the:
_nt	integer	without teacher treatment
_b	integer	baseline treatment
_p	integer	peer monitoring treatment
_i	integer	intensive monitoring treatment
mistake	integer	the number of mistakes in the:
_nt	integer	without teacher treatment
_b	integer	baseline treatment
_p	integer	peer monitoring treatment
_i	integer	intensive monitoring treatment
ncode	string	the code of the neighbor

[a] Since the code of the neighbor is known for every student, all of the variables for the neighbors can be assigned to the students, hence the list of variables can be duplicated.

Table 2: Treatment effect on scores

	Mean	Std. Err.	95% Conf.	Interval	N
no teacher	14.411	0.267	13.885	14.938	270
baseline	16.670	0.286	16.107	17.234	270
peer mon	16.796	0.280	16.245	17.347	270
intensive	16.860	0.321	16.228	17.493	258

Table 3: Average mistake ratio by treatments

	Mean	Std. Err.	95% Conf.	Interval	N
no teacher	0.049	0.004	0.041	0.058	270
baseline	0.019	0.003	0.013	0.025	270
peer mon	0.024	0.003	0.017	0.030	270
intensive	0.015	0.003	0.010	0.021	258

Table 4: Mistakes treatment effect

	Mean	Std. Err.	95% Conf.	Interval	N
mistake without teacher	0.736	0.071	0.597	0.875	270
mistake baseline	0.360	0.062	0.238	0.483	270
mistake peer mon	0.430	0.060	0.312	0.549	270
mistake intensive	0.236	0.040	0.157	0.315	258

Figure 3: Average scores of boys and girls by treatments