Fundació "la Caixa"

MAPS
Camins matemàtics al pensament computacional

Computational Thinking (MAPS): Impact Evaluation of an Implementation of the Exploding Dots Approach on the Development of Computational Thinking and Motivation for Learning Maths in Lower Secondary School Through an RCT

Evaluation Report

Published on August 26, 2025

Prof. Luis J. Rodríguez-Muñiz, Juan José Santaengracia de Pedro, Carmela Suárez González, Tania Iglesias Cabo, Prof. Celestino Rodríguez Pérez, Prof. Trinidad García, Prof. Irene Díaz, Prof. Belén Palop

educaixa.org

EduCaixa

EduCaixa
Fundació "la Caixa"

Partners:

mmaca Museu de Matemàtiques de Catalunya

GLOBAL MATH PROJECT
Uplifting Mathematics for All

Universidad de Oviedo

Education Endowment Foundation

with the support of

Generalitat de Catalunya

GOBIERNO DE ARAGON

EduCaixa
Fundació "la Caixa"

"la Caixa" Foundation, based in Spain, is one of the largest foundations in the world, and pursues the implementation of social, charitable, welfare, research, educational and cultural activities.

EduCaixa, the education programme within the foundation, offers educational programmes, activities, and digital learning resources with three clear goals: promoting student's development of 21st century skills, fostering the professional development of teachers, and generating and mobilising research evidence for schools.

For more information about EduCaixa and "la Caixa" Foundation or this report, please contact:

EduCaixa
Av. Diagonal 621-629
Torre 2, Pl. 3
08028 Barcelona
Spain

educaixa@fundaciolacaixa.org

www.educaixa.org
www.lacaixafoundation.org/en

Index

About the Evaluator

The impact evaluation of the implementation of the Exploding Dots approach on the development of computational thinking and motivation for learning maths (MAPS project), was carried out independently by a team from the University of Oviedo (Spain): Prof. Luis J. Rodríguez-Muñiz, Juan José Santaengracia de Pedro, Carmela Suárez González, Tania González Cabo, Prof. Irene Díaz, Prof. Celestino Rodríguez Pérez and Prof. Trinidad García, together with Prof. Belén Palop from the Complutense University of Madrid.

The lead evaluator was Prof. Luis J. Rodríguez-Muñiz.

Contact details:
Luis J. Rodríguez-Muñiz
University of Oviedo
Department of Statistics & OR and Mathematics Education
5.7. Fac. of Geology
33007 Oviedo (Asturias)
Spain

luisj@uniovi.es

Acknowledgements

We would like to thank all the teachers and students who participated in the project and, in particular, those who participated in the observation phases and interviews. A special thanks to Prof. Rubén Fernández-Alonso and Prof. Emilio Torres (University of Oviedo) for their valuable comments about the report.

Executive Summary

This project aimed to explore the impact of an intervention based on Exploding Dots (ED) on the computational thinking skills and attitudes towards mathematics of 1st-year secondary school students. The participants, aged 12-13, attended public and state-funded private schools in different regions of Spain. In total, 5262 students from 83 schools (42 in the intervention group and 41 in the control group) took part in the project.

The intervention was implemented in schools by trained teaching staff participating in the study. The training, delivered by a team from MMACA (Museum of Mathematics of Catalonia), consisted of ten in-person hours, delivered during six sessions over a two-day weekend. Additionally, they completed ten asynchronous learning hours, complemented by continuous support from the intervention team to facilitate adaptation to the specific needs of each school.

Teachers were given access to the ED website, which offers various virtual manipulatives. A virtual platform was made available to facilitate the monitoring of the course. Materials associated with the different training sessions were also made available and up to two online group sessions were scheduled. Each teacher received a Teacher's Guide and a Student Workbook for implementing the main six ED units in the classroom. These materials could be accessed via a shared drive created for the teachers. After completing their training, teachers implemented the intervention with a frequency of 1 hour per week during regular maths class hours for a total of 17 weeks, from October 2023 to February 2024.

The evaluation design consisted of a two-arm cluster randomised controlled efficacy trial, with randomisation at a school level. Two outcomes were measured: a primary outcome, regarding computational thinking skills (using an ad hoc instrument) and a secondary outcome, regarding attitudes towards mathematics (using an instrument adapted to the context). The scores of both instruments were compared before and after the implementation of the programme in line with the Statistical Analysis Plan (SAP) of the project. The evaluation process included in-class observations of the intervention group during the programme and was carried out in 34 of the 42 schools (November 2023 and January 2024). In addition, after the observation phases, interviews were held with students and teachers.

The timeline of the project was the following: all participating schools signed a Memorandum of Understanding (MoU) by 31 May 2023. The randomisation of the sample was done on 6 June 2023. The pretest was administered between 25 and 29 September 2023. The programme was implemented from 1 October 2023 to 15 February 2024. The posttest was administered from 19 February to 1 March 2024.

MAPS was commissioned by the "la Caixa" Foundation.

