Chapter 7

Effects of the project on students' outcomes and development of survey instruments

This chapter presents the most relevant effects of the OECD-CERI pilot project on students' outcomes and discusses the validation of the survey instruments. Furthermore, it introduces the initial findings of a class-level analysis, with a special focus on the most successful classes and the characteristics of their teachers, students and adopted pedagogical activities. Finally, it discusses the lessons learnt from the pilot phase. These are laid out by main topic and provide suggestions on how the survey operations and instruments can be improved in sight of the validation phase.

The OECD-CERI project

The OECD-CERI project brought together 13 teams from 11 different countries which recognised the importance of creativity and critical thinking for the future development of their students and partnered together in order to foster these skills with innovative pedagogical practices and a shared understanding of what creativity and critical thinking entail. The teams were from both OECD member countries (France [CRI and Lamap teams], Hungary, the Netherlands, the Slovak Republic, Spain [Madrid Community], the United Kingdom [Wales] and the United States [Montessori and Vista teams]) and from non member economies (Brazil, India, the Russian Federation and Thailand).

While the project mainly focused an intervention based on the development of pedagogical resources, a secondary objective was to develop instruments that would be used in a validation phase of the project. To this effect, the OECD developed and field trialled instruments for a quasi-experimental survey design, which consisted of selecting two samples of students and administering to them a set of questionnaires and tests for measuring several outcomes of interest (e.g. creative potential) and relevant explanatory variables (e.g. gender, age). One group was exposed to pedagogical practices aimed at fostering students' creativity and critical thinking (the intervention group), while the other one served as a reference (the control group). The administration of the questionnaires and tests took place twice for both groups of students: once before the intervention group started adopting the new pedagogies (at pre-), and once towards the end of the school year (at post-).

It is worthwhile highlighting what exactly defines the intervention group, as the juxtaposition between this group and the control group will represent the core of the analysis that will follow. The students in the intervention group were those whose teachers actively participated in professional development sessions organised for this project (see Chapter 5). These events provided teachers with common tools describing the main characteristics of pedagogical activities that would foster and assess creativity and critical thinking (see Chapter 2). Nonetheless, teachers in some teams were given complete autonomy concerning the practical development of the activities in terms of content and length. Therefore, not all students in the intervention group were exposed to the same pedagogical activities. Instead, all students in the intervention group had teachers who were part of dedicated professional development events and were exposed to pedagogical activities that were inspired by the principles discussed at these events.

The OECD developed most of the instruments and tests specifically for this project, which consisted of a student questionnaire, an achievement test in science and mathematics, and an achievement test in visual arts and music. All instruments were level-specific, meaning that those administered to primary students were simpler than those developed for secondary students. Furthermore, the project also adopted the EPoC test, a test developed by Todd Lubart, Maud Besançon and Baptiste Barbot which aims at measuring discipline specific creative potential in children and adolescents (Lubart, Besançon and Barbot, 2011[1]). In this chapter, findings at student level will also rely on some of the data obtained through the teachers' questionnaires.

The part of the OECD-CERI project involving the students took place between November 2015 and July 2017, a period that included two complete school years for both the northern and the southern hemispheres. The involvement of the local teams differed across school years, school level and administered instruments, but overall it allowed answers to be collected from more than 17 000 students. Hereinafter, "Round 1" and "Round 2" will be used to distinguish the stages of the project including respectively the first and second year of data collection. The assessments at pre- and at post- were carried out in each of the two rounds, and while some teachers took part in both rounds (33 out of 380, 15 in the intervention group and 18 in the control one), the vast majority of students did not. In fact, only 40 students did, and their Round 2 data were excluded from the analysis for the purpose of this report.

This chapter presents the initial findings describing the effect of the new pedagogical activities on students. These findings only represent a short-term assessment of the effectiveness of the intervention with students, as an assessment of longer term effects was not part of this pilot. Investigating the relevant factors that would allow a deeper evaluation of the findings (e.g. teachers gaining confidence with the new pedagogies over time, students being more able to process and assimilate the new pedagogies and concepts over longer periods) would only be possible through an assessment of longer term effects.

In light of a possible validation phase, this pilot represented an important field test for the different survey instruments, the administration in real life settings and the subsequent comparability of the data across countries. In order to ensure a consistent implementation of the survey operations, the OECD distributed guidelines and recommendations for the administration of the instruments to all teams in the form of a research protocol. At the end of the pilot, the refined research protocol was one of the outcomes of the project in terms of the development of instruments. In terms of instrument selection, teams independently chose which instruments to administer to their participants, usually following the structure of their interventions (i.e. teams focusing on creativity and critical thinking in mathematics prioritised the test in science and mathematics over the one in visual arts and music). Successively, teams were in charge of translating all instruments into their national language(s) and administering them according to the research protocol. The OECD co-ordinated the data collection and was in charge of the data analysis.

The research questions

For the pilot stage of the OECD-CERI project, the research questions addressed in this chapter covered two broad areas: survey issues and students' outcomes. The survey issues concerned all the challenges faced for a successful development of the survey instruments and implementation of the data collection. The students' outcomes, instead, focused on the effectiveness of the intervention at student level, and the aim was to establish whether the findings were encouraging for certain teams, levels, topics or for any combination of these. At this stage, the pilot was meant to constitute a proof of concept: finding out if the intervention worked under certain conditions would represent the rationale for a subsequent validation study.

The different instruments that were developed for this pilot represented one of the key outputs. Almost all instruments were used in the field for the first time, so there were some crucial questions needed to be clarified:

- Did the instruments measure the concepts for which they were built?
- Were the instruments capable of measuring meaningful changes in these concepts despite the relatively short period between the pre- and post-assessments?
- Did the instruments capture all the relevant information for a meaningful analysis?
- Could the instruments be further refined and improved?

In terms of survey design and survey management, the biggest challenges were the heterogeneity of stakeholders involved in the project and the narrow period within which the survey was organised and run. This meant that there were limited occasions for training and support in the use of the instruments and that, in some instances, operational responsibilities were assigned to people with limited prior experience in survey management and data collection. However, it will be informative for the validation phase to get an indication of whether such a complex design could be handled by in house staff rather than outside contractors. Some key questions in this regard were:

- Were survey operations carried out according to the research protocol?
- Were survey instruments used as expected?
- Was data collection carried out according to the research protocol?

Finally, in terms of students' outcomes, the questions were those that underpin most experimental studies:

- What types of effects can be identified?
- What is the role of context?
- Are there differential effects across subgroups of students?

Development and validation of the instruments

As mentioned above, students were administered up to four different instruments: a student questionnaire, the EPoC creativity test, an achievement test in science and mathematics, and an achievement test in visual arts and music. Students completed all relevant instruments twice: before the beginning of the intervention (at pre), and after the intervention was concluded, or the school year approached its end, (at post), ideally an interval of six months between the pre- and post-measurements. The following paragraphs provide a brief description of the characteristics and content of each instrument.

The student questionnaire, which presented only minor differences between its pre- and post-versions, contained several item batteries that allowed building a series of indices of interest, such as the index of positive learning feelings (Dormann, Demerouti and Bakker, 2018[2]; Schneider et al., 2016[3]) or the index of learning dispositions related to creativity and critical thinking (Carr and Claxton, 2002[4]). The questionnaire also included some anchoring vignettes about creativity and critical thinking (King et al., 2004[5]), which allowed evaluating students' understanding of these concepts and self-perception in terms of their creativity and critical thinking. Furthermore, the questionnaire collected some background information on the students and their households (e.g. gender, education level of the household), and on the students' activities inside and outside of school.

The EPoC test was developed to measure creative potential in children and adolescents in different domains of creative thinking and production: artistic-graphic; verbal-literary; social problem solving; scientific, mathematics and music composition (Lubart, Besançon and Barbot, 2011[1]). The test required individuals to produce work (e.g. drawings, stories, problem solutions) which was then assessed in a standardised way. In each domain, there were two types of tasks: divergent-exploratory thinking and convergent-integrative thinking (creative synthesis). The final creative potential measure included both aspects of creativity. For each domain, two equivalent booklets were developed, called booklet A and booklet B, making it therefore possible to conduct a pre post comparison. The EPoC test required 40-50 minutes to be completed.

The achievement test in science and mathematics was built by the OECD with released items taken from two large-scale surveys: Trends in International Mathematics and Science Study (TIMSS, carried out by the International Association for the Evaluation of Educational Achievement [IEA]) for primary students, and the Programme for International Student Assessment (PISA, carried out by the OECD) for secondary students. The test contained open- and closed-ended items on science and mathematics, some embedded questions on students' interest in these subjects, and some questions about teaching practices in their science and mathematics classes. As with the EPoC test, two booklets of equivalent difficulty were developed in order to allow pre post comparisons. Each booklet included 20 items that contributed to the final score at primary level, and 18 items at secondary level. Students at both levels were given 45 minutes to complete the test. Henceforward, this test will be referred to as the STEM test.

The achievement test in visual arts and music was built in house by the OECD. The test contained only closed-ended items on visual arts and music, some embedded questions on students' interest in these subjects, and some questions about teaching practices in their visual arts and music classes. As with the STEM test, two different tests were developed for primary and secondary students, and in both cases two equivalent booklets were developed in order to allow pre post comparisons. Each booklet included 53 items that contributed to the final score at primary level, and 82 items at secondary level. Students at both levels were given 30 minutes to complete the test. Henceforward, this test will be referred to as the VAM test.

All scores and indices discussed in this chapter are country-, level- and discipline-specific. Scores were computed as simple weighted scores, and the weights depended on the proportion of respondents that correctly answered each item in the different countries, educational levels and items. For STEM scores, these proportions were obtained from the TIMSS and PISA survey data (IEA, 2011 [6]; OECD, 2006[7]; OECD, 2012[8]). For VAM scores, instead, they consisted of the proportion of respondents that correctly answered each item in the single countries. If any of the weights were not available, international item- and level-specific weights were used. More complex methods, such as Item Response Theory (IRT) models, were also investigated for computing the achievement scores. However, due to the high correlation between the IRT scores and the simple weighted scores, it was preferred not to use IRT scores in order to ease the interpretation of the results. The majority of the indices were built by means of separate team- and level-specific factor analyses, with the remaining ones obtained by taking the simple average of two items. In the case of the factor analyses, it was assured that configural invariance was respected among all teams and levels. Furthermore, scores and indices depended on the discipline in which the students received the intervention. For example, if a student received the intervention in mathematics, their final STEM score would include only their score to the questions in mathematics. However, if their intervention was in a subject different from mathematics and science, their final STEM score would include their scores to the questions in both mathematics and science.

