Impact Evaluation of Mindspark

April 7th, 2014

Executive summary

About Educational Initiatives, Mindspark, and ASSET

Educational Initiatives (EI) is a prominent education organization in India with the mission of ensuring that every child learns with understanding. EI offers a scientifically designed test called ASSET that measures skills and deeper understanding of concepts, and an adaptive learning program called Mindspark to teach primary school students math according to their current level of understanding. Mindspark has been developed through student usage in classrooms over the past 5 years, and is currently used in approximately 900 government and private schools in India and the Middle East.

About the evaluation

IDinsight, a client-focused impact evaluation firm, designed an evaluation to answer these questions:

1. Primary: What is the impact of using Mindspark on student learning outcomes in maths?
2. Secondary: Is Mindspark more effective for:
    a. Advanced students or students with remedial needs?
    b. Solving basic or difficult questions?
    c. Students in early or later grades?
    d. Boys or girls?
3. Secondary: Is there a relationship between the intensity of Mindspark usage and learning gains?

The evaluation matched students in classes that used Mindspark to students with almost identical baseline test scores, in classes with similar average baseline test scores in schools, in the same states, using the same curriculum, charging similar fees. Impact was measured by comparing endline test scores after the students had used Mindspark for one school year. Extensive robustness checks are provided to address potential limitations of the evaluation method such as school or teacher effects, and all provide strong evidence that the estimated impact is accurate and unbiased.

Key findings

The impact of Mindspark was measured for the 2011-12 and 2012-13 school years, with an average usage period of 10 months and 27 hours per school year.

The results for each year were nearly identical: an impact of 0.19 s.d. on test scores in the first year, and 0.20 s.d. in the second year. These gains are statistically significant, and large relative to evaluations of similar interventions and to all schools that take ASSET. Mindspark schools' average increase was larger than that of 86% of all other schools that took ASSET in the same year.

Mindspark had a larger impact on fourth grade students (the youngest evaluated), with an average increase of 0.32 s.d.

Table of Contents

Evaluation design

About the intervention: Mindspark[1]

Mindspark is an adaptive learning software that:

1. Helps students learn Maths and language by employing a constructivist theory of learning by answering questions that are appropriate to their current understanding
2. Remediates common student misconceptions identified through 10 years of ASSET tests
3. Supports teachers in schools to teach to where the need is as opposed to what the curriculum tells them they should be teaching

Evaluation questions

1. Primary: What is the impact of using Mindspark on student learning outcomes in maths?
2. Secondary: Is Mindspark more effective for:
    a. Advanced students or students with remedial needs?
    b. Solving basic or difficult questions?
    c. Students in early or later grades?
    d. Boys or girls?
3. Secondary: Is there a relationship between Mindspark usage intensity and learning gains?

Evaluation method

The evaluation matched students in classes that used Mindspark to students in classes and schools that did not use Mindspark according to the following variables at baseline, prior to any Mindspark usage:

Table 1: Matching variables

Variable	Included in primary specification?	Matching range
Baseline test score (item response theory scaled ASSET)	Yes	4 specifications: divided students into 10, 20, 30, and 40 groups, matching within each group.
Class average test score	Yes	8 specifications: divided students into 0, 3, 6, 9, 12, 15, 18, and 21 groups, matching within each group. The average impact of the 32 different specifications (8 for class average baseline times 4 for student baseline) to avoid relying on arbitrary cutoffs.
Class	Yes	Exact
Language of instruction	Yes	Exact
Curriculum (board)	Not used for matching because it would reduce sample size, but controlled for statistically.	In a robustness check with exact matching on board, results remained the same.

[1] The language used in this section is based on and in some cases directly copied from the ASSET website: http://www.ei-india.com/what-is-mindspark/

Location (state, and city as robustness check)	Not used for matching because it would reduce sample size, but controlled for statistically.	In a robustness check with exact matching on state and city, results remained the same.
School fees	Not used for matching in primary specification due to missing data, but included as a robustness check	Within 0.5 standard deviations for each school (sufficient to produce balance). Results remained the same.

The matching technique employed was coarsened exact matching,[2] a method with very strong properties to reduce potential sources of bias that works effectively with large data sets. This process created a comparison group with near identical group means and standard deviations on these key variables. See the section "matching results" below.

Potential imitations of the evaluation

The ideal method for this type of evaluation would have been a randomized controlled trial. By randomly assigning some students (or schools) to use Mindspark while others to not use it, these two groups would be near identical on observable baseline characteristics such as test scores, but also on difficult to observe characteristics such as the motivation of school leadership, teachers, parents, and students.