This research project has been funded by the Education Endowment Foundation, in partnership with the BHP Foundation, as part of the "Building a global evidence ecosystem for teaching" project.

About the partners

The MMACA-Museum of Mathematics of Catalonia considers itself heir to a long educational tradition and has rooms dedicated to Emma Castelnuovo, Martin Gardner, Pere Puig and Adam and Lluís Santaló. All have been authorities on a method of teaching and the popularisation of mathematics based on discovery, reasoning and manipulation. One of the museum's objectives, among others, is to support the work of schools and promote collaborative projects with other national and international institutions that share similar goals.

The Education Endowment Foundation (EEF) is an independent charity dedicated to breaking the link between family income and educational achievement. It supports schools, nurseries and colleges to improve teaching and learning for 2 to 19-year-olds through a better use of evidence.

Main Findings

The main findings of the impact evaluation are summarised in Table 1 below.

Table 1: Key conclusions

Students in the intervention group performed moderately better in the posttest on computational thinking skills than those in the control group. However, the result is not statistically significant, indicating that the observed improvement could be due to chance. Both groups improved their performance in computational skills when comparing their pre- and posttest scores.
The attitudes towards mathematics of students in the intervention group were moderately better than those in the control group in the posttests. However, the result is not statistically significant, indicating that the observed improvement could be due to chance. Both groups worsened in attitudes towards mathematics when comparing their pre- and posttests.
During the observation phases and interviews, most students expressed a positive sentiment towards the ED intervention, emphasising that they found it entertaining, interesting and, in general, liked it more than the rest of their maths classes. However, some students reported feelings of boredom and fatigue at the end of the intervention.
In the observation phases and interviews held during the intervention, most teachers pointed out that the use of ED made their classes more inclusive, motivating students who normally did not dare to participate in maths class. They also pointed out that ED can be easily linked with the maths curriculum and suggested it would be more effective if incorporated into a more integrated year-long plan.
The fact that the ED intervention was carried out as an hour separate from the rest of the mathematics classes (which followed the normal curriculum) hindered it from being considered part of the subject, causing a positive attitude among the students towards the experience (as shown by the qualitative information) but not improving their overall attitude towards mathematics.
For the intervention to positively influence attitudes towards mathematics, ED should be completely integrated into the subject of mathematics.

EEF Security Rating

See Appendix A.

Additional Findings

There was no statistically significant impact of the ED project on students' computational thinking skills or their attitudes towards mathematics. The model included contextual variables, which, after analysis, were found to be significant. Follow-up measures indicated that the intervention was faithfully delivered by the team and participating teachers (91.96 % average degree of commitment). The resources provided by the intervention team were adapted to different contexts but maintained a high level of adherence (80.55 % average use of materials) and a high percentage of students participated actively (86.51 % average).

Regarding the primary outcome – the of the intervention on computational thinking – despite the moderate effect of the intervention on the impact intervention group, we cannot guarantee that this was not caused by random factors. While the definition of computational thinking and its component dimensions/skills is still under debate in research, there is some consensus about related skills, such as decomposition, pattern recognition, abstraction, modelling, algorithms and debugging.

The instrument used to research this outcome captured some of these skills, but the results did not show a clear impact of the intervention on computational thinking development. While there is a clear connection

between the construction of number sense (which underpins number systems: representation, estimation and arithmetic) and the abilities of abstraction and pattern recognition, the results highlight a complex and indirect relationship between these forms of thinking. Further research is needed to clarify the elements involved and their role in fostering the development of each.

When considering the contextual variables, a statistically significant effect was discovered of the socioeconomic status on computational thinking skills: high-level children performed significantly better than the rest, particularly compared to low-level children. Significant differences were also observed, although these progressively decreased, between medium-high and medium-low levels compared to low-level children. In addition, there was a moderate influence of gender, in favour of girls, and prior training in computer science or robotics also moderately increased performance in the computational thinking test. The gender effect is relevant because previous studies show that instruments for measuring these skills sometimes have a gender bias favouring boys, which was not the case in this study.

Regarding the secondary outcome – attitudes towards mathematics – some contradiction was observed between the results obtained from the quantitative (pre- and posttests) and the qualitative (observation phases and interviews) research instruments. While the questionnaires about attitudes towards mathematics always talked about "mathematics lectures" (unmodified standardised questionnaires), the observations made during the intervention and the interviews between researchers, teachers and students sometimes compared the programme with "the rest of the mathematics lectures". Therefore, even when the intervention had positive effects, these did not improve attitudes towards the mathematics course as a whole. Furthermore, qualitative results showed a fatigue effect, which is normal in maths courses. The effect of contextual variables on attitudes towards mathematics was greater than on computational thinking. In particular, gender had a high effect (boys had a more positive attitude than girls) and the effect of a high and medium-high socioeconomic status was also considerable.

It is important to note that since the sampling was not completely random, the results are not generalisable in a strictly inferential sense. However, the results can be considered quasi-representative, given the sample composition and size. In other words, the results provide a good approximation to similar contexts.