Between Round 1 and Round 2, the OECD carried out an initial assessment of the instruments' characteristics. The questionnaire and the STEM test saw only minor changes, while several items were removed and replaced in the VAM test. Additional details about the item-selection procedures, the instruments, the way that scores and indices were derived, and the validity checks that were run on them can be found in the Technical Annex.

The study group

Size of the study's populations

The initial sample of the OECD-CERI project pilot included 20 273 students – 8 949 primary school students and 11 324 secondary school students. The smallest sample size was that of the French (CRI) team (354) and the biggest corresponded to the Thai team with 5 021 students. A few schools and classes dropped out before the project even started, which reduced the sample of participating students to 19 129 (8 358 in primary school and 10 771 in secondary school). Of the 19 129 students, 17 291 participated in at least one assessment included in the data collection. With the exception of the Indian team, which achieved a response rate of 64%, the response rate across all the other teams was, on average, 95%.

Based on the high response rates and on the available information concerning a large part of the non-completion mechanisms, the analysis will make the assumption that the response mechanisms for all instruments followed a "missing completely at random" (MCAR) distribution – for further details see the work of Rubin (1976[9]). This is equivalent to assuming that the attrition (or the nonresponse) did not concern some groups of students more than others, or that there was not a selection bias. In most cases of classes or schools dropping out during the project or before the beginning of the data collection, local teams promptly informed the OECD about these occurrences and provided explanations for these events. For example, a few teams reported classes or entire schools dropping out because they had committed to too many research projects and were asked by their governing boards to drop the majority of them. In many other cases, the reasons for the missingness of post data were rooted in operational hiccups and in the misuse of the instruments, which were independent of the characteristics of the students.

As far as the STEM and VAM achievement tests are concerned, the scores of some students were excluded from the analysis because their response rate to the items in the tests was not above an agreed threshold. Such a decision aimed at excluding those scores that were at risk of depending exclusively on the (low) effort put into taking the test rather than on the students' ability. The selected threshold was 70%, which was established at the lowest possible value while maintaining the consequent data loss within acceptable limits. This implies, for example, that if a student left 7 or more items out of 20 blank, then their score was not considered as reliable, and it was thus excluded from the analysis. The overall data loss for the STEM test was 8% for primary students and 6% for secondary students. For the VAM test, data loss amounted to 20% for primary students and 4% for secondary students. However, most of these data losses were registered by teams that also faced relevant operational issues.

Of the 17 291 students who participated in at least one assessment, 12 265 completed at least one instrument both at pre and at post, with 5 703 at primary level and 6 562 at secondary level (Table 7.1). This corresponded to an overall completion rate of 71%, with only a minor difference between the two educational levels: 75% among primary students and 68% among secondary students. The highest completion rate was observed for the French (CRI) team (98%), while the Indian team had the lowest completion rate (28%). In terms of pre and post available data for the single instruments, the teams collected 8 986 questionnaires (with a completion rate of 67%), 7 953 EPoC creativity tests (with a completion rate of 75%), 7 376 STEM achievement tests with less than 30% of missing values (with a completion rate of 62%) and 1 500 VAM achievement tests with less than 30% of missing values (with a completion rate of 50%).

Table 7.1. Number of students who completed an instrument at the beginning of the project and share of those who also completed the corresponding instrument at the end of it, by team

	Questionnaires	EPoC creativity tests	STEM achievement tests	VAM achievement tests	Any instrument
Brazilian team	1 119 (51%)	628 (90%)	981 (31%)	x	1 248 (62%)
British (Wales) team	791 (75%)	821 (89%)	725 (86%)	x	852 (91%)
Dutch team	852 (69%)	652 (56%)	487 (63%)	348 (75%)	874 (73%)
French (CRI) team	325 (96%)	204 (99%)	319 (97%)	x	345 (98%)
French (Lamap) team	207 (0%)	361 (97%)	201 (19%)	x	364 (97%)
Hungarian team	1 272 (89%)	1 214 (62%)	1 286 (87%)	x	1 534 (85%)
Indian team	999 (31%)	x	1 280 (25%)	x	1 793 (28%)
Russian team	860 (66%)	1 310 (64%)	1 547 (41%)	740 (0%)	2 122 (50%)
Slovak team	563 (63%)	619 (90%)	423 (61%)	457 (64%)	652 (88%)
Spanish (Madrid) team	467 (0%)	x	361 (74%)	x	670 (51%)
Thai team	4 333 (86%)	3 645 (85%)	3 426 (84%)	456 (99%)	4 590 (95%)
US (Montessori) team	90 (0%)	242 (38%)	169 (53%)	x	253 (37%)
US (Vista) team	1 621 (51%)	938 (41%)	774 (30%)	246 (45%)	1 994 (58%)
Total	13 499 (67%)	10 634 (75%)	11 979 (62%)	2 247 (50%)	17 291 (71%)

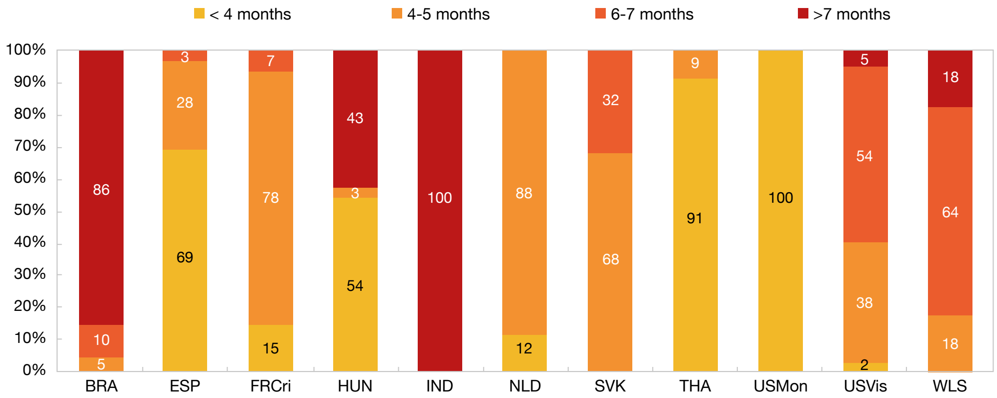
Notes: EPoC: Evaluation of Creative Potential; STEM: science, technology, engineering and mathematics; VAM: visual arts and music. STEM and VAM data only include those students who responded to at least 70% of the items included in the tests.

StatLink https://doi.org/10.1787/888934003307

The intervention with students

The research protocol recommended that the intervention with students take place between the pre- and the post-measurements, and that six to seven months elapse between the two measurements. Unfortunately, few teams were able to stick to the recommendation of the research protocol. As Figure 7.1. shows, only four teams managed to have a time frame of six months or more between pre- and post- data collections for at least 50% of their students, while some have time frames of three months. Some teams do not appear in Figure 7.1. as this information was not available. No remarkable differences could be observed between educational levels.

Figure 7.1. Share of students by length of data collection's time frame and team

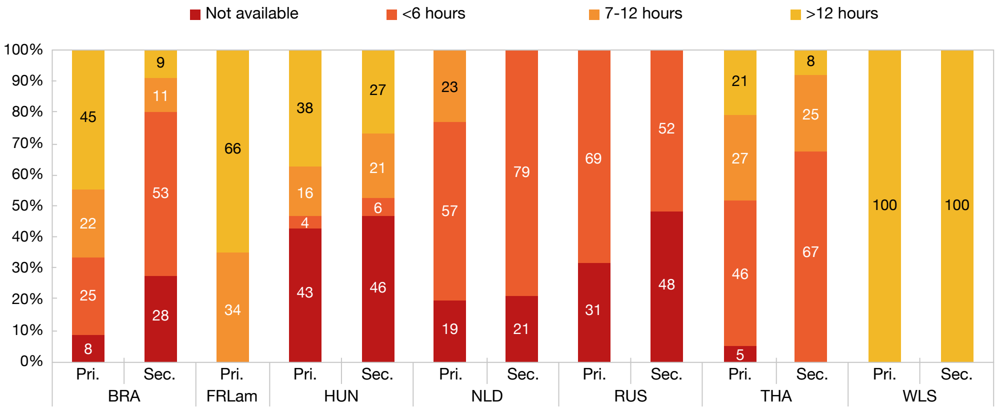


StatLink https://doi.org/10.1787/888934003326

In the context of this chapter, the duration of the intervention refers to the number of hours spent by students on the new pedagogical activities. This is substantially different from, and much shorter than, the duration of the project for teachers. In fact, the intervention with teachers also included meetings of the professional development plan and any out-of-class time spent by teachers to reflect on the new instruments (such as the rubric; see Chapter 2) and to devise and develop the new pedagogical activities.

The research protocol did not contain any explicit recommendation in terms of the duration of the intervention with students, as the local teams had to adapt to very different school contexts. In fact, substantial differences could be observed among teams (Figure 7.2.). Interestingly, the duration of the intervention was longer at primary level for almost all teams for which there was data available. A possible explanation may be based on the number of teaching hours with the same class available to each teacher, which is much higher in primary school than in secondary school. Consequently, a higher number of teaching hours with the same class allows teachers greater flexibility in the organisation of their teaching activities, making thus primary school teachers more likely to dedicate more hours to the project than their secondary school counterparts. Other factors that likely played a role in this context were the different disciplines in which the interventions were carried out (with significant variations in their allotment of instruction time) and the different teaching cultures that exist at primary and secondary level.

Figure 7.2. Share of students by duration of the intervention with students, team and educational level



StatLink https://doi.org/10.1787/888934003345

Taken together, Figure 7.1. and Figure 7.2. show that, from the available data, less than four months passed between the pre- and post-measurements for about 24% of the students, and that the intervention lasted less than six hours for almost 55% of the students. This figure should be taken with caution, as teams did not always time the duration of the teachers' interventions with students, nonetheless it shows that students' exposure to the new pedagogies was relatively short. (Taken at face value, this average corresponds to 1% of the average instruction time for six months in the participating countries [data adapted from OECD (2018[10])]). This highlights the pilot nature of the study and is a reminder that the objective of this phase was not to measure the efficacy of the intervention, but to develop instruments and field trial them. As a result, even where there are enough data collected, the real impact of the intervention could be underestimated due both to the short period elapsed between the two measurements and the limited exposure of the students to it. In fact, even by allowing that the innovative teaching practices might trickle and permeate some of the remaining instruction time, at least 90-95% of instruction time would still be delivered through the established teaching practices.

