IMPACT EVALUATION OF INOMA'S ONLINE EDUCATIONAL GAMES

Prepared by Armando Chacón and Pablo Peña

For Inoma A.C.

April 2013

The evaluation summarized in this document was performed with the data as provided by Inoma to the authors. The authors are responsible for methodological errors only. They were not remunerated for the evaluation by Inoma or any other organization. Supplemental materials can be requested from the authors at pablo@uchicago.edu.

I. Introduction

Inoma is a Mexican not-for-profit organization devoted to improving educational outcomes by complementing conventional education with online educational games. This note presents the results of an impact evaluation of the exposure of elementary school students in the Mexican state of Puebla to Inoma's online educational games. In its first stage—the subject of this evaluation—the games were designed to help students develop mathematical skills as measured by the national test Enlace.[1]

In order to evaluate the impact of its online games, Inoma carried out a randomized controlled trial (RCT) with students in grades 3 to 6 in a group of public schools in the metropolitan area of the city of Puebla (MACP), in the state of Puebla, Mexico. The RCT took place between February and June of 2012.

II. Sample

Inoma had the support of the local authorities of the state of Puebla to achieve systematic exposure of students to its online games. For the RCT, the state's Ministry of Education allowed students in treatment schools to play Inoma's games during class time, one hour per week. Schools already devote one hour per week to the development of students' digital capabilities—knowing how to use a computer and the Internet—in their "media classrooms" (*aulas de medios*). In treatment schools, that hour per week could be devoted to play Inoma's games, although it was not mandatory.

The state's Ministry of Education considered suitable for the RCT a total of 184 elementary schools located in the MACP.[2] According to official records all of them had media classrooms equipped with computers and Internet connectivity. In a first stage, Inoma randomized which of the 184 schools would be in a sample of 60 for the RCT. In a second stage, it randomly selected 30 schools that would receive the treatment. However, nine schools were thought to have dropped out of the trial—their representatives did not show up to the first information session—and the same number of additional schools was randomly selected to replace the apparent dropouts. As a result the sample grew to 69 schools: 33 schools in the treatment group and 36 in the control group.

Inoma later learned that most schools did not have adequate computers or Internet connectivity. Media classrooms had to be fixed in treatment schools. Nine schools were excluded from the treatment group because of the lack of infrastructure. Additionally, two schools in the treatment group refused to participate and dropped out of the trial. In order to keep balance in the RCT design, some schools randomly chosen were ex-

[1] The test is taken by students in grades 3 to 9 and 12 by the end of every academic year.

[2] Some schools have two "shifts" or *turnos*, one in the morning (*turno matutino*) and another in the afternoon (*turno vespertino*). For the purpose of the evaluation, separate shifts within the same school were counted as separate schools.

cluded from the control. The final sample had 44 schools, 22 in the treatment group and 22 in the control group.

Table 1 presents the sample used for the evaluation. The information displayed is based on the results of Enlace. It only considers schools whose students took Enlace 2012 in 4th, 5th or 6th grade, and Enlace 2011 in 3rd, 4th or 5th grade, respectively. Of the 4,138 elementary schools in the state of Puebla with students that took Enlace, 184 are in the MACP and were supposed to have operational media classrooms: 36 were assigned to the control group, and 33 were assigned to the treatment group. However, 14 control schools and nine treatment schools were excluded from the experiment, and two refused to participate.

Table 1. Sample composition

	Schools	Students
Elementary schools in Puebla (state):	4,138	350,097
Outside MACP or without conditions	3,954	309,978
Inside MACP and in conditions:	184	40,119
Schools selected for the RCT:	69	15,025
Original control:	36	7,167
Final control	22	4,396
Excluded from control	14	2,771
Original treatment:	33	7,858
Final treatment	22	5,584
Excluded from treatment	9	2,274
Dropped out of treatment*	2	337

MACP: metropolitan area of the city of Puebla. Different "shifts" within the same school were counted as separate schools. It only includes students who took Enlace 2012 in 4th, 5th or 6th grade, and Enlace 2011 in 3rd, 4th or 5th, grade, respectively.