There are other factors revealed by the observation phases and interviews that are not traceable in the information provided by the questionnaires, such as the impact of ED on students with learning difficulties and more able students. We were able to observe how students with learning difficulties, resulting from different educational support needs (low cognitive development, autism spectrum disorders, attention deficit hyperactivity disorder, etc.), participated in the same activity as the rest of the class and engaged with the ED problems that had been adapted to their needs. Furthermore, for many of them with anxiety or blocking attitudes towards mathematics, ED was the only mathematical task in which they took risks and tried to solve the problems. The same was true for other student profiles (sometimes coinciding with the above) with difficult economic and/or social situations at home (newly arrived students from other countries with different educational systems and/or languages, children of seasonal workers changing schools once or twice a year with little family support to study and even at risk of social exclusion). Many teachers pointed out that these students were motivated by the intervention to engage in mathematical tasks. As for more able students, the response was not uniform. While interviews and questionnaires revealed that they were more easily bored than other students when the teacher failed to enrich the activities or go beyond the task boundary, interviews also found examples of this enrichment and deeper connections.

Looking at all the results, a key factor for successful implementation is the level of integration of the ED activities in the mathematics course, so that students do not perceive them as tasks 'apart' from maths. In the follow-up questionnaire, 100 % of teachers acknowledged that "implementation and curriculum are loosely connected" and "I have sometimes been able to advance curriculum content through implementation". In the interviews, teachers acknowledged that the tasks were very close to the syllabus and could be followed either in parallel or jointly. For example, the Spanish mathematics curriculum includes

integer arithmetic in year 1. In the intervention, most teachers used the usual decimal system approach and then explained how it works with ED (with the anti-dots), but they could also approach the problem directly with ED.

Cost

The only cost of the programme is related to teacher training, as the materials are available for free. Based on the programme offered by the training partner MMACA, the cost is estimated at £8420 (€10 000) for 50 teachers, including follow-up support. This results in a cost of £168 (€200) per teacher and approximately £4.20 (€5) per student, assuming that each teacher works with 40 students (two classes).

Impact

Table 2: Summary of the impact on primary outcome(s)

Outcome/ Group	Effect size (95 % confidence interval)	Estimated months' progress	EEF security rating	No. of pupils	P Value	EEF cost rating
Primary outcome: CT skills	0.074 (–0.011, 0.160)	0	🔒🔒🔒🔒🔒			£ £ £ £ £
Secondary outcome: Attitudes towards mathematics	0.007 (–0.069, 0.084)	1	🔒🔒🔒🔒🔒			£ £ £ £ £

Introduction

Background

Low mathematical literacy in schools has been a recurrent and growing educational challenge in Spain, as evidenced by the latest TIMMS and PISA scores. In fact, maths and computer literacy rates vary considerably across classrooms, school types and socioeconomic strata, due to three main barriers: disengaged students (often blocked by maths anxiety or unmotivated due to low self-esteem), rigid teaching methods and insufficient teacher training in computer literacy and interdisciplinary methodologies. Addressing this gap with an integrated approach that combines maths and computer literacy is crucial for fostering equitable educational outcomes and preparing students for the demands of an increasingly digital world.

The ED intervention, evaluated as part of the MAPS project, could naturally connect mathematics with the ideas of computer science because it covers the arithmetic and algebra of school mathematics, but is not linked to a specific curriculum. Furthermore, ED has been designed to benefit all students in the class, especially those currently 'blocked' and struggling with learning maths.

The underlying hypothesis of MAPS is that arithmetic manipulation and number decomposition (main traits of ED), as the foundations of algebraic understanding, can positively influence the development of some dimensions of computational thinking (CT), particularly those not directly linked to algorithmics and programming.

It is important to note that there is a lack of consensus on the definition of computational thinking. Jeanette Wing, in collaboration with Cuny and Snyder (Cuny et al., 2010), defined CT as "the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent". Other authors have highlighted different relevant aspects of CT, such as the role of the computer in problem-solving (Barr & Stephenson, 2011), the application of computer science tools and techniques in various natural and artificial processes (Royal Society, 2012), the fundamental role of algorithmic thinking (Brennan & Resnick, 2012) and the importance of problem-solving (Grover et al., 2020) or the relevance of data (Palop et al., 2024).

In addition to this lack of consensus, scientific literature often indicates which components, dimensions, facets, or skills constitute CT. These components are understood as parts of the reasoning process inherent to CT, rather than as a series of cognitive factors. Among the most cited definitions and conceptions of CT, three stand out, all sharing the components of abstraction, algorithms and problem decomposition: Barr and Stephenson (2011), who proposed the collection, analysis and representation of data, parallelisation and simulation; Selby and Woollard (2014), who highlighted automation, evaluation and generalisation; and Angeli et al. (2016) and Shute et al. (2017), who considered generalisation, debugging and iteration. Regardless, CT has gained relevance in the international arena and was considered, together with mathematical competence, in the PISA 2022 assessment framework (OECD, 2018), characterised by dimensions such as: abstraction, algorithmic thinking, automation, decomposition and generalisation.

