Can a Curriculum that Teaches Abstract Reasoning Skills Improve Standardized Test Scores?

Donald Green
Yale University

Dan Gendelman
Eshcolot

August 23, 2004

Abstract: A pilot study was conducted in order to determine whether a curriculum that teaches strategic principles improves performance on standardized tests. Two classrooms in a low-SES Israeli school participated in an experiment. The more academically advanced classroom was exposed to a series of strategy games; the less advanced classroom was both exposed to games and taught a series of principles designed to guide their thinking in complex choice situations. These principles were taught with special reference to strategy games and then reinforced in the course of ordinary schoolwork. We find that in the wake of this intervention, the less advanced class outperformed the more advanced class on computerized tests of game-playing performance and on paper-and-pencil tests of verbal and math ability.

Introduction

The present study examines the effectiveness of a school-based curriculum called The Mind Lab that attempts to improve students' ability to reason strategically. The Mind Lab is a program that provides instructors and game-based teaching materials to elementary schools. The instructional program is designed to enhance strategic reasoning by drawing analogies to real-life situations. For example, when teaching children to reason through games that present complex sequencing problems, the lesson draws an analogy to a formidable journey that seems overwhelming unless it is broken down into a series of more manageable steps. The idea behind the analogies is to provide easy-to-remember heuristics that have meaning both in games and in life.

Previous research (Green and Gendelman 2003) has shown that the Mind Lab curriculum improves performance in game-playing ability, as gauged by computerized tests. The Green and Gendelman experiments randomly assigned children to treatment and control conditions. Both experimental groups were presented with an introduction to the rules of various games. Children in the treatment group were taught strategic principles useful to the solution of certain types of game-related puzzles, while children in the control group spent the equivalent time period practicing the games. Green and Gendelman find statistically significant increases in performance among those students assigned to receive instruction in strategic principles.

This evaluation extends this line of investigation. In addition to examining whether the analogy-based approach used by the Mind Lab improves performance in strategy games, we consider a more general treatment that integrates strategic reasoning into the day-to-day curriculum of a classroom. The outcome variable in question is not simply game performance but performance on standardized academic tests.

This hypothesis is tested by means of a panel study in which two classrooms of third grade students in a low SES Israeli school were tracked over time. The classroom containing more advanced third graders, as indexed by their standardized scores at the start of the year, was assigned to a control group that played a series of abstract strategy games over a period of several weeks. Strategic principles were not part of their curriculum. The classroom containing less advanced third graders was assigned to the treatment group. They received analogy-based instruction designed to organize their strategic approach to each game and to convey the relevance of strategic principles for thinking in forms of schoolwork.

This essay is organized as follows. We begin by providing an overview of the research design. After describing the population under study and the experimental intervention, we present the statistical model used to estimate the treatment effects. Next, we present results showing that the pedagogic approach used in the Mind Lab significantly improved performance in game-playing performance. Prior to instruction, students in the control group slightly outperform students in the treatment group; after instruction, the treatment group significantly outperforms the control group. We then examine the effects of the experimental intervention on math and verbal scores. Although the control group slightly outperformed the treatment group on math and verbal pretests, the treatment group pulled ahead in both post-tests. The gains in verbal performance were statistically significant, as were the gains in both verbal and math scores summed together.

Research Design and Analysis

Subjects. This study took place in an elementary school in the town of Migdal Haemeq. This town is noteworthy for its high proportion of Ethiopian and Russian immigrants, accounting for nearly half of the town's population. This town's population contains a large fraction of poor families, and the average family income is approximately $12,000. In Migdal Haemeq, just 35% of high school age students graduate with a full diploma. The Shalom school was characterized as a "troubled school," where teachers faced the daunting task of instructing children whose average achievement level was well below average. The teacher in the experimental classroom was eager to try the proposed intervention in the hopes of improving the morale and performance of her students. She was assisted by an instructor from the Mind Lab program, who, from January through May of 2004, helped introduce children in both classes to the computerized strategy games studied here.