Characteristics of the study population

This section discusses the main characteristics of the population that took part in the OECD-CERI project. In order to provide further context for the description of the different realities in which the single teams operated, PISA 2015 data (OECD, 2015[11]) were used as a reference value. While collected in 2015, which is two to three years before the data collection for this project, PISA 2015 data provide nationally representative estimates for some of the variables that were also collected in this pilot. By including them, readers can see how the samples participating in this project compared to their respective national populations. However, these comparisons should only be used as simple indications, as teams were not required to be working with nationally representative samples.

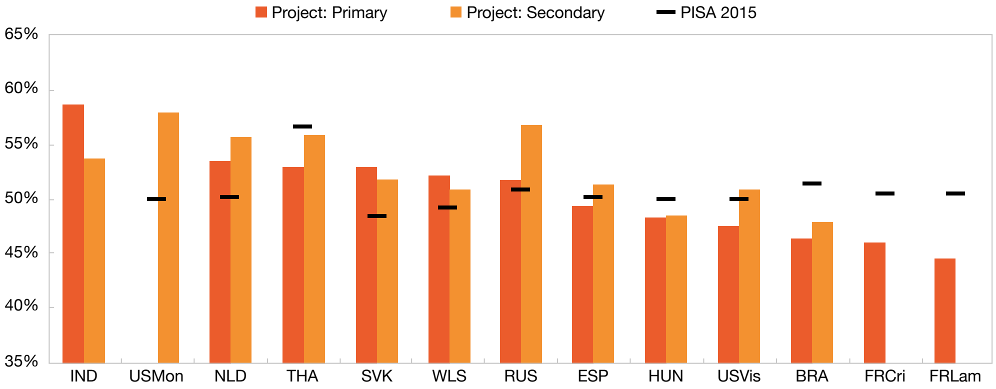
Age

The research protocol recommended local teams to include in their samples 3rd grade classes for primary students and 8th grade classes for secondary students. Nevertheless, the age of the students still showed some variation across the different teams. For primary students, the average age across teams was 8.8 years, with a minimum average of 8.0 years for the US (Montessori) team and a maximum average of 10.1 years for the Brazilian team. For secondary students, the average age across teams was 13.5 years, with a minimum average of 12.5 years for the Indian team and a maximum average of 14.1 years for the Russian team. A very small number of students in the sample belonged to high schools (216 – 174 for the Brazilian team and 42 for the Hungarian one). Due to the limited size of this group, these students were analysed as part of the secondary students.

Gender

The share of girls in the samples was relatively uniform across the participating teams, with a minimum of 45% observed in the French (Lamap) team and a maximum of 58% in the US (Montessori) team. Figure 7.3. presents the share of girls across the participating teams by educational level. This share was similar across the two educational levels for most of the teams, with the exception of the Thai, Russian and Indian teams.

Figure 7.3. Share of girls in the different samples and PISA 2015 reference data, by team and educational level

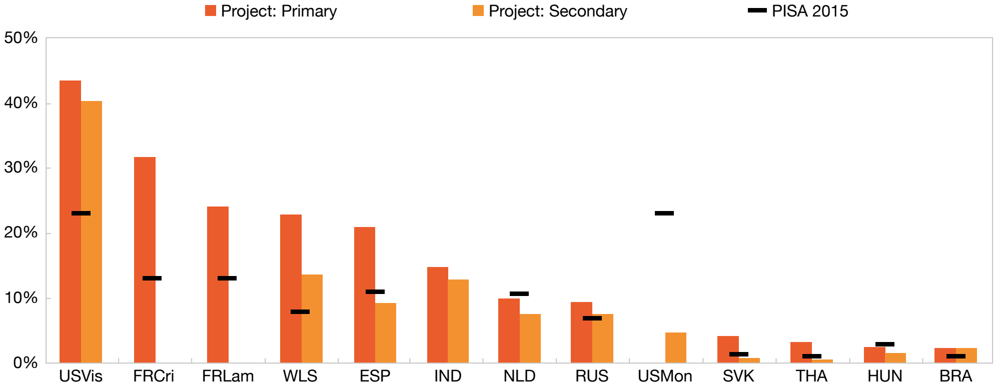


StatLink https://doi.org/10.1787/888934003364

Immigrant background

According to the PISA definition (OECD, 2015[12]), students are identified as having an immigrant background if both of their parents were born abroad (i.e. regardless of the student's country of birth). In this pilot, the share of students with an immigrant background varied significantly across teams, with a minimum of 1% for the Thai team and a maximum of 44% for the US (Vista) team (Figure 7.4.). For a few teams, the share was substantially higher than the PISA 2015 values, but having sample populations with similar characteristics to those of the respective national populations was not a requirement. Possible explanations for these differences could include, but are not limited to: the fact that teams participating in this project were working with more diverse realities in terms of immigrant background than the average schools in their countries; and a real increase in terms of magnitude of the population of students with an immigrant background, which in this project was measured three to four years after the data collection for PISA 2015.

Figure 7.4. Share of students with an immigrant background in the different samples and PISA 2015 reference data, by team and educational level



StatLink https://doi.org/10.1787/888934003383

Figure 7.4. highlights a very low presence of students with an immigrant background for some teams (less than 2.5%), at least according to the PISA definition. However, this definition exclusively takes into account the country of birth of the students' parents (as a couple). When using the information about the country of birth of the students and of each of the parents (collected through the project's student questionnaire), the data highlights several different student profiles, which differ even more when considering the information provided by the variable describing the main language spoken at home by the students. The options in this case were "Main language of the country", "Secondary language of the country" and "Foreign language".

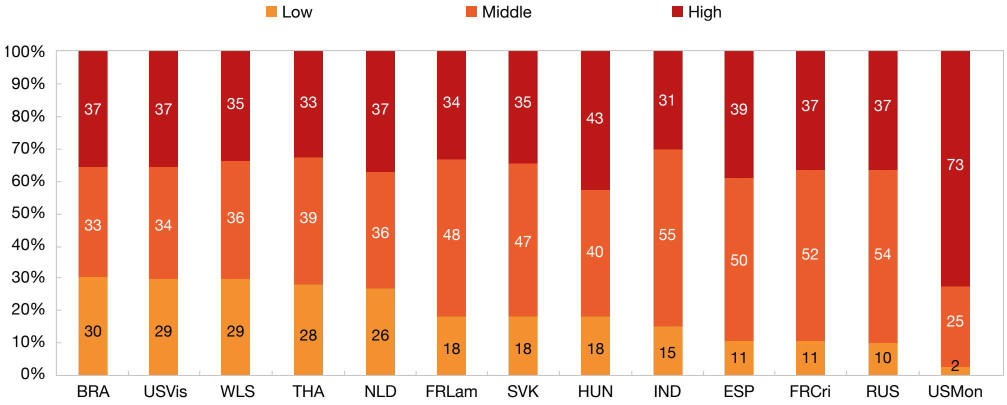
In order to have meaningful results for all teams, a new index describing the immigrant background of the students was built by using all the information provided by the variables about the country of birth of the students and their parents, and the language spoken at home by the students. The

Technical Annex illustrates the differences between the PISA variable and the new index. For the rest of the analysis, the variable used to describe the immigrant background of students will be the new index.

Socio-economic status

The socio-economic status index was built in a country- and level-specific way. At primary level, the index only included information about the possession of books of the households. At secondary level, it also included information about the highest level of education of the parents. The index divided the students into three groups according to their underlying socio-economic status (low, average and high) and aimed at including at least 15% of the students in the low and high categories for every country. Nonetheless, this was not always possible due to the underlying discrete nature of the data, and the shares of students in these two groups sometimes varied significantly across teams (Figure 7.5.). Students in the low category ranged from 2% for the US (Montessori) team to 30% for the Brazilian team, while those in the high category ranged from 31% for the Indian team to 73% for the US (Montessori) team.

Figure 7.5. Share of students with a low or high socio-economic status, by team



StatLink https://doi.org/10.1787/888934003402

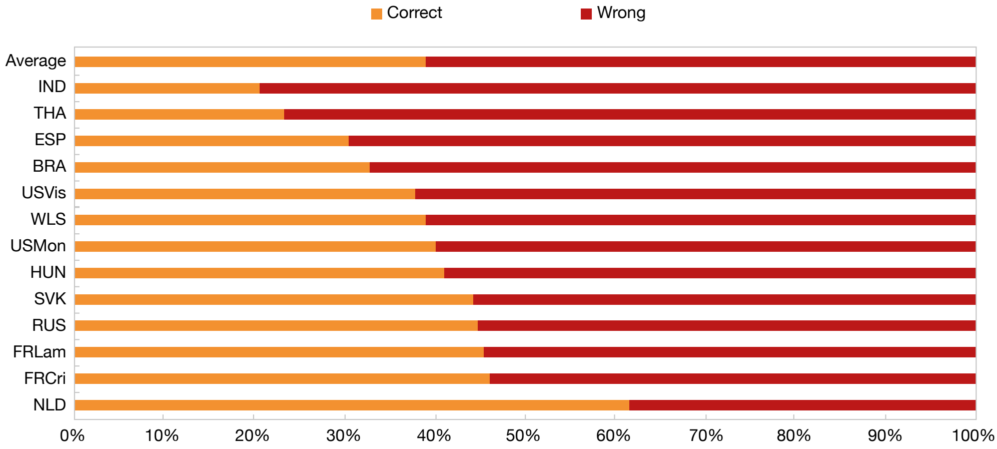
Initial understanding about creativity and critical thinking

The student questionnaire contained two sets of anchoring vignettes that allowed investigation of the students' self-perception of their creativity and critical thinking (the language for primary students consisted of a simplified version of the one used for secondary students). For each skill, these vignettes described three characters with different levels of creativity or critical thinking. Students were first asked to evaluate the level of creativity or critical thinking of the characters (from "Not at all" to "Very much"), and then asked to identify themselves with one of the characters. This allowed assessing the extent to which the students had a correct understanding of these skills by looking at those who correctly ranked the different vignettes.[2]

Furthermore, it was possible to evaluate both their relative and absolute self-perception of creativity and critical thinking. The relative self perception was defined as the level of these skills that the students assigned to the character with whom they identified themselves. For example, if they believed that the character that they identified themselves with was "Very creative", then their relative self perception would result as "Very creative". The absolute self-perception, instead, was given by the a priori level of creativity or critical thinking of that character, as established when preparing the vignettes. If students identified themselves with the character with the lowest creativity, for example, then their absolute self-perception would result as "Little creative", regardless of level of creativity that they assigned to the character. If the students had a perfectly clear understanding of creativity and critical thinking, then the correlation between relative and absolute self-perception should have been close to 1.