The methodology employed in this evaluation, matching with difference in differences analysis, controls for observable characteristics, but leaves open the possibility of systematic differences in unobserved characteristics. They key potential sources of bias are why some schools and some classes decided to purchase and use Mindspark, while others did not. These school and teacher "effects" were directly tested with "placebo treatments" which show that school and teacher effects do not seem to exist. Therefore, the comparison group seems to be a strong estimate of the counterfactual, so these findings are robust and conclusive.

Another potential limitation is in the external validity of these findings. Schools in this study were all English medium private schools, mostly urban, and mostly charging medium or high fees. It is very likely that other such schools that decided to use Mindspark would experience similarly large gains in tests scores. It is not known to what extent these results would be found in different schools, such as government schools or "affordable" private schools.

Test instrument & outcome variables[3]

ASSET (Assessment of Scholastic Skills through Educational Testing) is a scientifically designed, skill-based assessment test. Rather than testing rote learning, through multiple-choice questioning, it focuses on measuring how well students have learned skills and concepts underlying the school syllabus. The test provides information on the strengths and weaknesses of individual students, entire classes and schools.

ASSET features

- For students of classes 3-10
- Core Subjects: English, Maths and Science

[2] Iacus, King and Porro, 2011, Causal Inference without Balance Checking: Coarsened Exact Matching, Political Analysis 20(1).
[3] The language used in this section is based on and in some cases directly copied from the ASSET website: http://www.ei-india.com/asset/

- Optional Subjects: Social Studies and Hindi
- Based on the CBSE, ICSE, IGCSE, and major state boards' curriculum, but tests conceptual understanding
- Detailed skill-wise feedback with customized letter for every student and teacher
- Administered in schools every August or December
- Provides a benchmark of the student's, teacher's and school's performance with peers all over the country

Note that ASSET was not designed to measure the impact of Mindspark, as it existed before Mindspark was conceived, reducing the risk of "teaching to the test."

Data description

Educational Initiatives (EI) provided all data to IDinsight.

ASSET data

EI provided IDinsight with all of the ASSET test data from 2010–2013, administered twice per year (though most school only test students once per year). As explained in the timeline below, only test data from the August test administration was used.

ASSET is normalized by grade level across all test takers, with mean 500, standard deviation 100, minimum of 200, maximum of 800.

This database included raw and scaled test scores for 428,129 students in grades 3-9 in 590 schools.

Mindspark data

EI provided IDinsight a list of which classes in which schools used Mindspark for each school year. In addition, EI shared detailed data from the Mindspark management information system that reveals how often and for how long students used the software.

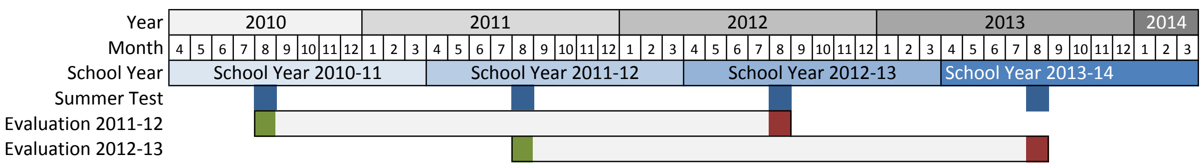
School information data

EI shared their database of 3,495 private schools in India, containing the schools' locations, curriculum (board), medium of instruction, schools size, and an estimate of how much the school charges in fees.

ASSET and Mindspark timeline

The following figure shows the timing of the typical school years, summer ASSET administration, and 2 evaluation timelines, both of which are included in this report. Green indicates a baseline test, red the endline test:

Figure 1: ASSET testing, school year and evaluation timeline



Data included in this evaluation

From the complete data sets, the following data was used in the evaluation after the matching process:

Table 2: Matched data

School year	Mindspark / comparison	Schools	Grades (can include multiple sections per school)	Students
2011-2012	Mindspark	18	52	4,559
	Comparison	143	311	13,727
2012-2013	Mindspark	15	29	2,543
	Comparison	63	94	2,323

For the main impact measure, the results from 32 different matching specifications were averaged. For the subgroup analysis, students were matched within 20 distinct groups by student test score and 12 distinct groups for class average test scores.