Outcome measures. During each experimental session, children played strategy games using a computer interface. The interface provided an overview of the rules, offer a series of examples, and then present a series of puzzles to be solved by the student. Because the games were played on the computer, data on the quality and quantity of play were easily gathered for each student during all three tests. The quality of play was gauged by the number of puzzles solved.

In addition, students were given standardized tests in math and verbal ability. The math test was developed by the school district. Since the school district did not have a standard verbal test, the teaching aide from the Mind Lab developed one based on other tests of verbal ability for third graders. The treatment and control groups were tested, under identical testing conditions, before the study began (September 2003) and after it concluded (May 2004).

Overview of Study Design. The Mind Lab curriculum was introduced over several weeks. The first session consisted of a pre-test in which both the treatment and control groups were taught a pair of games, Rush Hour and Four-in-a-Row. The results of this session provide a benchmark for assessment of subsequent treatment effects. The second session marked the beginning of the experimental intervention; the treatment and control groups were segregated, and the treatment group was taught strategic principles relevant to the game Rush Hour. This lesson stressed the importance of breaking down a complex problem into its component parts, and working in a methodical fashion to solve a formidable challenge. The treatment and control groups tested their performance in both games. The second session also saw the introduction of a new game, Touchdown. Treatment and control groups were taught the rules and, with similar background preparation, tested their ability against a series of computer puzzles. The third lesson introduced students in the treatment group to strategic principles relevant to Touchdown. This lesson offered an accessible introduction to the idea of backward induction, solving a puzzle by working backward from the endpoint to the current position. During this session, students were tested again on Rush Hour and Touchdown. The fourth session saw the introduction of a new game, Lunar Lockout, but no special instruction on principles relevant to it. Students were tested on Touchdown and Lunar Lockout. Thus, we have pre-tests and post-tests for the games Rush Hour, Four-in-a-Row, and Touchdown. In fact, for Rush Hour and Touchdown, we have two successive post-tests.

It should be stressed that throughout the experimental period, the teacher sought to integrate the Mind Lab lessons into the curriculum. Consider, for example, the way that two strategic ideas, the “Detective Method” and the “Ladder Method” were integrated into the curriculum of the treatment classroom. The Detective Method is premised on the idea that asking questions can be a tool to clarify one’s goals and identify the obstacles that prevent one from attaining them. Before every written exercise, students were encouraged to first think about their “detective questions” and say why the answer to the question will help them realize the objective of the exercise. The Ladder Method emphasizes the importance of breaking down a complex problem into its component parts and identifying secondary objectives that will serve one’s primary objective. Students were urged to use the Ladder Method when engaging math problems. For example, one child remarked that "In math lessons I know that I have to solve every stage, if I only miss one part then everything will go down.” When confronted with a problem like 37 plus 15, a student said “I should do it with the Ladder Method and not try to guess the answer in one step. First, I fill in a nice numbers like 37 + 13 = 50, and now I have another 2 so the answer is 52.” This method also applied to verbal reasoning. As one student noted, “When I have to read a text I should ask myself, ‘What is the meaning of the first part of the text and then the second part?’”

Session-by-Session results. A total of 19 students (from the less advanced class) were assigned to the treatment group, and 16 students (from the more advanced class) were assigned to the control group. Due to the vagaries of student attendance, some of the students tested in the pre-test phase of the experiment were absent when subsequent tests were administered. The number of observations varies from 15 to 19 in the treatment group and from 13 to 16 in the control group.

The assignment of classrooms to experimental groups produced a treatment group that lagged behind the control group in terms of pre-intervention test scores. For example, the control outperformed the treatment group on a standardized math test used in the local school district (treatment mean = 67.8, control group mean = 70.9), on a standardized verbal test devised by the Mind Lab instructors (treatment group mean = 60.2, control group mean = 70.9), and in the Rush Hour pretest (average number of puzzles solved by the treatment group = 4.4, average numbers solved by the control group = 5.7).