Table 1 also shows the number of students in grades 4, 5 and 6 in 2012 who took Enlace in 2012 and 2011—when they were in grades 3, 4 and 5. Over 350,000 students took Enlace, and over 40,000 were in the MACP. In the set of schools deemed suitable for the RCT there were over 15,000 students. In the final control group there were 4,396 students and in the final treatment group there were 5,584.

III. Treatment

Since media classrooms were not operational before the RCT, Inoma had to fix them in treatment schools. Thus, the treatment was not limited to the exposure to online educational games. It also involved fixing the media classrooms to have operational computers and Internet connectivity. In theory, fixing the media classrooms might have had an impact not attributable to playing Inoma's online games.

With the purpose of having evaluation results shortly after starting, Inoma created Enlace proxies.[3] The proxies had between 14 and 19 questions and were applied to both control and treatment schools in February, April and June of 2012.[4] In principle, the application of Enlace proxies alone could result in a better performance in the actual test—students get more practice and teachers become more aware of students' proficiency level. Both the control and the treatment groups were exposed to the Enlace proxies. Therefore, the control group received

[3] Enlace results are available several months after the tests are taken.

[4] Enlace Mathematics tests had between 53 and 71 questions in 2012, depending on the grade.

some treatment in the form of more practice for Enlace.

Playing Inoma's games was a non-mandatory treatment. For that reason, only an intent-to-treat impact can be estimated. The treatment group was "offered" the treatment, but not necessarily all students in the group took it. The impact is averaged across all individuals offered the treatment, including those who voluntarily did not take it. The relevance of the intent-to-treat estimate depends on the actual intervention expected in practice. If the actual policy were not mandatory—or were mandatory but not enforced—then the intent-to-treat estimate would be informative on that policy.

Inoma could monitor which students in treatment schools were registering in its online platform. After noticing a modest number of student registrations in treatment schools by mid-April, Inoma introduced an incentive to increase take-up. It offered Notebook computers to the teacher and the school recording the highest take-up. The additional incentive must be considered as part of the treatment.

The type of interaction that occurred between teachers and students while using Inoma's games is unknown. It could well be that students were left alone to explore and figure out the games by themselves, or that teachers actively coached them. Whatever was said and done by teachers in the game sessions could have had an impact on students' attitude and level of engagement—e.g. did they perceive it as an obligation or recreation? That aspect of the treatment is a black box.

IV. Metric of impact

The Ministry of Education of the state of Puebla shared with Inoma the Enlace scores of all elementary schools students from 2011 and 2012. Since the intervention took place between February and June 2012, the changes in test scores between 2011 and 2012 can be used as a metric of impact. Students in 3rd grade could not be included in the evaluation because Enlace covers grades 3rd to 9th and 12th. In 2011, 3rd-graders were in 2nd grade and therefore were not tested.

There are several considerations about the use of Enlace test scores as a metric of impact. First, Enlace's difficulty might vary from year to year. Consequently, both the mean and the variance of the test scores could change even if students' skills remained constant. In order to control for variations in difficulty, the metric of impact was defined as changes in standardized test scores. In other words, test scores for each year and each grade were "de-meaned" and divided by their standard deviation. Then, for each student we computed the difference between her standardized score in 2011 and 2012. The average difference is zero, and any difference is defined in standard deviations of Enlace.

Second, Enlace is not perfect and provides only a noisy measure of what it intends to capture. For the sake of brevity of the test, a relatively small set of questions has to be selected from a much

larger pool of theoretically equivalent questions. From the perspective of students, there is an element of luck involving their performance in Enlace.

Some students get luckier than others in a particular year because they studied more some topics that appeared in the test. In other words, there is noise in Enlace that causes "regression to the mean" across tests: some of the lucky students that performed well one year are expected to perform worse the next. The same holds at the other end of the distribution. Some of those who performed poorly are expected to do better just because they were unlucky the previous year. The empirical strategy to evaluate the impact of Inoma's online games must account for such regression to the mean.