Data analysis

Following the matching process, the main data analysis technique is difference-in-differences[4] with the following specification:

$test\ score_{endline,\ i,j} = \beta_0 constant + \beta_1 Mindspark_{i,\ j} + \beta_2 test\ score_{baseline,\ i,\ j} + \beta_3 class\ test\ score_{baseline,\ c,\ j} + \beta_4 curriculum_j + \beta_5 state_j + \varepsilon_{i,j}$

Where variables are defined as:

- $test\ score_{endline,\ i,j}$: test score of student i in school j at endline
- $Mindspark_{i,\ j}$: this is a binary variable equal to 1 if the student was in a class that was assigned to use Mindspark
- $test\ score_{baseline,\ i,\ j}$: test score of student i in school j at baseline
- $class\ test\ score_{baseline,\ c,\ j}$: mean test score of class c in school j at baseline
- $curriculum_j$: the curriculum (board) used by school j
- $state_j$: the Indian state of school j
- $\varepsilon_{i,j}$: error term for student i in school j at baseline

In addition to the main specification, alternate specifications are included as robustness checks, in which fewer or more controls are employed, including controls for class, gender, city, and fees. These were excluded from the main analysis because of some missing data for the latter 3, which caused reductions in the sample size, and because controlling for class reduced precision. The findings did not change when including these as controls, showing that the comparison group was already strong.

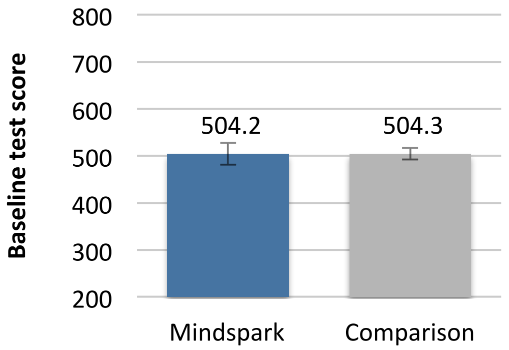
Standard errors were clustered at the school level. CEM matching weights were employed, which did not affect the estimated impact, but increased the precision.

Matching results

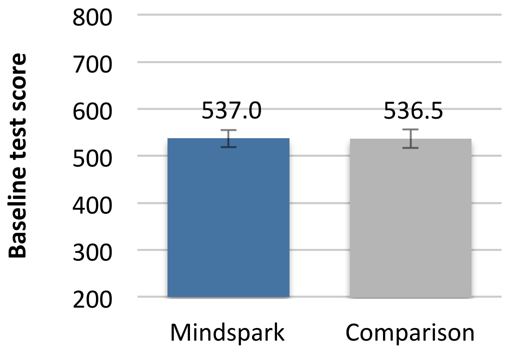
The following graphs demonstrate how well the matching process worked in creating two nearly identical groups. As a reminder, for the main impact measure, the results from 32 different matching specifications were averaged. The graphs presented below are for the specification when students were matched within 20 distinct groups by student test score and 12 distinct groups for class average test scores.

[4] Note that instead of subtracting baseline test score values, this specification statistically controls for baseline levels, which maximizes statistical power.

Graph 1: Balance in 2010 baseline test scores

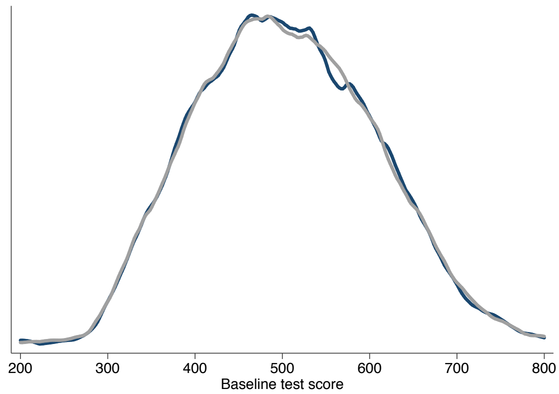


Graph 2: Balance in 2011 baseline test scores

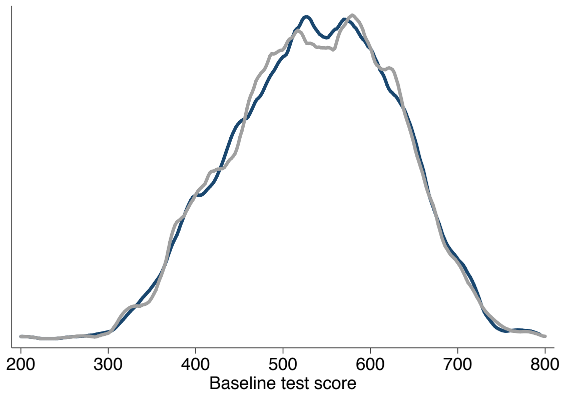


The following two graphs show how closely the comparison student distribution (gray) was to the treatment student distribution, (dark blue) far beyond just having similar means and standard deviations.[5]