Most instruments for measuring CT, such as the well-known instrument by Román-González et al. (2017), focus mainly on the programming/algorithmic dimensions. These were not suitable for measuring the dimensions of interest in this present study, specifically those related to the impact of basic arithmetic (number positioning and representation) on CT. Therefore, an instrument was developed (the Bebras-Based Assessment for Computational Thinking, BBACT) based on Bebras items (Dagiene & Dolgopolovas, 2022), to focus on these dimensions and avoid an overrepresentation of programming as well as a reliance on any programming language (even block-based ones). In this way, students who had received, for example, Scratch training (a visual programming language, see http://scratch.mit.edu) had no advantage

over the rest. As mentioned, research literature has not yet converged on a single definition for CT, possibly given the broad scope of the paradigm, and a new instrument to refine measurements was needed for this project and welcomed in the field. Studies on the dimensionality of Bebras (Palts et al., 2017) show that there is a mix between factorial and conceptual dimensions, possibly caused by the lack of clear definitions of many of the terms used. In contrast, the instrument used for attitudes towards mathematics is well-known and widely used in the literature.

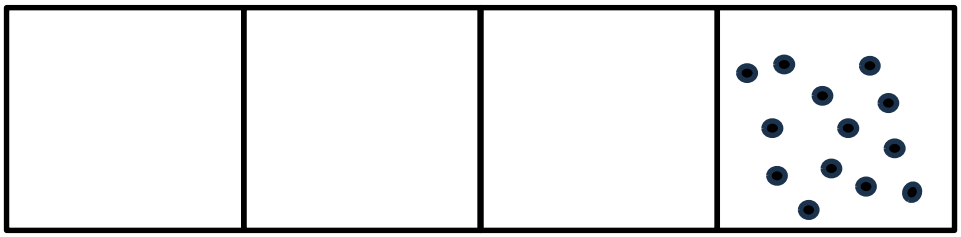
To measure the impact, we used the instrument developed before and after the intervention. In addition, given the relevance of collecting not only quantitative but also qualitative information, we tried to visit all the schools in the intervention arm to verify how the programme was being implemented.

About Exploding Dots

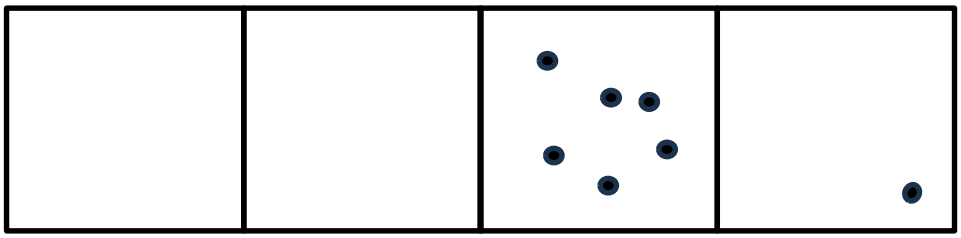
Exploding Dots is an innovative maths learning approach for students in primary and secondary school that was created in the U.S. by James Tanton and is today used by over 6.8 million students and teachers in 168 countries. As previously mentioned, ED focuses mainly on number decomposition and representation, but it also naturally bridges the gap between number and algebraic representation. This makes it an interesting tool to explore the MAPS hypothesis.

ED is a pedagogical approach that introduces the concept of a 'machine' that produces visual representations of numbers and arithmetic operations by processing dots inside boxes according to a specific place value system. Each machine consists of a sequence of boxes plus a rule on two numbers, *m* and *n*. In an *n*←*m* machine, a fixed number m of dots in a given box 'explode' into a number n of dots in the next box to the left, modelling the structure of positional number systems.

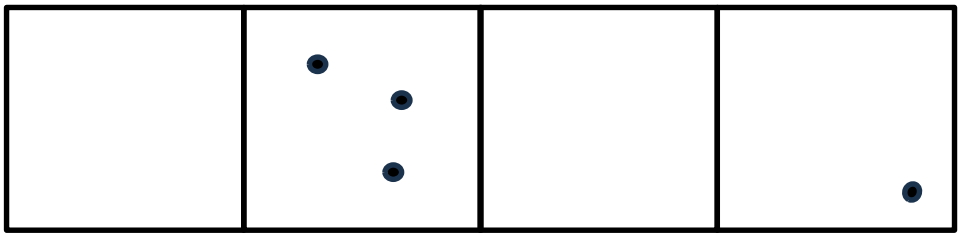
For example, the representation of the number 13 in a 1←2 machine is obtained by placing 13 dots in the first box on the right.