Figure 7.6. shows the different baseline understanding of creativity that students had across teams based on the rankings of the vignettes. The average percentage of students that were able to correctly rank the three vignettes on creativity at the beginning of the project was about 40%, but the differences were substantial across teams, ranging from 61% for the Dutch team to 21% for the Indian team. With the exception of the Dutch case, though, the highest percentages were all around 45%, and no remarkable differences were observed between primary and secondary students.

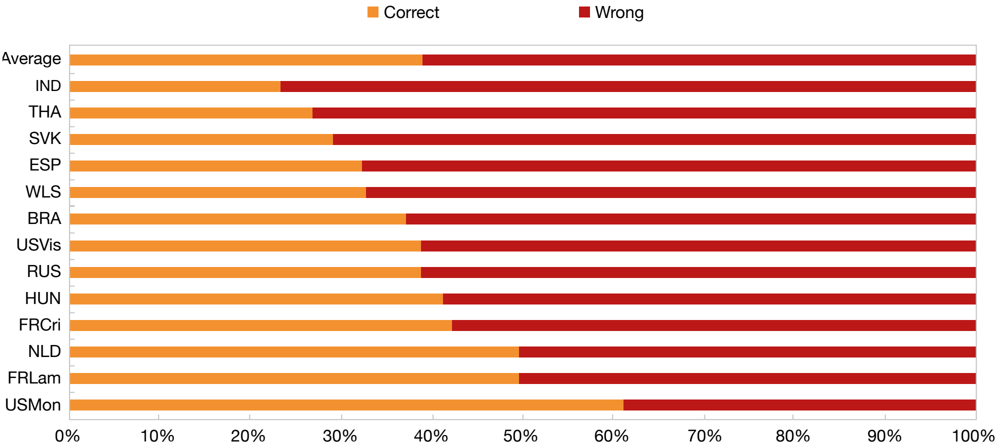
Figure 7.6. Share of students who correctly ranked the vignettes on creativity at the beginning of the project, by team



StatLink https://doi.org/10.1787/888934003421

Figure 7.7., instead, shows the shares of students correctly ranking the vignettes on critical thinking at the beginning of the project. Here, the highest percentage was observed in the US (Montessori) team (61%), while the lowest was in the Indian team (23%). The average percentage was about 40%, but substantial variation emerged between the primary and secondary levels: the average for secondary students was 47%, while that for primary students was only 30%. This difference and the ones that emerged when considering the ranges of these percentages (a minimum of 28% and a maximum of 61% for secondary students versus a minimum of 17% and a maximum of 50% for primary students) suggest that primary students did not have a clear understanding of the different levels of critical thinking presented in the vignettes.

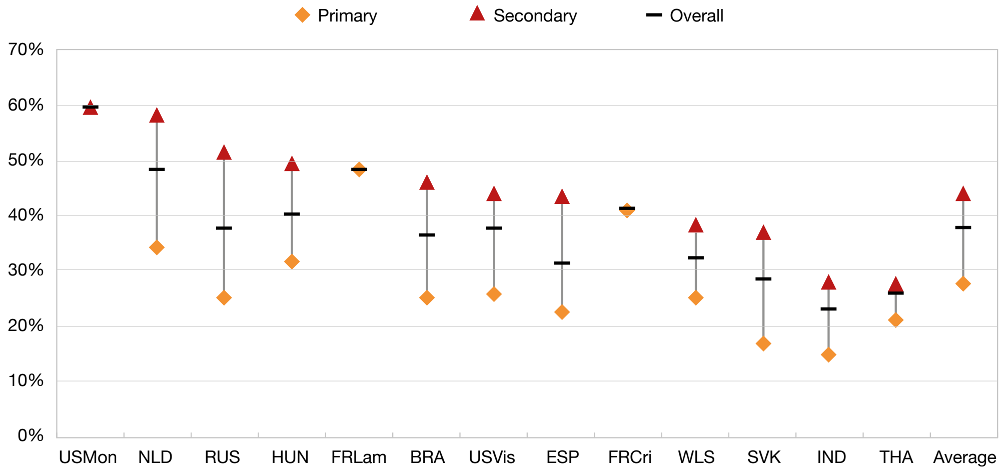
Figure 7.7. Share of students who correctly ranked the vignettes on critical thinking at the beginning of the project, by team



StatLink https://doi.org/10.1787/888934003440

Figure 7.8. presents the same data as Figure 7.7. , but breaks the data down into primary and secondary students in order to show the differences between the responses of the two groups to the vignettes on critical thinking.

Figure 7.8. Share of students who correctly ranked the vignettes on critical thinking at the beginning of the project, by team and educational level

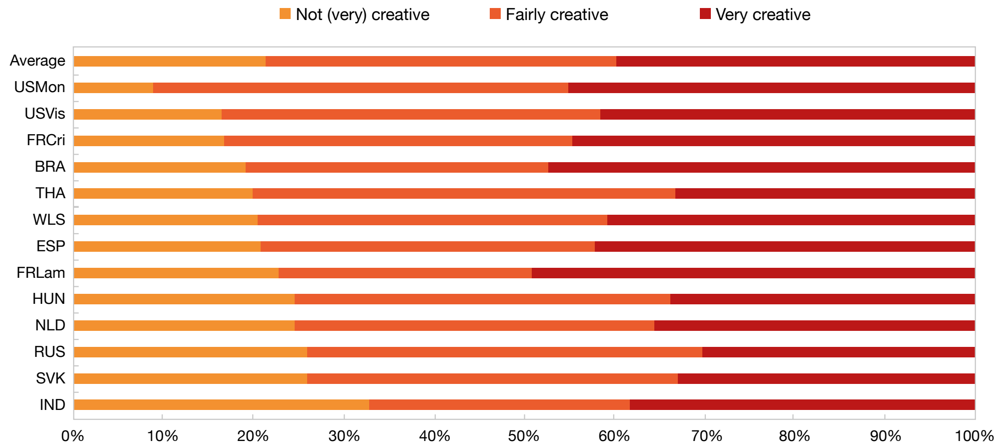


StatLink https://doi.org/10.1787/888934003459

Students' self-perception of their own creativity and critical thinking

As mentioned earlier, vignettes can also be used to investigate the students' self-perception of their creativity and critical thinking. In terms of self-perception of their creativity (Figure 7.9.), the highest percentage of students identifying themselves as being very creative was registered in the French (Lamap) team (49%), while the lowest was in the Russian team (30%). Interestingly, the percentage of students at the other extreme, i.e. identifying themselves as not at all or not very creative, was similar for both teams (23% and 26%, respectively), while for other teams this percentage showed remarkable variations. The minimum was observed for the US (Montessori) team (9%), while the maximum was for the Indian team (33%). However, both the US (Montessori) and the Indian teams seemed to represent exceptional cases, as the share of students identifying themselves as not at all or not very creative varied between 17% and 26% for the other teams. The correlation between relative and absolute self-perception for creativity was around 0.3 for both primary and secondary levels of education.

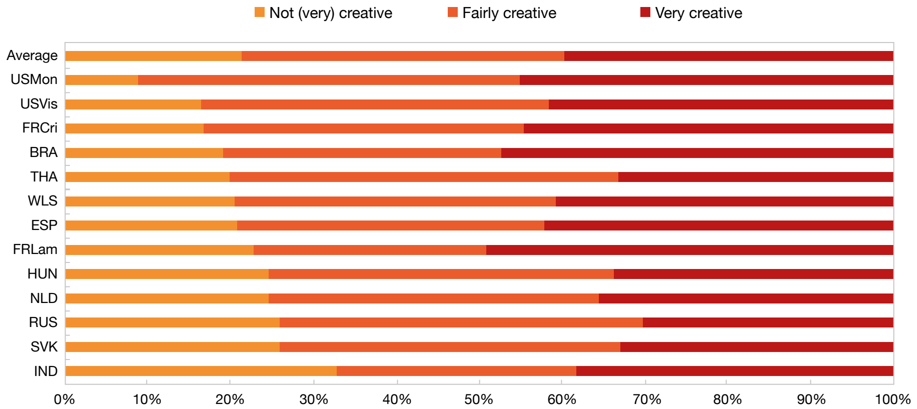
Figure 7.9. Students' relative self-perception of their creativity at the beginning of the project, by team



StatLink https://doi.org/10.1787/888934003478

When considering the students' relative self-perception of critical thinking (Figure 7.10.), the highest percentage of those identifying themselves as being very critical thinkers was registered in the French (Lamap) team (46%), while the lowest was in the Slovak team (23%). A similar variation could be observed in the percentage of students identifying themselves as not at all or not very critical thinkers, which ranged from 18% for the US (Montessori) team to 42% for the Indian team. In the case of critical thinking, looking at the correlation between relative and absolute self-perception seems to confirm the analysis in the previous paragraphs, i.e. that primary students did not have a clear understanding of the different levels of critical thinking presented in the vignettes. In fact, while the correlation was around 0.3 for secondary students (the same value observed in the case of creativity), it was close to 0 for primary students. Possible explanations for this difference for primary students may lie in the simplified language that was adopted for their vignettes (which may not have been as effective as intended), in the natural development of children (as developmental psychology shows that abstract thinking tends to develop during adolescence) or in the presence of a more common understanding of creativity than of critical thinking between children and adolescents.