Third, in some respects the unit of analysis is the classroom—not the student. Students in the same classroom share teachers and resources. Thus, shocks to the performance in Enlace are not independent across students in the same classroom. In order to account for potential correlation of error terms, standard errors in the regression analysis must be clustered by classroom.

Fourth, the contents of Enlace are not exactly the same every year. It is possible that one year's version is more closely related to the contents of Inoma's games than others. That variation cannot be accounted for with a one-year evaluation. Inoma might have gotten lucky or unlucky in 2012. That limitation must be borne in mind when analyzing the results.

Fifth, there are some skills measured by Enlace that Inoma is not intending to improve. In other words, Inoma's games are focused on a subset of what Enlace measures. However, the use of the whole test should not pose a problem. If anything, Enlace is a noisy measure of the subset of skills targeted by Inoma. Since the noise is on the left-hand side of the equation—the dependent variable—it should not introduce a bias in the estimates of the impact.[5]

Enlace proxies' scores could have been used as the metric of impact for the evaluation. However, the use of actual Enlace test scores is favored for several reasons. First, there was non-negligible attrition between the three proxies. An important fraction of students were not properly identified from one test to the next, reducing the sample that could be used in a non-random way.[6] In the case of Enlace, students are properly identified by their CURP.[7] Second, proxies' test scores lack external validity. Inoma created, applied and graded the proxies. In contrast, Enlace is created and graded by the federal Ministry of Education, and it is applied by each school. Third, the proxies might not have been taken very seriously, particularly in control schools. Since nothing was at stake, their comparability with Enlace is not clear. Enlace

[5] Measurement error does produce an attenuation bias when present in right-hand side variables.

[6] Of those who took the February proxy, less than 80% were identified among the takers of the June proxy.

[7] CURP stands for Unique Population Registry Key.

results do have implications for schools—part of the teachers' compensation is determined based on Enlace test scores. Finally, the results using Enlace proxies cannot be easily interpreted. They do not directly translate into commonly used units. Using Enlace means that an impact equivalent to x standard deviations is well defined. Using the proxies, an impact of x standard deviations does not map into changes in Enlace standard deviations—which is a reference point for other interventions.

V. Empirical strategy

Because of the issues that arose with the RCT design that resulted in the exclusion of some schools and the refusal to participate of others, we decided to take a quasi-experimental approach. Instead of comparing means across the final treatment and control groups, we opted for a regression analysis including all schools in the state of Puebla and applied a difference-in-differences (DD) technique. In sum, we compare the change in the performance of students in the treatment group to the change in performance of students in the control group, but we do not exclude any school.[8]

In the DD approach the first difference is defined across years (the change in standardized test scores from 2011 to 2012) and the second is defined across groups (treatment versus control). In order to control for regression to the mean in the test scores of each student, we include in the regressions a third-degree polynomial in the standardized test score of 2011, and we use all students in the state of Puebla. Standard errors are clustered by classroom—in some schools there are several classrooms of the same grade.

Separate regressions are run for boys and girls, and for every grade available (4th, 5th and 6th). For the sake of completeness (and as a robustness check) the results in Spanish are explored in addition to Mathematics.

The regressions consider all schools in the state of Puebla, including those that were excluded from the treatment or control groups and the two that dropped out of the treatment group. Fixed effects are included for each of those categories.

The DD approach implies that no controls for time invariant student or school characteristics are needed. Examples of those characteristics on the side of the student are: educational attainment of the parents, material resources of the household, preferences, IQ. On the side of the school, those characteristics might include: location, facilities, curricula, training of the teachers, staff incentives. All those potential determinants of performance in Enlace are differenced out. The regression equation is:

$$\Delta y_i = \beta_0 + \beta_1 y_i + \beta_2 y_i^2 + \beta_3 y_i^3 + \theta d_i^s + \varphi d_i^{ec} + \gamma d_i^{et} + \mu d_i^{dt} + \lambda d_i^t + \varepsilon_i \quad (1)$$

where y is the standardized score in 2011 and Δy is the change in standardized

[8] If the RCT was properly implemented, then a comparison of means and the DD approach would identify the same parameter. That is not the case if the RCT was not properly carried out.