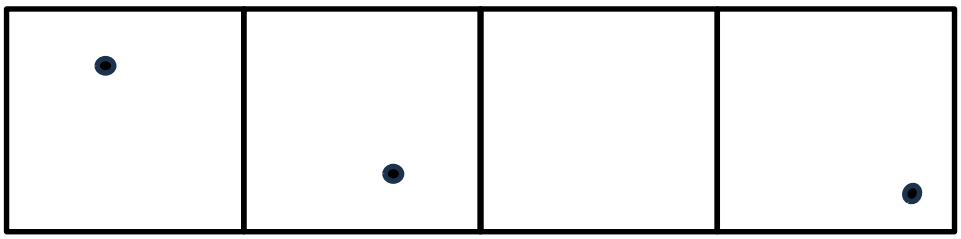
By the 1←2 rule, six pairs of dots in the rightmost box will explode and move as single dots to the next box to the left.



The same will happen again to the six dots in the second box: three pairs will explode as three single dots in the next box to the left.



The machine will ultimately arrive at a final configuration that cannot evolve more. Below is the representation obtained.



The final configuration represents the binary number 1101. This process allows students to observe and abstract the structure of base conversion, building an intuitive bridge, in this case, to binary logic and digital systems in computing.

This visual model can be used for addition, multiplication, division and subtraction. For instance, in the case of addition, it not only reinforces the concept of carrying through visual and rule-based transformations, but also mirrors how operations are broken down into rule-based steps, essential in algorithmic thinking. Learners abstract the 'carry' operation as a rule that can be applied regardless of numbers.

The system provides an intuitive and visual approach to abstract mathematical ideas, making it particularly effective in developing early number sense and supporting abstraction.

The hypothesis that ED improves abstraction skills in computational thinking is plausible given the close correlation between the cognitive processes involved in this approach and those involved in computational tasks. By translating between dot patterns and numerical expressions, students engage in typical processes used in computing – recognising patterns, applying rules and generalising structures. As students move from concrete representations to symbolic reasoning, they develop the ability to model complex systems, reflecting how abstraction works in algorithms and data representation. We believe that, by making the implicit logic of place value and arithmetic operations visible, ED provides a basis for understanding core computational concepts. Therefore, it is reasonable to hypothesise that sustained engagement with this method could lead to measurable improvements in abstraction skills within the realm of computational thinking.

Concerning the evidence or theoretical basis in expecting ED to particularly benefit students from disadvantaged backgrounds, ED is built around the concept of 'low threshold, high ceiling', a concept coined by NRICH. The ED intervention designed for this evaluation is built on three principles:

1. Maths education is a crucial component for the development of programming and computer-science skills. This follows from the constructivist approach of Seymour Papert's (1980) seminal work. One of the possible approaches, probably the most common, to algebraic thinking is to consider it as the generalisation of numerical thinking: variables are considered as unknown numbers in a numerical expression that becomes an equation (Adamuz-Povedano et al., 2021). Algebraic thinking has a relevant intersection with CT (Bråting & Kilhamn, 2021). Therefore, the question naturally arises as to how numerical manipulation and decomposition might influence the dimensions of CT.

2. ED uses manipulative materials (physical and virtual) that teach how numerical representation and agile numerical thinking permeate the entire mathematics curriculum, from primary school base-ten arithmetic to working on decimals in upper primary, and beyond. The concept of the 'machine' provides a natural setting to connect essential mathematics ideas with automatic computation. Freudenthal (1991), Sarama and Clements (2009) and Drjvers (2013), for example, provided evidence of the effectiveness of using both physical and virtual manipulatives.
3. The student-centred approach to inquiry-based learning (IBL) is based on constructivist theory and harnesses children's natural curiosity to learn, focusing on confidence building, critical thinking and problem-solving (Caswell & LaBrie, 2017). While it is difficult to generalise results due to the different forms of IBL, evidence of an impact on students' academic performance and satisfaction has been accumulating over the past decade (Lazonder & Harmsen, 2016; Zafra-Gómez, 2015).

Intervention

In this section, a brief description of the intervention is provided.

Table 3: Description of the educational intervention using the TIDIeR checklist

Brief Name: Provide the name or a phrase that describes the intervention.	An intervention to assess the impact of Exploding Dots on the development of computational thinking skills and the affective dimensions of mathematics education among 1st-year secondary school students.
Why: Describe any rationale, theory or goal of the elements essential to the intervention.	Mathematics, being a fundamental competence, has the lowest student performance and causes the most frustration among teachers in Spain. Despite this, its teaching remains focused on calculations and symbolic manipulations, overlooking two key facts: computers can perform these tasks and, in particular, that the essential skill today is knowing how to instruct a computer to perform a calculation that solves a given problem, which is broadly known as computational thinking. Although the curriculum attempts to address this need and several efforts have been made to integrate computational thinking into the maths class, most current efforts focus on programming and algorithms, neglecting the importance of a solid foundation in the numerical and algebraic sense, as addressed in Papert's seminal work on mathematics education. Exploding Dots offers a powerful and intuitive approach to developing numerical and algebraic thinking from a fundamental and abstract perspective. This intervention aims to assess how a deep understanding of numbers and algebra contributes to the development of key computational thinking skills, using Exploding Dots as the core method.
What (Materials): Describe any physical or informational materials used in the intervention, including those provided to participants or used in intervention delivery or training of intervention providers. Provide information on where the materials can be accessed (such as online appendix, URL).	Physical and digital materials from Exploding Dots, including interactive digital tools available through the Global Math Project website, printed resources available at the EduCaixa website and specific materials developed for MAPS.
What (Procedures): Describe each of the procedures, activities and/or processes used in the intervention,	This intervention was implemented in classrooms during the first half of the school year (October-February), combining guided teaching with