Figure 7.10. Students' relative self-perception of their critical thinking skills at the beginning of the project, by team



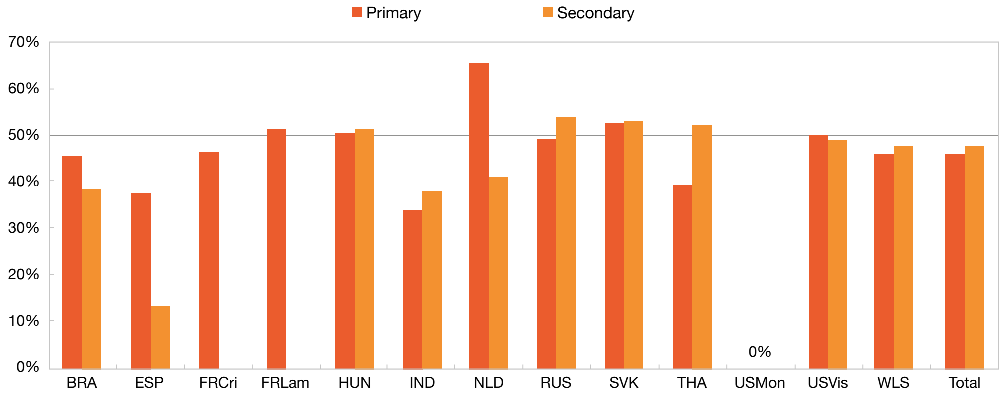
StatLink https://doi.org/10.1787/888934003497

Control and intervention groups

So far, the descriptive analysis has looked at all the students who participated in the project altogether. As mentioned earlier, though, the project followed a quasi experimental design, which implied splitting students into two groups: those in the control group and those in the intervention one. The research protocol recommended recruiting control groups that were broadly comparable to the intervention groups, notably in terms of academic achievement and socio-economic status. Furthermore, it wanted control and intervention groups of similar size.

Significant differences between these groups emerged when their profiles were analysed by team and educational level. This was true for the underlying distribution of some of the main socio-demographic variables of interest and for the sample sizes of the two groups. Overall, out of the 7 620 primary students, 3 486 were enrolled in the control group, while 4 134 were part of the intervention group, leading to an average 46% share of controls. Across teams, though, this share varied between 34% and 66% (Figure 7.11.). On the other hand, the 9 657 secondary students were divided into 4 558 control students and 5 099 intervention students, which led to an average 47% share of controls. As in the case of primary students, the share of controls showed a stark variation across teams – between 13% and 54%. Most of the cases with a substantial unbalance between the two groups occurred due to issues of various kinds faced during the school recruitment process or data collection operations.

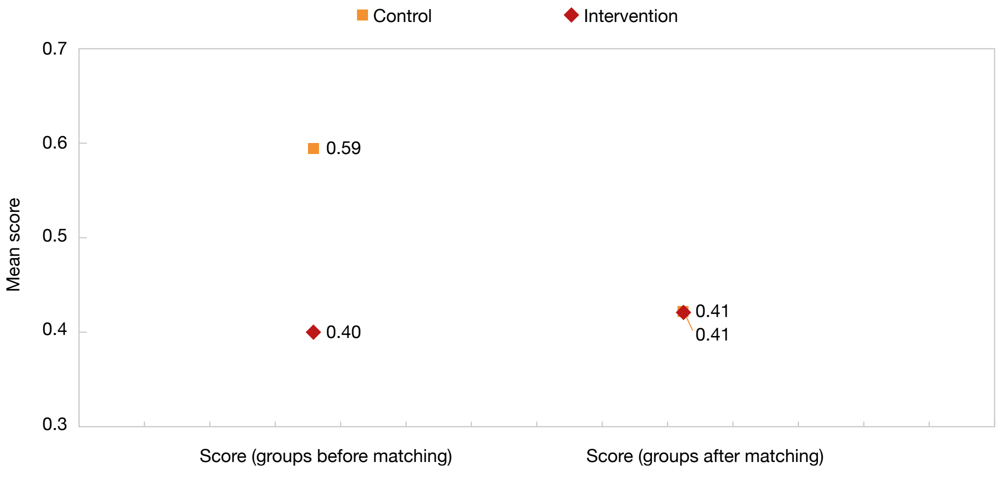
Figure 7.11. Share of control students, by team and educational level



StatLink https://doi.org/10.1787/888934003516

In terms of socio-demographic variables, local teams were encouraged to select schools, teachers and students from a variety of possible backgrounds when possible (e.g. in terms of school size, socio-economic status, achievement level), and to ensure that control and intervention groups would be broadly comparable. In spite of the teams' efforts, these recommendations proved difficult to achieve in practice and there were substantial differences between the two groups. In order to reduce their impact on the findings to a minimum, the first step of the analysis consisted of a propensity score matching (Rosenbaum and Rubin, 1983[13]). This technique realigns the baselines of the two samples by attributing different weights to the students in the control group. A specific set of weights was computed for each of the survey instruments. The weights for the questionnaire were computed in order to adjust potential imbalances in gender, socio-economic status and age. The weights for the tests, additionally, also adjusted potential imbalances in the baseline of their main variable of interest (namely the EPoC, STEM and VAM scores). Figure 7.12. shows an example of the effect of the propensity score matching.

Figure 7.12. Example of the effect of the propensity score matching on the STEM scores at the beginning of the project for primary students in the Thai team



StatLink https://doi.org/10.1787/888934003535

The objective of the propensity score matching was to obtain two groups of controls and interventions that, once weighted, can be assumed to be largely equivalent (at the beginning of the project). Its main drawback is that it implies some data loss if some students did not reply to any of the explanatory variables used for building the propensity scores. In this case, it becomes impossible to compute their propensity score, and they need to be excluded from the analysis. However, this would have ultimately happened anyway when running any analysis including the explanatory variables used for the matching. In order to minimise data losses, the propensity score matching included only the variables listed in the previous paragraph (mostly available for all students) and sometimes accepted minor imbalances between the groups rather than excluding a significant number of students from the analysis. As an indication, the data loss due to the propensity score matching amounted to 3% of the overall data for the questionnaires, 11% for the EPoC creativity tests, 6% for the STEM achievement tests and 4% for the VAM achievement tests. More information about this analysis can be found in the Technical Annex.

Measuring the effects of the intervention with students

The main objective of this pilot was to develop instruments for a possible validation phase of the study. This process included the actual collection and analysis of data, first to assess the validity of the instruments, but also to understand the effects of the intervention, even with a small statistical power.

The following sections discuss the results associated with being in the intervention group on students and the factors that affected these results in a consistent way across countries. The first section introduces the methods used for the analysis, while the second presents the results that were consistently observed across countries for all students. The focus then shifts to specific subgroups of students in order to explore whether the intervention had a differential effect for some sub-populations. All results account for the possible confounding effects of a set of 13 explanatory variables.

Methods

The results derived from a set of multivariate models that investigated the effect of the intervention with students in terms of pre post change for several outcomes of interest. All models were computed with cluster robust standard errors at school level, thus accounting for the hierarchical structure of the data. Furthermore, all models included a set of control variables, such as age, gender, socio-economic status, subject and the value of the outcome of interest at the beginning of the project. When available, the time between pre- and post- data collections and the duration of the intervention with students were also included among the control variables.

The analysis was limited to a discussion of the results in terms of the direction of the findings (positive or negative) due to the high heterogeneity in the fieldwork carried out by the different teams. The current pilot constituted a proof of concept, so the actual magnitude of the findings would have had only limited importance for the assessment of the effectiveness of the intervention with students. Hence, rather than focusing on the size of the different coefficients, the analysis focused more on the patterns that could be observed across different teams and educational levels.

The threshold for statistical significance was not set at the conventional level of 0.05. Initially suggested by Fisher in 1926, the threshold of 0.05 implies that out of 100 tries, the result of interest will be observed in at least 95 occasions. In other words, the conclusion would not be true in 1 out of 20 tries. As the same author pointed out, this threshold was deemed appropriate for establishing "experimentally established" scientific facts (Fisher, 1926[14]), which was not among the objectives of this pilot. Furthermore, as stated by the American Statistical Association in 2016, "scientific conclusions and business or policy decisions should not be based only on whether a p value passes a specific threshold" (Wasserstein and Lazar, 2016[15]).

As the same authors suggest to use the significance "as a tool to indicate when a result warrants further scrutiny", and in order to comply with the explorative and informative – rather than evaluative – nature of this report, the threshold for statistical significance was set at the level of 0.2, meaning that the results presented as statistically significant in this analysis would be observed in at least four out of five tries. All considerations about the need for a more restrictive definition of the statistical significance will be discussed in the context of the future validation study, if any, just like those on the magnitude of single effects. For information, the share of results discussed in Table 7.2. that were also significant at the level of 0.1 amounted to almost 80%, while the share was 65% for those discussed in Table 7.3.

This pilot provided a substantial wealth of data, collecting up to more than 2 000 variables for each of the participating students. For the purpose of this chapter, 36 outcomes constituted the focus of interest. Of these, 18 were taken from the questionnaire (8 of which from the section on vignettes); the remaining 18 were taken from the EPoC, STEM and VAM tests (6 outcomes from each). In order to select the most relevant explanatory variables, the effects of 29 variables on the outcomes of interest were initially explored. Of the initial 29 variables, 13 were retained, which included the time between pre-and post- data collections and 3 main groups of variables concerning, respectively, the students' background, their responses to the vignettes, and their teachers' practices and beliefs.

The final analysis consisted of team- and level-specific models investigating: 1) the effect of the intervention with students after controlling for the above-mentioned set of control variables; and 2) the effect of the interaction of the 13 explanatory variables of interest with the intervention (while still maintaining the control variables in the models). The number of models varied significantly across teams due to the differences in data availability. For point (1), they ranged from 34 for the Slovak team to 3 for the French (Lamap) team, while for point (2) they ranged from 413 for the Thai team to 36 for the French (Lamap) team. The Spanish and US (Montessori) teams were excluded from the multivariate analysis due to data availability issues. Table 7.2. illustrates the findings related to point (1).

Overall results of the intervention with students

The aim of this pilot was to implement new pedagogical activities that would benefit students on several dimensions concerning creativity and critical thinking: their creativity potential, their understanding of these concepts, the use of teaching practices related to these skills by their teachers, their learning dispositions and learning approach towards these skills, etc. Furthermore, it was important to measure the potential effects of this intervention with students in terms of more established metrics, such as scores in achievement tests focusing on STEM or VAM subjects.

For this to be true, students in the intervention group would need to show more improvements in the outcomes of interest than their counterparts in the control group. Additionally, it would be desirable for these findings to be consistent across countries, even if only limited to some subjects, topics, educational levels or other relevant variables.

The intervention with students seemed to have a positive effect: of all the 268 models, 25% showed a statistically significant positive effect while only 18% showed a significant negative effect, for a net total of 7%. The overall impact of the intervention was similar across educational levels, as the net totals were around 7% for primary and secondary students taken separately.