As shown in Table 1, the second meeting saw the treatment group surpass the control group in terms of average performance on both of the games on which they had received instruction. The treatment group, whose scores had been poorer than the control group in the pre-test, now outperformed the control group in Rush Hour (treatment group average = 6.6, control group average = 5.1). The computer continued to dominate students in 4-in-a-Row, although it should be noted that both students who won any games at all against the computer were in the treatment group, which had earlier received instruction on relevant strategic principles. There was no difference in performance on the game Touchdown, which the students played initially without any instruction in strategic principles (treatment group average = 5.3, control group average = 5.4).

Table 1: Performance of Treatment and Control Groups, by Game and Experimental Session

(Table entries are means with Ns in parentheses)

	Session	Control Group Mean	Treatment Group Mean	Significance (two-tailed)
Pretest Four-in-a-Row	1	0 (16)	0 (19)	N/A
Posttest Four-in-a-Row	2	0 (16)	.13 (15)	.14
Pretest Rush Hour	1	5.7 (16)	4.4 (19)	.09
Posttest Rush Hour	2	5.1 (16)	6.6 (15)	.06
Posttest Rush Hour	3	8.0 (14)	9.0 (16)	.17
Pretest Touchdown	2	5.4 (16)	5.3 (14)	.91
Posttest Touchdown	3	8.1 (15)	9.1 (16)	.33
Posttest Touchdown	4	10.3 (15)	11.7 (15)	.15
Pretest Lunar Lockout	4	4.1 (16)	4.2 (17)	.81

The third session saw the students hone their skills on Rush Hour and Touchdown, and no new games were presented. The treatment group outperformed the control group in both games. In both cases, the mean in the treatment group was one point higher, indicating that on average students in the treatment group solved one additional puzzle. The fourth session gave both groups one more opportunity to test their skills in Touchdown, and the treatment group continued to solve more puzzles than the control group. When both groups were presented with a new game with no special instruction, however, the two groups performed almost identically. The near-equivalence of the two groups in the pretest for both Touchdown and Lunar Lockout may be interpreted in two ways. One interpretation relates these scores to the baseline and concludes that instruction in strategic principles helped buoy the scores of the treatment group, which prior to any of the interventions lagged behind the control group. Alternatively, one could ignore the baseline scores of the two groups and, seeing no difference in average scores, conclude that the instruction did nothing to distinguish the pretest scores of the treatment group.

Estimation. In this section, we attempt to gauge the effects of the experimental treatment more precisely. Two statistical models are considered. The first is simply a linear regression model of the outcome measure (posttest scores) on the independent variable, experimental group. Let Y represent a vector of post-test scores. Let X denote a dummy variable scored 1 if the student was assigned to the treatment group. Let U represent a vector of disturbances. The regression model is

$$Y = a + Xb + u,$$

which turns out to be equivalent to a comparison of average outcomes in the treatment and control groups.

Given the small size of this panel study, pre-test scores may play a potentially useful role in estimation. By reducing the amount of unexplained variation in the dependent variable, these covariates improve the precision with which the experimental effects are estimated. Let P represent a matrix of pre-test game scores, and let S denote pre-intervention standardized test scores. Thus, P and S comprise the pre-test verbal score, the pre-test math score, and the pre-test score of the relevant strategy game. The regression model is

$$Y = a + Xb + Sc + Pd + u.$$

The central hypothesis of this study concerns the parameter *b*: if the treatment improves test performance, *b* is positive. Thus, a one-tailed test will be used to gauge the statistical significance of the result against the null hypothesis that the treatment did nothing to improve scores.

Results: Curriculum Affects Game Performance

Table 2 reaffirms the findings of Green and Gendelman (2003). The Mind Lab curriculum improves game-playing performance. This pattern holds for both Rush Hour and Touchdown, although the effect is most apparent statistically when the two games are analyzed jointly. Without controlling for covariates, we find that the treatment group completed an average of 6.5 more puzzles for both games. Controlling for covariates, this number jumps to 7.7. Both numbers are statistically robust ($p < .01$).