including any enabling or support activities.	hands-on activities. Teacher training sessions and school visits for observations were carried out.
Who provided: For each category of intervention provider (such as psychologist, nursing assistant), describe their expertise, background, and any specific training given.	Maths teachers who had previously received 20 hours of training in Exploding Dots. The training covered theoretical foundations, implementation in the classroom and follow-up of the intervention.
How: Describe the modes of delivery (such as face-to-face or by some other mechanism, such as internet or telephone) of the intervention and whether it was provided individually or in a group.	Classroom teaching, supported by digital resources. Activities were conducted in small groups of 3 to 4 students as advised during the training, with a clear protocol for working in groups.
Where: Describe the type(s) of location(s) where the intervention occurred, including any necessary infrastructure or relevant features.	Public and state-funded private schools in Catalonia, Andalusia and Aragon (Spain). Basic infrastructure was required, with access to digital devices only during the pre- and posttests.
When and how much: Describe the number of times the intervention was delivered and over what period of time, including the number of sessions, their schedule and their duration, intensity, or dose.	The intervention was implemented for 17 weeks, lasting one hour per week.
Tailoring: If the intervention was planned to be personalised, titrated or adapted, then describe what, why, when and how.	Adapted to different levels of mathematical competence. Teachers could adjust the difficulty of activities based on student needs.
How well (planned): If intervention adherence or fidelity was assessed, describe how and by whom, and if any strategies were used to maintain or improve fidelity, describe them.	A monitoring system was designed that included school visits and surveys for teachers and students to assess implementation fidelity. Each teacher's implementation plan was assessed by the intervention team through the Google Classroom platform. Teachers received feedback and corrections and their plans were given a score by the intervention team.
How well (actual): If intervention adherence or fidelity was assessed, describe the extent to which the intervention was delivered as planned.	Implementation varied slightly depending on the school context. Overall, the intervention was implemented as planned, with positive feedback from both students and teachers.

The intervention team was composed of two members of the MMACA, both mathematicians with experience as secondary school teachers and teacher trainers within the Department of Education of Catalonia. They are also involved in the promotion of mathematics in informal learning environments, such as the MMACA.

The first part of the intervention was teacher training. It consisted of ten in-person hours, delivered during six sessions over a weekend (two days) and ten hours of asynchronous learning, with follow-up by the intervention team to facilitate adaptation to the characteristics of each school. The weekend sessions were held in three different locations (Barcelona, Zaragoza and Sevilla). For the asynchronous work, the teachers had access to the ED website, where all the virtual manipulatives were available. A virtual space was also