For primary students, the intervention seemed to be particularly beneficial in terms of scores in the achievement tests. More specifically, positive and significant effects were observed for:

- STEM test scores (for four teams out of nine)
- VAM test scores (for two teams out of three).

Table 7.2. Positive and negative statistically significant results associated with the effect of the intervention with students

Instrument	Index or item	Models with positive results	Models with negative results	Total models
STEM test	Teaching practices in STEM (P)	1	1	8
	Teaching practices in STEM (S)	3	0	6
	Interest in STEM (P)	2	3	9
	Interest in STEM (S)	2	2	7
	Score (P)	4	0	9
	Score (S)	1	1	7
VAM test	Teaching practices in VAM (P)	1	0	3
	Teaching practices in VAM (S)	1	0	2
	Interest in VAM (P)	0	0	3
	Interest in VAM (S)	2	0	2
	Score (P)	2	0	3
	Score (S)	2	0	2
Vignettes	Ranking CR vignettes (P)	2	0	10
	Ranking CR vignettes (S)	3	1	8
	Ranking CT vignettes (P)	0	0	10
	Ranking CT vignettes (S)	2	2	8
	Rel. self-perc. CR (P)	4	2	10
	Rel. self-perc. CR (S)	2	3	8
	Rel. self-perc. CT (P)	1	2	10
	Rel. self-perc. CT (S)	3	2	8
EPoC test	Overall score (P)	1	1	8
	Overall score (S)	3	0	6
	Convergent score (P)	2	3	9
	Convergent score (S)	2	2	7
	Divergent score (P)	4	0	9
	Divergent score (S)	1	1	7
Questionnaire	Learning dispositions (P)	1	0	3
	Learning dispositions (S)	1	0	2
	Positive feelings (P)	0	0	3
	Positive feelings (S)	2	0	2
	Single interest (P)	2	0	3
	Single interest (S)	2	0	2
	Parental engagement (P)	2	0	10
	Parental engagement (S)	3	1	8
	School belonging (S)	0	0	10
	Learning approach (S)	2	2	8
TOTAL	Primary students	34	24	130
	Secondary students	33	25	138
	All students	67	49	268

Notes: P = primary; S = secondary; CR = creativity; CT = critical thinking; Rel. self-perc. = relative self-perception. All models included a set of control variables, such as age, gender, socio-economic status, subject, the value of the outcome of interest at the beginning of the project, and, when available, the time between pre- and post- data collections and the duration of the intervention with students. The positive or negative results columns include those models for which the statistical significance of the intervention was less than 0.20.

StatLink https://doi.org/10.1787/888934003554

In both cases, none of the models showed a negative and significant effect of the intervention.

For secondary students, findings tended to be more scattered across the different variables of interest. Nonetheless, positive and significant effects were observed for:

- the use of teaching practices related to creativity and critical thinking during STEM classes (for three teams out of six)
- students' interest in VAM subjects (for two teams out of two)
- VAM test scores (for two teams out of two).

No negative and significant effects were observed for any of these variables.

In terms of scores in the EPoC creativity test, the intervention with students had mixed effects. Its effect among primary students was positive for four teams out of ten, but it was also negative for three of the remaining teams. For secondary students, instead, one team out of seven showed a positive intervention effect and three showed a negative effect. These heterogeneous effects persisted even when looking at EPoC subscores, as divergent and convergent scores presented similarly mixed results.

The time elapsed between pre- and post-measurements also had a positive impact on the amount of positive and significant results (not shown). The longer the time period between the pre- and post-measurements, the better the observed effects. Of the 227 models that included this variable, 33% showed a positive impact of longer time frames while only 15% showed a negative impact, for a net total of 17%. This effect was similarly distributed among primary and secondary students (net totals of 15% and 19%, respectively). This is in line with the recommendations of the research protocol and with the evidence of action research in education, which suggests the need of a sufficiently long time frame in order to be able to measure significant changes in the outcomes of interest. At primary level, EPoC scores were the outcome mostly associated with an improvement when the time between pre- and post- data collections was longer (for three teams out of eight). Longer time frames also seemed to have positive effects on a few indices for secondary students, such as learning dispositions related to creativity and critical thinking, positive learning feelings, and ability to correctly rank the vignettes on critical thinking. In all cases, these effects were measured in two or three teams out of seven, while no significant negative changes were observed.

Results for specific subgroups of students

When looking at the effects of the intervention with students on the different groups of students (Table 7.3.), the interest shifted to finding out whether some sub populations (e.g. girls) benefited particularly from the intervention. To do so, in each of the models presented in the previous section, the interactions between the intervention and the different sub populations of interest were included one at a time.

After considering all the outcomes of interest, the results showed that the following groups of students seemed to consistently benefit more from the intervention across countries:

- students whose teacher believed that creativity could be taught at school when the intervention started (net total of 9%)

- students who correctly ranked the vignettes on critical thinking at the beginning of the project (who showed positive differential results in 18% of the models and negative differential results in 11% of the models, for a net total of 7%)
- students who did not correctly rank the vignettes on creativity at the beginning of the project (net total of 6%).

Table 7.3. Positive and negative significant results associated with the effect of the intervention with students for the different subgroups of interest

Variable	Positive results	Negative results	Overall models	Percentage positive results	Percentage negative results	Net total	Net total (primary)	Net total (secondary)
Gender: Female	34	44	268	13%	16%	-4%	-5%	-2%
Socio-economic status: Low	41	43	265	15%	16%	-1%	-6%	6%
Socio-economic status: High	35	39	268	13%	15%	-1%	-4%	2%
Immigrant background (project definition)	52	33	261	20%	13%	7%	2%	13%
Relative self-perc. of CR at pre: Low	51	29	268	19%	11%	8%	3%	15%
Relative self-perc. of CR at pre: High	27	34	268	10%	13%	-3%	2%	-8%
Relative self-perc. of CT at pre: Low	38	44	267	14%	16%	-2%	-8%	5%
Relative self-perc. of CT at pre: High	35	49	266	13%	18%	-5%	-4%	-7%
Correct ranking of CR vignettes (at pre)	29	45	268	11%	17%	-6%	-6%	-6%
Correct ranking of CT vignettes (at pre)	48	30	267	18%	11%	7%	6%	8%
Longer time between pre- and post-data collections	39	61	220	18%	28%	-10%	-13%	-7%
Higher index of practice change	34	48	256	13%	19%	-5%	-3%	-8%
Teacher correct rank. of CR vign. (at pre)	17	16	81	21%	20%	1%	-6%	13%
Teacher correct rank. of CT vign. (at pre)	14	3	40	35%	8%	28%	42%	6%
Teacher believed CR could be taught (at pre)	15	10	56	27%	18%	9%	4%	13%
Teacher believed CT could be taught (at pre)	7	12	34	21%	35%	-15%	-39%	13%
Discipline of teacher: STEM (vs. VAM)	6	18	59	10%	31%	-20%	9%	-27%
Discipline of teacher: STEM (vs. other)	9	4	27	33%	15%	19%	x	19%
Discipline of teacher: VAM (vs. other)	12	17	62	19%	27%	-8%	-19%	15%

Notes: Self-perc. = self-perception; CR = creativity; CT = critical thinking; vign. = vignettes; rank. = ranking. "Overall models" indicates the number of instances in which it was possible to investigate the effect of the interaction between the intervention and each variable across the 13 teams and the 36 outcomes of interest. Besides the interaction between the intervention and each of the variables, all models also included a set of control variables, such as age, gender, socio-economic status, subject, the value of the outcome of interest at the beginning of the project, and, when available, the time between pre- and post- data collections and the duration of the intervention with students. The reference group for "Socio-economic status" and for the variables describing the relative self perception of the students' creativity and critical thinking at the beginning of the project is "Average". In the case of "Discipline of teacher", the category "Other" groups together all subjects other than STEM and VAM subjects.

StatLink https://doi.org/10.1787/888934003573

Furthermore, additional interesting positive results could be observed by looking at the two educational levels separately. In particular, in secondary schools, the intervention seemed to work better for:

- students with an immigrant background (net total of 13%)
- students who had a low, and then average, relative self-perception of their creativity at the beginning of the project (net totals of 23% and 8%, respectively)
- students who had a low, and then average, relative self-perception of their critical thinking at the beginning of the project (net totals of 12% and 7%, respectively)
- students whose teacher correctly ranked the vignettes on creativity at the beginning of the project (net total of 13%).

For primary students, instead, the intervention seemed to work better for:

- students who had an average relative self-perception of their critical thinking at the beginning of the project (net total of 9% for low self-perception and of 4% for high self perception)
- students whose teacher correctly ranked the vignettes on critical thinking at the beginning of the project (net total of 48%).

Surprisingly, the intervention seemed to have a negative effect at primary level for those students whose teacher believed that critical thinking could be taught at school at the beginning of the project (net total of 39%), while this effect was positive at secondary level (net total of 13%).

In terms of discipline, the interactions between the subject of the intervention and the intervention itself were often not available due to the survey design of the local teams (e.g. all teachers carried out the intervention in the same subjects, all control teachers belonged to one discipline and all intervention teachers to another one). However, 121 models were estimated, 53 for primary students and 68 for secondary students. It was observed that the intervention seemed to work particularly well in subjects other than STEM and VAM (mostly interdisciplinary interventions) for primary students (net total of 19%) and in VAM subjects for secondary students (net total of 42%).

For primary students, the positive effect of the intervention in interdisciplinary projects mostly concerned:

- parental engagement (in two models out of four)
- the students' positive learning feelings (in two models out of four)
- the students' understanding of creativity (ability to correctly rank the vignettes on creativity – in three models out of four)
- the students' curiosity (share of students learning only what they were interested in – which decreased in two models out of three).

For secondary students, instead, the most frequent positive effects associated with an intervention in VAM subjects concerned:

- the students' relative self-perception of their creativity and critical thinking (in two models out of four for both skills)
- the students' learning dispositions related to creativity and critical thinking (in two models out of four)
- the students' learning approach related to creativity and critical thinking (in two models out of four)
- the students' feeling of belonging in school (in two models out of four).

A snapshot of class-level analysis

An additional way to look at the data is to shift the attention from the students to the classes. By doing so, it is possible to use additional information originating from the teachers' questionnaires on the characteristics of the teachers themselves and of the learning environments, and to focus on it. Most of this information had to be otherwise excluded from the student level analysis due to the limited availability of teacher questionnaires.