Table 2: Effects of Treatment on Posttest Game Scores, with and without Covariates

	Rush Hour	Rush Hour	Touchdown	Touchdown	Both Games	Both Games
Treatment	3.0 (1.4)	3.6 (1.5)	2.6 (2.0)	3.9 (1.4)	6.5 (2.9)	7.7 (2.1)
Pretest Rush Hour Score		.4 (.3)				.3 (.6)
Pretest Touchdown Score				1.7 (.3)		2.2 (.6)
Pretest Verbal Score		-.04 (.05)		-.04 (.04)		-.08 (.06)
Pretest Math Score		.04 (.05)		.07 (.04)		.07 (.07)
N	27	26	30	25	26	24

Results: Curriculum Affects Standardized Test Performance

That instruction improves performance in strategic games is a conclusion that is neither new nor altogether surprising. What sets this evaluation apart from previous research is that it estimates the effects of the Mind Lab curriculum on standardized test perfomance. Two tests were administered, one math test that was commonly used among nearby school districts and a standardized verbal test that was developed in collaboration with the classroom teacher.

Our regression model predicts post-test scores on each test with a dummy variable marking the experimental treatment and pre-test scores in math, verbal, and Rush Hour. Note that all of the pre-test covariates were measured before any Mind Lab instruction occurred. Although intuition suggests that strategic thinking might be more readily generalized to math as opposed to verbal reasoning, our results suggest that the effects are at least as strong for verbal scores. Table 3 reports that the treatment group scored on average 11.9 points higher than the control group (SE=5.1, $p = .014$), which is slightly more than half a standard deviation improvement in the verbal post-test. The results were also positive for math, with the treatment showing a 4.8 point gain. However, given the standard error of 3.1, this effect falls just shy of statistical significance ($p = .065$). Summing the two post-tests together to form a composite score and repeating the regression estimation shows a statistically significant 17.5 point gain (SE=7.5, $p=.015$).

The estimates in Table 3 vary markedly, depending on whether we control for baseline covariates. Without controls for baseline performance, the effects are weakly positive. When controls are included, the treatment effects become much stronger. This pattern reflects the contrasting baseline scores of the two groups. After the intervention, the treatment slightly outperformed the control group. However, when one takes into account the substantial gap between these two classrooms prior to the study, the effects of the Mind Lab curriculum become much larger. In other words, the Mind Lab curriculum appears to have closed the substantial gap between the two classrooms.

Table 3: Effects of Treatment on Posttest Game Scores, with and without Covariates

	Verbal	Verbal	Math	Math	Verbal + Math	Verbal + Math
Treatment	1.3 (6.1)	11.9 (5.1)	1.4 (4.7)	4.8 (3.1)	1.3 (9.6)	17.5 (7.5)
Pretest Verbal Score		.66 (.13)		.18 (.08)		.85 (.20)
Pretest Math Score		-.03 (.12)		.47 (.08)		.42 (.18)
Pretest Rush Hour Score		.91 (1.06)		.82 (.70)		1.81 (1.57)
N	30	30	34	33	30	30

Discussion

The data presented above indicate that the Mind Lab curriculum had two statistically significant effects. First, as expected, it improved game-playing performance, as gauged by the number of puzzles each child solved during each session. Prior to the experimental intervention, the control group enjoyed a slight edge in game-playing performance; after the intervention, the treatment group significantly outplayed the control group.

Second, and more importantly, the treatment group outperformed the control group in standardized testing. The results were more decisive for verbal than for math scores; nevertheless, the findings lend surprising support for the idea that instruction in strategic reasoning improves academic scores.

Additional research is needed to understand why the Mind Lab curriculum improved verbal and math scores. One possibility is that the curriculum helped students in the treatment group negotiate standardized tests – they might have become more sensitive to the strategy of picking the best option from a field of choices. Another possibility is that games made ordinary schoolwork more fun for the children in the treatment group, enhancing their attentiveness to their lessons. A small post-intervention interview of the children lends some support to both of these hypotheses, as children expressed widespread enthusiasm for the Mind Lab curriculum, claimed that it gave them confidence, and often said that it helped them in their all around academic performance. It remains to be seen whether the promising results in this pilot study can be replicated in other grades and school environments.

References

Green, Donald P., and Dan Gendelman. 2003. Teaching Children to Think Strategically: Results from a Randomized Experiment. Unpublished manuscript, Institution for Social and Policy Studies at Yale University.