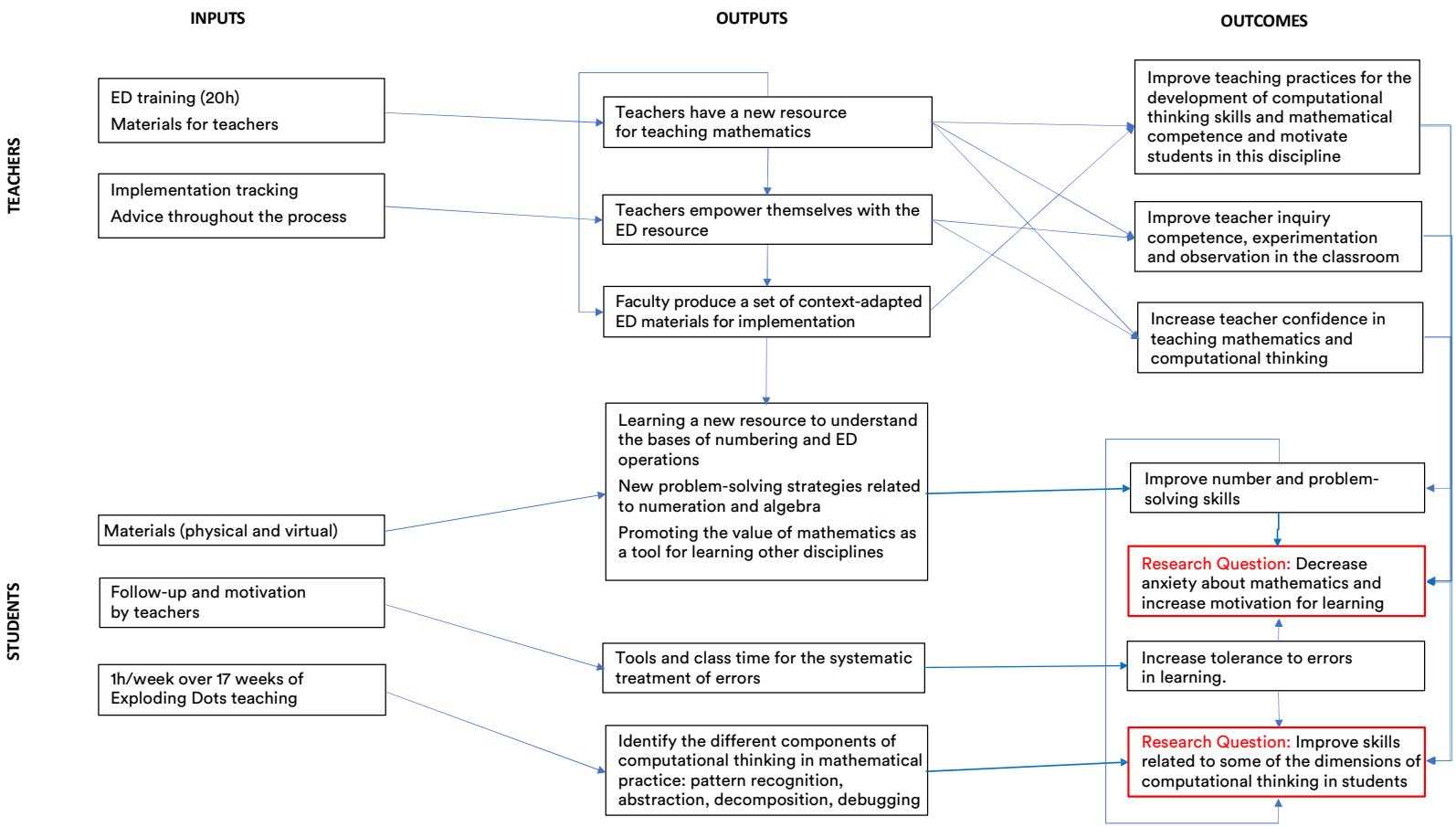
available to facilitate the monitoring of the course. The documents associated with the different sessions were posted and opportunities for online group sessions were provided. Each teacher was given a Teacher's Guide and a Student Workbook for classroom implementation of the six main ED units. These materials were uploaded to a drive made available to them. After the in-person training, teachers adapted the materials to their context and established their calendar. The intervention team assessed the adapted materials, also providing suggestions for improvement, to ensure that the implementation conditions were respected, in particular, the distribution of one hour per week during the half-year.

During recruitment, one school withdrew from the study due to teacher absence. A few issues arose in some of the schools during the intervention, such as power outages (specifically, during the pretest in one school, so the instruments were administered on paper) and the irregular attendance of some of the students, which led to missing data. This is discussed in more detail in the results section.

One school split the classroom into three smaller groups with a different timetable than the one stipulated, so it was eliminated from the study. Another school added a teacher rather than a trained one, so it was also eliminated. In addition to power outages, some technical issues were detected during the pretest, such as some images in two items that, depending on the device, were not properly displayed. This was fixed for the posttest and during the pretest, the images in PDF format were sent to the participating schools, solving the problem. Support for students with special needs was inconsistent: in some schools, the teacher included them in the study, while others chose to exclude them, resulting in unequal treatment between the groups.

The intervention followed the logic model displayed in the figure below.

Figure 1: Logic model of the intervention



Evaluation Objectives

The primary objective was to measure students' computational thinking skills. This was done using an ad hoc instrument (the BBACT test), designed by the evaluation team. The secondary outcome was to measure attitudes towards mathematics, using an instrument taken from the Spanish version of the Attitudes Toward Mathematics Inventory (the IAM test).

Impact Evaluation

Primary question

Does the Exploding Dots intervention improve students' computational thinking skills, as measured by the BBACT test, for the intervention group compared to the control group?

Secondary question

Does the Exploding Dots intervention improve students' motivation and enjoyment in learning mathematics, as measured by the IAM test, for the intervention group compared to the control group?

Implementation and Process Evaluation

Pre-intervention:

RQ1) Have teachers actively participated and engaged in the training?
RQ2) Did the training generate exchanges of points of view and discussion among the teachers?
RQ3) Have families been informed? What was their reaction?
RQ4) Have teachers mastered the teaching resources based on ED?
RQ5) Have teachers created teaching materials based on the resource?
RQ6) How different is Exploding Dots from 'business as usual'?

Early intervention:

RQ7) Did the intervention team solve the doubts that arose during the process?
RQ8) Has the intervention team been receptive to the contextual situations of the schools?
RQ9) Has the implementation been linked to the contents and competencies of the maths course?
RQ10) Has contact with the team been fluid and dynamic during the implementation?
RQ11) Have the control groups changed their methods in relation to maths and computational thinking?
RQ12) Did the control groups implement the pretest?