Graph 3: Balance in 2010 baseline test score distributions



Graph 4: Balance in 2011 baseline test score distributions



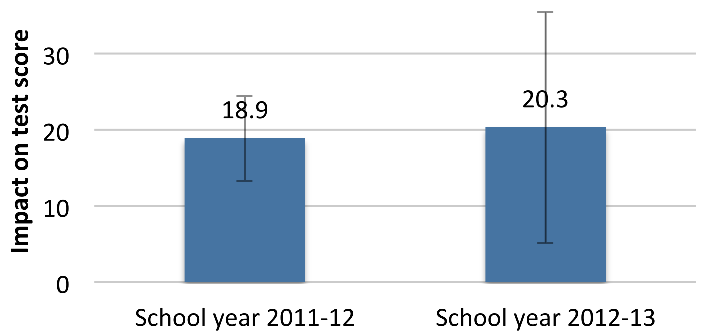
Results

Primary results

In both evaluations, Mindspark had a positive, statistically significant, and practically meaningful impact. Students in classes that used Mindspark scored 18.9 points higher (0.189 s.d.) than matched students that did not use Mindspark in the 2011-12 school year. In the next year the estimated impact was virtually the same, 20.3 points (0.203 s.d.). The results are presented below, and the full table of results for all 32 specifications is given in the appendix:

[5] Graphs were made with kernel density smoothing, bandwidth 0.2.

Graph 5: Primary results: impact and 90% confidence intervals



Robustness of the findings

Extensive mechanisms were employed to check the robustness of the findings. First, results were found to be stable, and not dependent on

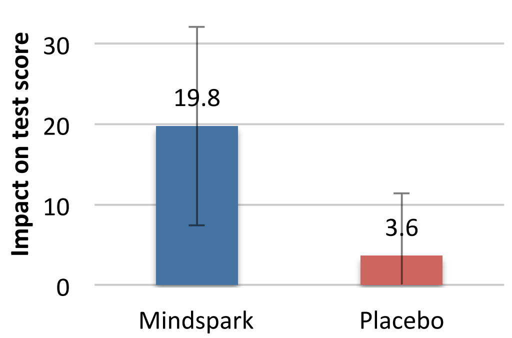
1. Different cutoff points for matching variables (see the appendix)
2. Inclusion or exclusion of control variables (results available upon request)

In addition, extra tests were done to test the robustness of the findings. One concern of education studies that do not use randomized controlled trials is that there might be school or teacher effects – that is, the schools and teachers that use Mindspark might have already been on a faster learning trajectory, even without Mindspark. Both of these possibilities are investigated below:

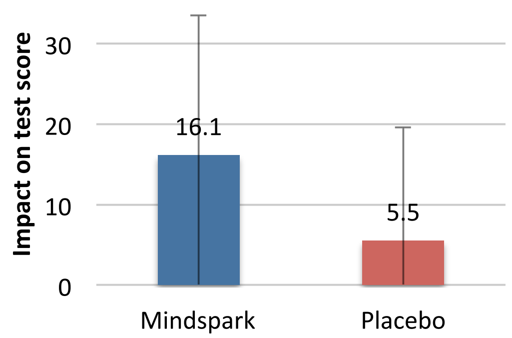
School effects

To investigate the possibility of school effects, "placebo treatments" were created. Many of the schools that used Mindspark did not use it in every grade. So students in schools that *did* use Mindspark but in grades that *did not* use Mindspark were matched to similar students in comparison schools that did not use Mindspark, and the difference between their learning gains was calculated. The graphs below show these placebo impacts in red, next to the impact of true treatment students within the same schools in blue. The placebo impacts were very small and not statistically different from zero, while the true treatment effects remained similar to the treatment effects measured overall, providing strong evidence that there were no school effects.

Graph 6: Impact of Mindspark compared to placebo impact for non-Mindspark grades in Mindspark schools, 2011-12



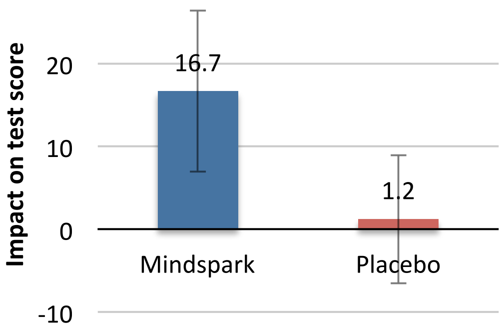
Graph 7: Impact of Mindspark compared to placebo impact for non-Mindspark grades in Mindspark schools, 2012-13



Teacher effects

To investigate the possibility of "teacher effects" another set of placebo treatments were created. In this case, the change in test scores for Mindspark treatment grades was measured in the year before that school purchased Mindspark. Assuming that the majority of teachers do not move grades or schools year to year, this captures the effect of the same teachers on a different group of students that did not use Mindspark. Once again, the placebo effect was not statistically different from zero, providing strong evidence that there were no teacher effects.