score between 2011 and 2012.[9] The subscript indicates student *i*. d^s is a dummy for students in schools in the RCT sample of 69 schools, d^{ec} is a dummy for students in schools excluded from the control group, d^{et} is a dummy for students in schools excluded from the treatment group, d^{dt} is a dummy for students in the schools that dropped out of the treatment group, and d^t is a dummy for students in schools in the final treatment group.

The coefficient λ is the parameter of interest: the impact of the treatment. It is estimated comparing the final treatment and control groups. The estimates of φ and γ could be interpreted as decoys. If the randomization and the treatment were properly implemented, the estimates of φ and γ should not be statistically different from zero. Otherwise, they could be interpreted as evidence of selection into the sample.

VI. Results

Table 2 shows the regression results for Mathematics and Table 3 shows the regression results for Spanish. Both tables have the same layout. The top panel displays the results for boys and the bottom panel displays the results for girls. Each panel shows three columns, one for each grade analyzed: 4th, 5th and 6th. All students in the state of Puebla with Enlace test scores in 2011 and 2012 were included in the regressions in Tables 2 and 3.[10] The specification of the regressions is described by equation (1) above. The significance of the coefficients was computed using robust standard errors clustered by classroom.[11]

Table 2. Impact of the exposure to Inoma's online games on Enlace Mathematics

Explanatory variable	Dependent variable: change in standardized test score 2011-12		
	4th	5th	6th
Boys			
Final treatment	-0.004	0.180 **	0.114 *
Excluded from treatment	0.072	0.347 ***	0.152
Dropped out of treatment	-0.263 **	0.089	-0.629 **
Excluded from control	0.004	0.155 *	0.053
In RCT sample (control)	-0.024	-0.092 *	-0.078 *
R square	0.208	0.226	0.202
Observations	62,221	56,352	56,786
Clusters	5,683	5,385	5,350
Girls			
Final treatment	0.038	0.115	0.149 **
Excluded from treatment	0.055	0.381 ***	0.272 ***
Dropped out of treatment	-0.266 **	0.030	-0.504 **
Excluded from control	0.102	0.243 **	0.056
In RCT sample (control)	-0.058	-0.130 **	-0.139 ***
R square	0.223	0.215	0.203
Observations	62,195	56,184	56,696
Clusters	5,674	5,399	5,360

*p<.1; **p<.05; ***p<.01. p-values for one-sided tests using robust standard errors (clustering by classroom). All specifications include as controls a polynomial of degree three in the standardized score in 2011.

The first result to notice is that the coefficient on the dummy "final treatment" is positive and significant in some cases. A positive impact cannot be rejected at 95% of confidence for boys in 5th grade and girls in 6th grade. Additionally, it cannot be rejected at 90% confidence for boys in 6th grade. The point estimates

[9] Standardized tests have mean zero and standard deviation one. Standardization is done for each test and each grade separately.

[10] For students in 4th grade, the tests scores of 4th grade are compared to those from 3rd grade. The analogous comparison applies for 5th and 6th grades.

[11] We identify classrooms using the variable *grupo* within the same grade and school.

that are significant are not negligible. They imply meaningful improvements: between 0.114 and 0.180 standard deviations.

A puzzling result is the estimates for students in schools excluded from the control or treatment groups, and the school that dropped out of the treatment. In four cases they are positive and significant at 95% confidence, and the point estimates range between 0.243 and 0.381. In other words, on average students in schools originally in the RCT and later excluded did better than students in the final treatment group in some cases. Those findings pose questions regarding the validity of the experimental design.

Table 2 also shows that the RCT sample is not representative of all schools in Puebla. The coefficient on being in the RCT sample is negative and significant at 95% confidence for girls in 5th and 6th grades.