By using single classes as units of interest for the analysis, it is possible to identify those that showed the most promising results and look at the commonalities between them. Furthermore, in some cases it was possible to link these data with those of the specific new pedagogical activities, thus providing readers with useful benchmarks both in terms of class characteristics and of specific interventions with students.

Methods

The class analysis focused only on a few variables of interest: EPoC, STEM and VAM scores; interest in STEM and VAM subjects; use of teaching practices related to creativity and critical thinking; proportion of students not only learning what they were already interested in; ability of the students to correctly rank the vignettes on creativity and critical thinking; and use of a learning approach related to creativity and critical thinking (only for secondary students). The analysis isolated the top 25% of classes in terms of change pre post for each of these variables – separately for controls and interventions and by level – and then compared this group with the rest of the classes in order to identify its distinctive characteristics.

Two important differences existed between the class level analysis and the student level one presented in the previous sections. The first was that the class level analysis was not carried out for each local team separately due to the wide variation in the number of participating classes for each of the teams. For the purpose of this exercise, carrying out the analysis for all teams together still allowed meaningful conclusions to be drawn from the data. The second difference was that the class level analysis considered variables that could not be included in the student level analysis. Some of the most relevant were: the number of teaching hours with the class per week, whether the teacher felt prepared for fostering students' creativity and critical thinking, the seniority of the teacher, and the class climate. The full list of explanatory variables used for the class level analysis can be found in the Note 5.

Overall, 753 classes participated in the pilot but, in order to ensure a minimal reliability of the estimators, only those with at least five students were included in the analysis. Therefore, the final sample consisted of 732 classes. Class-level data were either derived from the teachers' questionnaires or consisted of class-level averages based on the answers to the students' questionnaire. In the latter case, averages were estimated separately for each variable of interest and only if any of the following conditions was satisfied: the class had a response rate of at least 50%; or the class had at least ten valid answers.

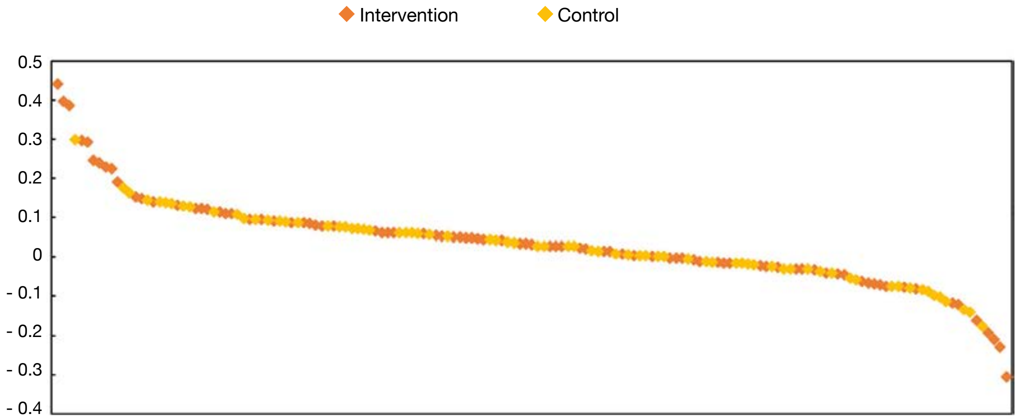
Effects of the intervention with students by outcomes of interest

By looking at the distribution of control and intervention classes in terms of pre post changes in the variables of interest, it was possible to pinpoint those variables for which the intervention with students lead to the most satisfactory results. This turned out to be the case of:

- STEM scores, at primary level (shown in Figure 7.13.)
- the ability to correctly rank the vignettes on critical thinking, at primary level
- the use of teaching practices related to creativity and critical thinking, at both levels.

In the case of Figure 7.13. , for example, the classes with the highest pre post changes appear on the left side of the chart, and here it is possible to notice how the intervention classes, i.e. the circles, outnumber the control ones, i.e. the crosses. Among the first 10 classes, for example, only 1 belongs to the control group, and only 6 control classes appear among the first 20 classes. This suggests a positive effect attributable to the intervention. The conclusions that could be drawn by simply looking at a series of figures like this one were broadly consistent with those that emerged from the student-level data analysis described in the previous sections.

Figure 7.13. Pre-post change in STEM scores at class-level, by group



Note: Classes are sorted by descending magnitude of pre-post change.

StatLink https://doi.org/10.1787/888934003592

Effects of the intervention on students by class profile

Primary students

In addition to the previous findings, the analysis highlighted that, at primary level, the interventions seemed to benefit those classes where the learning climate was challenging at the beginning of the project. When looking at the top performing classes, the share of those with a challenging learning climate was often more than double among the interventions than among the controls, and it was 78% higher on average.

Top performing intervention classes also showed slightly better performances in terms of STEM scores at the beginning of the project and higher socio-economic status than their control counterparts. In terms of teachers' profiles, most top performing classes had teachers with lower qualifications, were less senior, and felt less prepared for fostering students' creativity and critical thinking when the project started.

Secondary students

For secondary students, the intervention still seemed to benefit those classes with a challenging learning climate at the beginning of the project, but the difference was less remarkable than for primary students (only 28% higher on average).

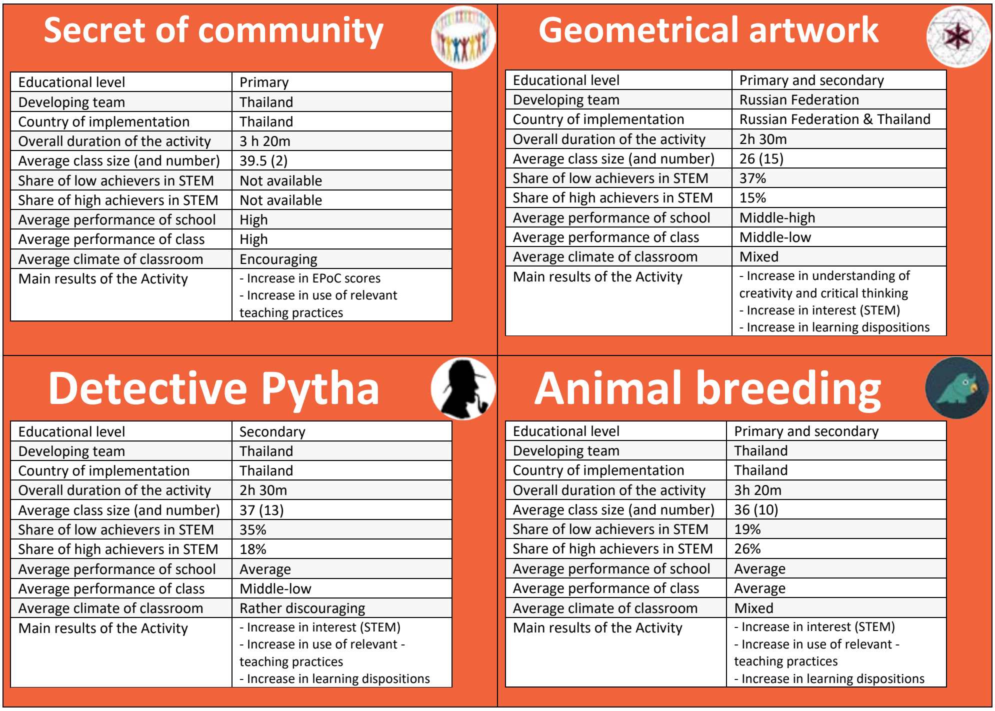
A longer time between pre- and post- data collections seemed to be positively linked with better results, and for almost all outcomes of interest the top performing intervention classes had, at the beginning of the project, teachers who believed that creativity and critical thinking could be taught in schools more than what their counterparts did in top performing control classes.

Activity-level analysis

Finally, the design of the study makes it possible to look at the specific effects that some pedagogical activities seemed to achieve across the different teams. Those listed in Figure 7.14. were selected because they showed very positive results (in the top quartile) in more than five classes (except "Secret of community", which was only used in two classes but that in both cases had excellent results for all variables of interest available). The figure presents the overall profile of the activities, which includes their main characteristics, those of the classes where they were implemented and their most relevant results.

In some cases, local teams provided the OECD with detailed descriptions and lesson plans for some pedagogical activities implemented in the field. Most of these materials were included in the OECD repository of lesson plans, after they were peer-reviewed (Chapter 4).

Table 7.14. Profiles of the most successful pedagogical activities



Secret of community

Educational level	Primary
Developing team	Thailand
Country of implementation	Thailand
Overall duration of the activity	3 h 20m
Average class size (and number)	39.5 (2)
Share of low achievers in STEM	Not available
Share of high achievers in STEM	Not available
Average performance of school	High
Average performance of class	High
Average climate of classroom	Encouraging
Main results of the Activity	- Increase in EPoC scores - Increase in use of relevant teaching practices

Geometrical artwork

Educational level	Primary and secondary
Developing team	Russian Federation
Country of implementation	Russian Federation & Thailand
Overall duration of the activity	2h 30m
Average class size (and number)	26 (15)
Share of low achievers in STEM	37%
Share of high achievers in STEM	15%
Average performance of school	Middle-high
Average performance of class	Middle-low
Average climate of classroom	Mixed
Main results of the Activity	- Increase in understanding of creativity and critical thinking - Increase in interest (STEM) - Increase in learning dispositions

Detective Pytha

Educational level	Secondary
Developing team	Thailand
Country of implementation	Thailand
Overall duration of the activity	2h 30m
Average class size (and number)	37 (13)
Share of low achievers in STEM	35%
Share of high achievers in STEM	18%
Average performance of school	Average
Average performance of class	Middle-low
Average climate of classroom	Rather discouraging
Main results of the Activity	- Increase in interest (STEM) - Increase in use of relevant - teaching practices - Increase in learning dispositions

Animal breeding

Educational level	Primary and secondary
Developing team	Thailand
Country of implementation	Thailand
Overall duration of the activity	3h 20m
Average class size (and number)	36 (10)
Share of low achievers in STEM	19%
Share of high achievers in STEM	26%
Average performance of school	Average
Average performance of class	Average
Average climate of classroom	Mixed
Main results of the Activity	- Increase in interest (STEM) - Increase in use of relevant - teaching practices - Increase in learning dispositions

The class level analysis is an interesting alternative type of analysis, which allows deeper interpretations of the results and is more robust against data missingness. Furthermore, the main results emerged at student level could be found again also in this context. Consequently, it seems that this type of analysis should be an interesting avenue to explore in action research in education. For it to be fully informative, though, it would be recommendable to have a reasonable sample size of classes, allowing a separate class level analysis by country and educational level. This should be taken into account when planning the analytical strategy of a validation phase of the project.