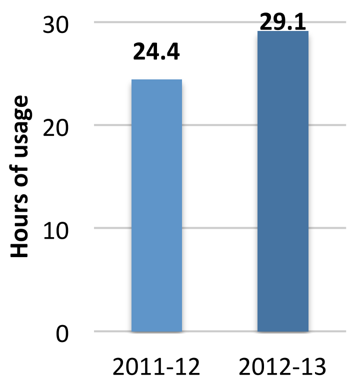
Graph 8: Impact of Mindspark compared to placebo impact for the same teachers in the year prior to treatment



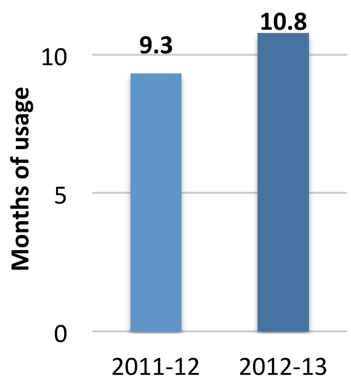
Exploring usage

In classes that the schools purchased Mindspark for, 84% of students used the software at least once in the 2011-12 school year, and 91% in the 2011-12 school year. The graphs below present average total usage hours during the evaluation period, average months with at least some (greater than 0) usage, and the average hours used per month. This shows that while there were 2 years between the baseline and endline, on average students used Mindspark for the duration of one school year.

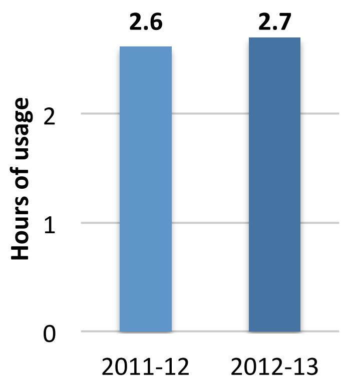
Graph 9: Average usage hours during evaluation



Graph 10: Average usage months during evaluation



Graph 11: Average usage hours per month during evaluation



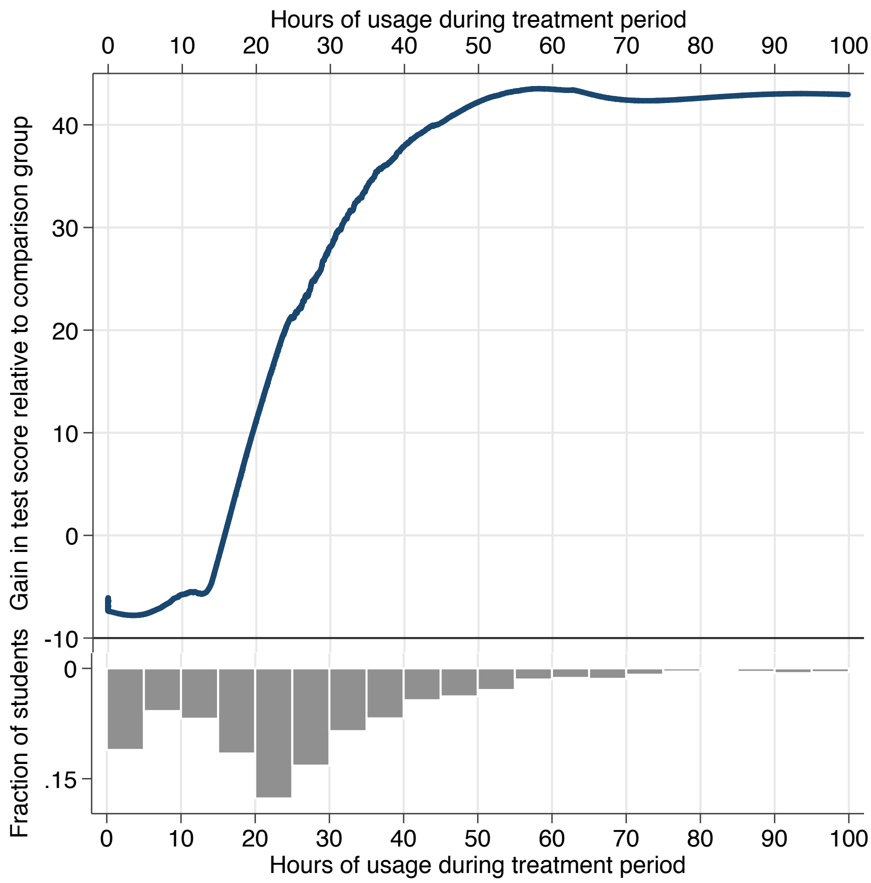
The following graph displays the non-parametric estimate of impact based on usage hours during the treatment period. Students that had below average usage had negative or zero increase in test scores on average. Students using more than that demonstrated strong increases in test scores. Beyond 60 hours per year there seems to be no additional, or even declining impact, however there were few students with usage beyond that level.