In sum, the results regarding the impact of Inoma's online games on Mathematics are not conclusive but promising. There seems to be a positive impact from having been granted access to Inoma's online games.

Table 3 shows a similar analysis for Spanish. Although there was no a priori reason to expect an impact in the performance in Spanish, the results show some evidence of improvements. The estimate of the impact is positive and significant at 95% confidence for boys in 6th grade, and at 90% confidence for boys in 5th and girls in 6th. There is also evidence of students in schools excluded from the treatment doing better in the case of 5th-graders. The impact on Spanish is surprising. In principle, it is possible that Inoma's games freed up resources from the study of Mathematics to the study of Spanish.

Table 3. Impact of the exposure to Inoma's online games on Enlace Spanish

Explanatory variable	Dependent variable: change in standardized test score 2011-12		
	4th	5th	6th
Boys			
Final treatment	0.011	0.154 *	0.132 **
Excluded from treatment	0.000	0.199 **	0.006
Dropped out of treatment	-0.242 **	0.420 **	-0.485 ***
Excluded from control	-0.055	0.038	-0.003
In RCT sample (control)	0.001	-0.007	-0.086 **
R square	0.240	0.210	0.216
Observations	62,221	56,352	56,786
Clusters	5,683	5,385	5,350
Girls			
Final treatment	0.048	0.099	0.124 *
Excluded from treatment	0.035	0.198 **	0.112
Dropped out of treatment	-0.143	0.409 **	-0.486 ***
Excluded from control	0.034	0.072	-0.013
In RCT sample (control)	-0.007	-0.023	-0.050
R square	0.214	0.194	0.198
Observations	62,195	56,184	56,696
Clusters	5,674	5,399	5,360

*p<.1; **p<.05; ***p<.01. p-values for one-sided tests using robust standard errors (clustering by classroom). All specifications include as controls a polynomial of degree three in the standardized score in 2011.

VII. Caveats and concerns

There are several caveats and concerns about the results presented in Tables 2 and 3. First and foremost, it is not clear what the intensity of the treatment was but it was probably modest. Time of exposure was short—a few sessions with only a handful of games. Inoma's platform was a new technology and there might be a learning curve for teachers trying to promote it. Most likely there was not enough time to properly exploit it.

Second, it is unclear to what degree the whole experiment was perceived by teachers and principals as a policy for which results they would be accountable, or just an inconsequential, rather academic experiment in which they played a passive if not indifferent role. It is unclear what their expectations were in terms of the consequences of the evaluation. For instance, would they keep the new responsibility of using Inoma's online games only if the students performed well? Was this perceived as a strategy to substitute traditional methods that would jeopardize the status of current teachers? Those could be factors affecting the interpretation of the results as a policy guide.

Third, the probable contact between school principals of control and treatment groups casts doubt on the purity of the experiment. Some principals were aware that they were in "some sort of competition." That perception could have affected their behavior but it is not clear in which direction.

VIII. Discussion of findings

The comparison of the final treatment and control groups alone provides evidence of a positive effect of the exposure to Inoma's online games. In spite of the short time of exposure to the treatment and its non-mandatory status, a positive impact cannot be rejected at 95% confidence for some students. The point estimates in those cases are not negligible: 0.114 and 0.180 standard deviations in Mathematics test scores.

A wide variety of interventions have been empirically studied in the academic literature in terms of their impact on educational outcomes. Besides interventions that affect the availability of desks, teacher knowledge of the subjects they teach, and teacher absenteeism, there is not much empirical guidance on what actually works. In fact, the more methodological rigor studies have, the more likely they are to reject a positive effect of the intervention under study.[12] In that context we see the evidence on Inoma's online games as very encouraging.

Inoma took a huge step in the right direction by attempting an experimental evaluation of the impact of its online games. The results presented here indicate that important academic gains resulting from playing those games cannot be ruled out.

[12] See Glewwe, Paul, Eric A. Hanushek, Sarah D. Humpage, Renato Ravina, "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010", NBER Working Paper No. 17554, October 2011.