Conclusions

Following the results of the data analysis, the feedback received from the local teams and the evidence that was collected on the behaviour of the instruments, this pilot seems to confirm that the adopted instruments and analytical strategy are appropriate to evaluate the effects of the intervention with students. As the previous sections showed, the instruments and the analytical strategy allowed capturing both positive and negative effects, and then identifying context related factors that influenced them. Furthermore, they allowed for some flexibility in the possible types of analysis, depending on whether the interest lied with students or classes.

The major findings that emerged from data analysis of this initial pilot are the following:

- The intervention with students seemed to have an overall positive effect: of all the models that were estimated, 25% showed a statistically significant positive effect while only 18% showed a significant negative effect, for a net total of 7%. This overall effect was similar across educational levels. If the specific effects are considered, these varied across levels:
    - for primary students, the effect of the intervention seemed to be particularly beneficial in terms of STEM and VAM test scores
    - for secondary students, it seemed to be particularly beneficial in terms of the use of teaching practices related to creativity and critical thinking during STEM classes, of the students' interest in VAM subjects and of VAM test scores.
- Some sub groups of students seemed to benefit particularly from the intervention, in a way that was fairly consistent across countries:
    - students whose teacher believed that creativity could be taught at school when the intervention started
    - students who correctly ranked the vignettes on critical thinking at the beginning of the project
    - students who did not correctly rank the vignettes on creativity at the beginning of the project.

- Finally, some additional positive effects varied significantly within teams in line with the high diversity of the local realities. These are detailed in Chapter 8.

The existence of common results suggests that engaging with teachers and local policy makers to actively foster creativity and critical thinking is possible, and that it can lead to significant and replicable improvements in several outcomes of interest for students. Furthermore, these results are particularly relevant due to the robustness of the findings, which accounted for the possible confounding effect of variables such as gender and socio-economic status.

The effect of the intervention with students seemed to be focused on the achievement test scores among primary students, for whom the results showed clearer patterns compared to secondary students. Two main factors may explain this: on the one hand, primary students might be more receptive to the more open pedagogies used by the teams due to their relatively little previous experience with other more established practices. On the other hand, primary school teachers spend much more time with the same pupils than secondary ones, and this might facilitate a more comprehensive adoption of the new teaching practices beyond what was considered as intervention time.

When considering specific subgroups of interest, the findings of this pilot were particularly encouraging for secondary students. These subgroups mostly belonged to populations that could be defined as disadvantaged either in terms of cultural resources (e.g. low socio-economic background) or from a cognitive point of view (e.g. poor understanding of the concepts of creativity and critical thinking), and many interventions were designed with some of these subgroups as targets. The success of some interventions in closing the existing gaps between these subgroups and the general population in terms of outcomes of interest constitutes a highly relevant and evidence-based result for targeted education policies.

The results were also positive in terms of the development of instruments. In most cases, all the items included in the questionnaire were retained, and configural invariance was observed across teams. This suggest that the instruments could effectively measure the concepts for which they were built (e.g. the different indices of interest). In some cases, the results held only at secondary level, but this could be expected, as the complexity of some indices may not be fully grasped at primary level. The ability of the instruments to measure meaningful changes in the outcomes of interest across countries could not be thoroughly tested due to the variation in the fieldwork conditions across the different teams. However, since they actually measured some significant changes in all countries across the different outcomes, evidence would suggest that they are capable of doing so. Yet, further research in this direction would be recommendable.

Finally, the pilot showed that most costs due to survey management and data collection could be internalised when staff with previous experience in action research are available. If not, some teams resorted to external consultants (often sourced through contacts with academia), which also ensured a good quality of the outputs while maintaining costs at a relatively manageable level. The majority of the teams that managed to adhere the most to the research protocol, use the instruments as intended and promptly communicate crucial details concerning the data collection to the OECD were those where experienced researchers were given the lead for the management of survey operations. As the current context of action research in education mostly relies on limited available resources, the findings of this pilot acquire significant relevance from a policy-making perspective.

Notes

1) For STEM scores, the correlation between simple weighted scores and IRT scores was 0.93 for primary students and 0.84 for secondary students. For VAM scores, these correlations were 0.77 and 0.69, respectively.

2) The correct order of the vignettes was defined as follows: the vignette with the highest level of creativity had to be ranked above or at the same level of the vignette with the average level of creativity; the vignette with the average level of creativity had to be ranked above or at the same level of the vignette with the lowest level of creativity; the vignette with the highest level of creativity had to be evaluated as "Fairly creative" or "Very creative"; and the vignette with the lowest level of creativity had to be evaluated as "Not very creative" or "Not creative at all". The same procedure was applied to the vignettes on critical thinking.

3) The weights consist of the probability of being part of the intervention group given each student's values for a set of explanatory variables.

4) In Table 7.3., the 121 models can be obtained by summing the 59 models for the row "Discipline of teacher: STEM (vs. VAM)" and the 62 models for "Discipline of teacher: VAM (vs. Other)". The 27 models of the row "Discipline of teacher: STEM (vs. Other)" should not be considered for the total, as they are already included in the 59.

5) Teacher's teaching hours with the class per week; performance of the class relative to the country as reported by the teacher; whether the teacher felt prepared for fostering students' creativity and critical thinking; whether the teacher believed that creativity and critical thinking could be taught in schools; the teacher's seniority; the teacher's level of education; the subject matter; the average socio-economic status of the students' household as seen by the teacher; the share of females in the classroom; the immigrant background of students; the class climate; the time between pre- and post- data collections; the duration of the intervention with students; the EPoC, STEM and VAM scores; the proportion of students only learning what they are interested in; and change in use of teaching practices related to creativity and critical thinking as perceived by the students.

6) Classes with a challenging engagement climate were identified as those where the teacher agreed or strongly agreed with at least one of the following items: "When the lesson begins, I have to wait quite a long time for students to quiet down" or "It is difficult to keep the group concentrated for more than a few minutes". Classes were also included in this group if the teacher disagreed or strongly disagreed with at least one of the following items: "Students in this class take care to create a pleasant learning atmosphere" or "Students in this class are generally active and eager to participate in class activities and discussion".

References

Carr, M. and G. Claxton (2002), "Tracking the Development of Learning Dispositions", *Assessment in Education: Principles, Policy & Practice*, doi: 10.1080/09695940220119148, pp. 9-37, http://dx.doi.org/10.1080/09695940220119148. [4]

Dormann, C., E. Demerouti and A. Bakker (2018), "A model of positive and negative learning", in Zlatkin-Troitschanskaia, O., G. Wittum and A. Dengel (eds.), *Positive Learning in the Age of Information: A Blessing or a Curse?*, Springer VS, Wiesbaden, http://dx.doi.org/10.1007/978-3-658-19567-0_19. [2]

Fisher, R. (1926), "The Arrangement of Field Experiments", *Journal of the Ministry of Agriculture of Great Britain*, Vol. 33, pp. 503-513. [14]

IEA (2011), *The TIMSS Assessments website*, https://timssandpirls.bc.edu/timss2011/international-released-items.html (accessed on 12 June 2019). [6]

King, G., C. J. Murray, J. A. Salomon and A. Tandon (2004), "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research", *American Political Science Review*, Vol. 98/1, pp. 191-207, http://dx.doi.org/DOI:10.1017/S000305540400108X. [5]

Lubart, T., M. Besançon and B. Barbot (2011), *EPOC: Évaluation du potentiel créatif des enfants*, Editions Hogrefe, Paris, France. [1]

OECD (2018), *Education at a Glance 2018: OECD Indicators*, OECD Publishing, Paris, https://dx.doi.org/10.1787/eag-2018-en. [10]

OECD (2015), *PISA 2015 Database*, https://www.oecd.org/pisa data/2015database/ (accessed on 12 June 2019). [11]

OECD (2015), *PISA 2015 Technical Report*, OECD Publishing, Paris. [12]

OECD (2012), *Compendium for the cognitive item responses*, https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm (accessed on 12 June 2019). [8]

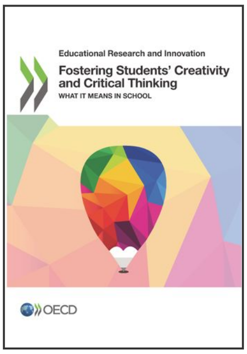
OECD (2006), *Compendium for the cognitive item responses*, https://www.oecd.org/pisa/data/database-pisa2006.htm (accessed on 12 June 2019). [7]

Rosenbaum, P. and D. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, Vol. 70/1, pp. 41-55, http://dx.doi.org/10.1093/biomet/70.1.41. [13]

Rubin, D. (1976), "Inference and Missing Data", *Biometrika*, Vol. 63/3, pp. 581-592, http://dx.doi.org/10.2307/2335739. [9]

Schneider, B., J. Krajcik, J. Lavonen, K. Salmela-Aro, M. Broda, J. Spicer, J. Bruner, J. Moeller, J. Linnansaari, K. Juuti and J. Viljaranta (2016), "Investigating optimal learning moments in U.S. and finnish science classes", *Journal of Research in Science Teaching*, doi: 10.1002/tea.21306, pp. 400-421, http://dx.doi.org/10.1002/tea.21306. [3]

Wasserstein, R. and N. Lazar (2016), "The ASA Statement on p-Values: Context, Process, and Purpose", *The American Statistician*, doi: 10.1080/00031305.2016.1154108, pp. 129-133, http://dx.doi.org/10.1080/00031305.2016.1154108. [15]



Educational Research and Innovation

Fostering Students' Creativity and Critical Thinking

WHAT IT MEANS IN SCHOOL

OECD

From:

Fostering Students' Creativity and Critical Thinking

What it Means in School

Access the complete publication at:

https://doi.org/10.1787/62212c37-en

Please cite this chapter as:

Vincent-Lancrin, Stéphan*, et al.* (2019), "Effects of the project on students' outcomes and development of survey instruments", in Stéphan Vincent-Lancrin*, et al.*, *Fostering Students' Creativity and Critical Thinking: What it Means in School*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/ccff92ba-en