These findings are *not causal*: for example, it cannot be claimed that if a teacher makes all of their students use Mindspark for 50 hours that they were increase their test score by 40 points (0.4 s.d.). It could be that the students that used Mindspark more than others were more motivated, had better focus, had better organized teachers, and so on, and there factors that led them to use Mindspark more could be related to faster learning even without Mindspark. While it is impossible to control for motivation and focus and other such unobservable factors, it was possible to see if there is a relationship between baseline test scores and usage: there was no statistical relationship, suggesting that stronger students were not more likely to use Mindspark more.

Ultimately the conclusion is that student who used Mindspark more did better, but only up to 60 hours per year. Beyond that there were no observed learning gains, however, as the bottom part of the graph shows, there were very few students that studied more than 60 hours.

Students that used Mindspark fewer than 15 hours actually did worse than their matched peers that did not use Mindspark at all. This suggests that there is a selection effect driving the slope of the curve, as less motivated or less focused students use Mindspark just a few hours, while more motivated / focused students used Mindspark more.

Graph 12: Increase in test scores by usage hours, adjusted for expected increase



The relationship between usage and test score gains was also investigated parametrically, and this function was estimated: learning gain = $1.7*\text{use\_hours} - 0.01*\text{use\_hours}^2 - 13$, while also controlling for all other parameters in the primary specification given earlier. According to this function, students with median usage (24 hours) would experience a gain of 22.7 points, students at the 75th percentile of usage (36 hours) 35.7 points, and students at the 90th percentile (54 hours) 49.8 points.

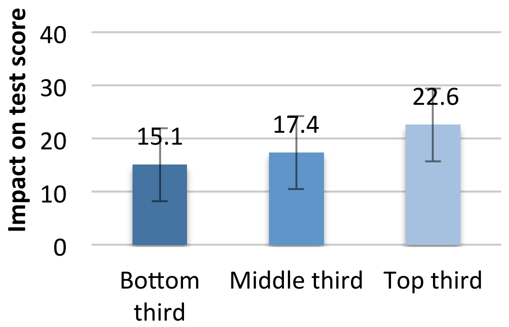
Secondary results

In addition to the overall impact, tests for differential impact were conducted to understand where Mindspark was more effective. None of these findings were statistically significant so should only be used as suggestive evidence, perhaps to generate new hypotheses for testing.

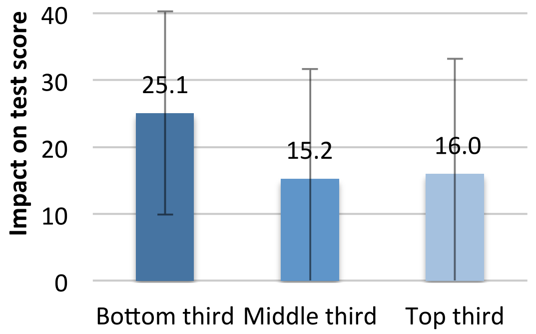
By class rank

In both evaluation years, Mindspark was impactful for students in the bottom, middle and top third of their classes based on baseline test scores, though in the 2011-12 school year the result for students in the middle and top third was not statistically significant, due to a smaller sample size and larger variance. There appeared to be larger impact for students in the bottom third in 2012-13, but this result was not statistically significant.

Graph 13: impact based on students' baseline rank within their class, 2011-12



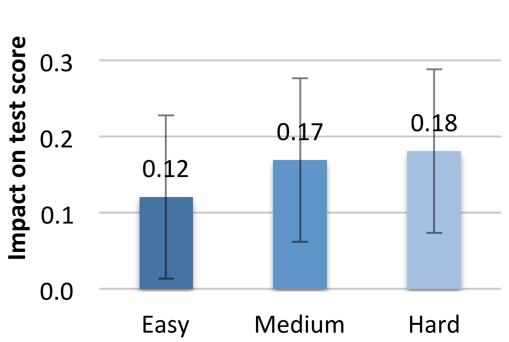
Graph 14: impact based on students' baseline rank within their class, 2012-13



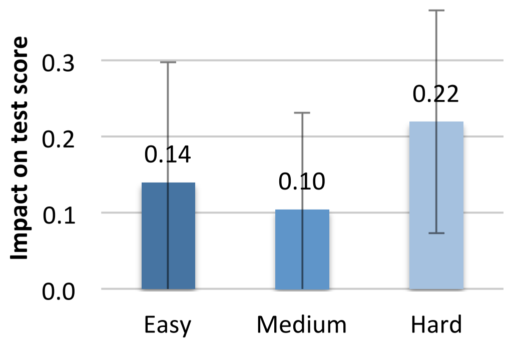
By question difficulty level

In both evaluation years, Mindspark students outperformed their peers on "easy", "medium", and "hard" questions.[6]

Graph 15: impact by question difficulty level, 2011-12



Graph 15: impact by question difficulty level, 2012-13

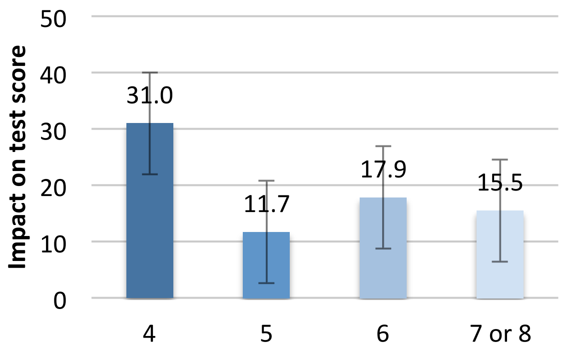


[6] Easy was defined at questions that at least 55% of test takers got correct, while hard was when fewer than 35% of test takers were correct.

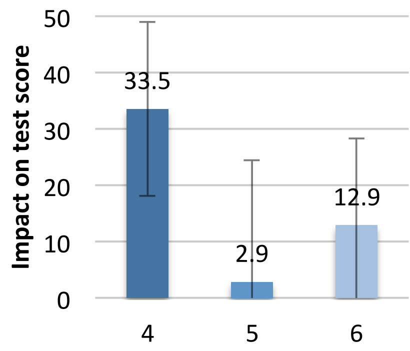
By grade

There is suggestive evidence that Mindspark is more effective for younger students, but while the estimated effect size was larger for 4th grade, this difference was not quite statistically significant. In some cases grades were combined or omitted because the sample size was too small for robust results.

Graph 16: impact by grade level, 2011-12



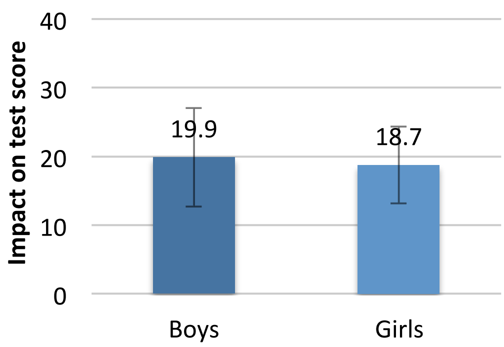
Graph 17: impact by grade level, 2012-13



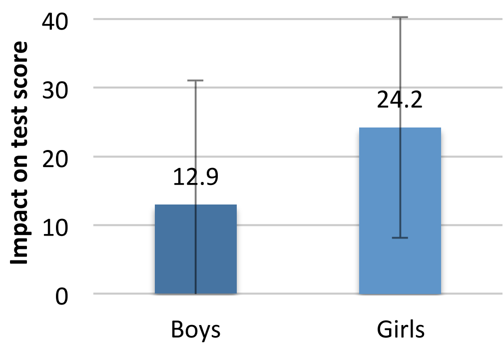
By gender

Mindspark appeared to have a bigger effect for boys in 2011-12, and for girls in 2012-13, but the differences between genders were not statistically significant.

Graph 18: impact by gender, 2011-12



Graph 19: impact by gender, 2012-13

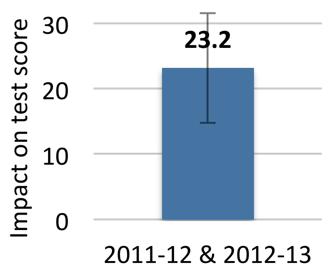


Two-year results

The impact on students that used Mindspark for both school years in the evaluation was only slightly larger than the one-year impact, at 23.2 points (0.232 s.d.). This suggests that Mindspark is most impactful in the first year of use, and has a positive but limited impact beyond one year.

It is also possible that Mindspark continues to be an important instructional tool, but does not produce gains in *conceptual* understanding beyond the first year, which is what is captured by ASSET.

Graph 20: impact over 2011-12 and 2012-23 school years



Discussion and Recommendation

Results in context

Compared to other evaluations of primary school interventions

Mindspark has a large impact on student learning relative to similar interventions.

Three recent reviews of the rigorous evidence on the impact of education interventions in the developing work provide useful comparisons for this evaluation.[7] Looking at similar interventions, the studies find that the mean effect size for "computers and instructional materials" is 0.15 s.d.[8], for "learning materials" 0.16 s.d.[9], and for "pedagogical innovations," including computers: 0.16 s.d.[10]

At 0.19 to 0.2 s.d., the impact of Mindspark is larger than average for these 3 categories, which are the 3 most impactful categories on average in 2 of the 3 studies.

Compared to gains experienced by non-Mindspark schools

Compared to other schools in the evaluation sample, the gains experienced in schools using Mindspark were large. The average increase between the baseline and endline was calculated, and the average Mindspark gains was at the 86th percentile. That means only 14% of schools that did not use Mindspark experienced gains that were are large as Mindspark schools did.

Recommendation

We believe the findings in this report are conclusive, and a randomized controlled trial is not required to document impact in English language private schools similar to those in this evaluation. However, it is yet to be proven that Mindspark can be as effective in affordable private schools and in government schools, or in Mindspark Centres. IDinsight encourages EI to commission a rigorous evaluation to demonstrate impact in these settings, to learn whether Mindspark can be an effective learning tool for those who need it most.

---

[7] The 3 sources are:
McEwan, 2013. Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. Working Paper
Krishnaratne, White, and Carpenter, 2013. Quality education for all children? What works in education in developing countries. Working Paper 20. New Delhi: International Initiative for Impact Evaluation (3ie)
Kremer, Brannen, and Glennerster, 2013. The Challenge of Education and Learning in the Developing World. Science.

[8] McEwan, 2013

[9] Krishnaratne et al., 2013

[10] Kremer et al. 2013

Appendices

Results tables – all matching specifications used

Table 3: Impact estimates for 2011-12 school year, 32 different matching specifications

		Individual level matching groups				
		10	20	30	40	Avg.
Class level matching groups	0	18.9	18.8	19.0	18.9	18.9
	3	19.2	19.1	19.0	18.6	19.0
	6	18.9	18.4	18.4	17.5	18.3
	9	20.1	19.9	19.7	18.9	19.6
	12	18.7	18.5	18.1	16.9	18.0
	15	20.3	19.9	19.3	18.2	19.4
	18	21.4	20.2	20.2	19.8	20.4
	21	18.4	17.5	17.3	17.5	17.7
	Avg.	19.5	19.0	18.9	18.3	18.9

Table 4: Impact estimates for 2012-13 school year, 32 different matching specifications

		Individual level matching groups				
		10	20	30	40	Avg.
Class level matching groups	0	24.8	24.9	24.7	24.6	24.7
	3	23.7	23.7	24.1	23.4	23.7
	6	17.3	16.2	16.8	16.1	16.6
	9	23.2	21.7	22.4	22.0	22.3
	12	21.1	19.1	18.4	19.0	19.4
	15	16.9	14.3	11.7	12.2	13.8
	18	21.3	18.3	18.2	19.2	19.3
	21	24.6	22.9	21.3	22.9	22.9
	Avg.	21.6	20.1	19.7	19.9	20.3

About IDinsight

IDinsight is a development consulting organization that helps policymakers and managers make socially impactful decisions using rigorous evidence. IDinsight's core service tailors experimental evaluation methodologies – including, but not limited to, randomized controlled trials – to the priorities of policymakers and managers.

IDinsight also offers policy design consulting and scale-up support to complement evaluation activities for clients who want to maximize social impact through evidence-based decision-making.

IDinsight has offices in India, Uganda and Zambia serving government, NGO and social enterprise clients working in education, health, nutrition, agriculture, governance, sanitation, and finance.

IDinsight is led in India by Partners Ronald Abraham, Dr. Neil Buddy Shah, Andrew Fraker, and Senior Manager Deeptha Umapathy.

For more information, please visit www.IDinsight.org. For questions on this report please contact Andrew Fraker: andrew.fraker@idinsight.org.