During the intervention:

RQ13) Have students been actively involved and motivated during the implementation?
RQ14) Have students positively assessed the experience?
RQ15) Has the implementation gone according to the schedule?
RQ16) Have the designed materials been used in the implementation?
RQ17) Have teachers solved students' doubts during the implementation?

Post-intervention:

RQ18) What is the teacher's perception of the effectiveness of the programme?
RQ19) Did the teachers identify any unexpected issues during the implementation of the programme? What did schools need to have in place for the programme to go as planned and what were the challenges they faced?
RQ20) What is the teacher's perception of the support received from the school?
RQ21) Unintended consequences – is there any evidence of a displacement of provision?

The protocol and SAP of the project can be found **here**.

Ethics and Trial Registration

Once schools agreed to participate, they signed a Memorandum of Understanding (MoU). The MoU provided details of the project and its ethical aspects. It also confirmed that any information obtained from the students would be anonymised. The families gave their consent for their children's participation by signing a form that detailed all aspects of the intervention, as well as the anonymisation and storage procedures of the data. The regional education authorities also agreed to the participation of their schools. The Ethics Committee of the University of Oviedo approved the evaluation design in its meeting on 9 November 2022 (reference 15_RRI_2022).

Data Protection

All the collected data was treated as confidential and used solely for the study. The evaluation and intervention teams did not use student or school details in any report. The only student data required for the study was age and gender, ensuring student anonymity. The data requested about the schools was for statistical purposes: to determine the type of school, location and size. Also of interest to the study, was knowing the students' prior experience with programming, computing or robotics, as well as the teachers' training in these subjects.

The data collected was used solely for statistical purposes for the study and included only the age and gender of students. It was used by the evaluation team and MMACA to prepare the report, with a data-sharing agreement signed.

Now that the data has been analysed, it will be held by the evaluation team for a maximum of two years, solely to conduct secondary analyses. It will be securely stored on the servers of the University of Oviedo.

Project Teams

Evaluation Team

University of Oviedo: Luis J. Rodríguez Muñiz (principal investigator and coordinator of the evaluation team), Juan José Santaengracia, Carmela Suárez González, Irene Díaz, Celestino Rodríguez Pérez, Trinidad García, Taina Iglesias Cabo.

Complutense University of Madrid: Belén Palop

MERG Research Group, **UNIMODE Research Group** and **ADIR Research Group**

Intervention Team

MMACA – Museum of Mathematics of Catalonia

Sergio Belmonte – Coordinator of the training team, design of the intervention and trainer

Eulàlia Tramuns – Design of the intervention and trainer

Commission & Organisation Team

EduCaixa – "la Caixa" Foundation

Marta García Matos – Conceptualisation and coordination

Anton Aubanell – Conceptualisation

Marià Cano Santos – Support in conceptualisation and coordination

Ana Municio Zúñiga – Support in coordination and contact with participating educational centres

Methods

Trial Design

The impact of ED on students' CT skills was evaluated through a two-arm, clustered RCT with randomisation occurring at the school level. The 83 recruited schools were randomised to either the intervention (42 schools) or control (41 schools) group in June 2023.

To be included in the randomisation process, recruited schools were required to sign a Memorandum of Understanding (MoU) by which they committed to provide the data necessary for the study, apply the pretest and posttest, have the teachers attend the training and implement the intervention programme in all first-year secondary classrooms at their schools (intervention group only).

Once the participant schools were determined, the selection of the two arms of the study was performed by simple randomisation with stratification. The variables considered for stratification included the geographic area in terms of the population size of the village/town/city in which the school was located (< 50 000, 50 000-100 000, 100 000-250 000, > 250 000 inhabitants) and the region (Catalonia, Aragon, Andalusia), the socioeconomic status of the families (low, medium-low, medium, medium-high, high), the type of school (public, state-funded private or private) and the school's previous syllabus in computer science topics (yes/no).

Following randomisation, the teachers nominated for each intervention school participated in ED training. They then taught ED curriculum materials to more than 5100 students from October 2023. The intervention ended in February 2024.

The main aspects of the trial design are shown in the table below.

Table 3: Trial design

Trial design, including number of arms		Two-arm cluster randomised controlled efficacy trial
Unit of randomisation		School
Stratification variable (s) (if applicable)		Geographic area, socioeconomic index, school syllabus in computer science topics, type of school (public, state-funded private or private).
Primary outcome	Variable	Computational thinking skills
	Measure (instrument, scale, source)	Bebras-Based Assessment for Computational Thinking (BBACT) test score
Secondary outcome(s)	Variable(s)	Attitudes towards mathematics: lack of confidence in the future, perceived competence, perceived utility, intrinsic motivation, achievement motivation, lack of interest in mathematics, anxiety and feelings
	Measure (instrument, scale, source)	Subset of the Attitudes Toward Mathematics Inventory in its Spanish version (IAM) test score
Baseline for primary outcome	Variable	Computational thinking skills
	Measure (instrument, scale, source)	BBACT test score

Baseline for secondary outcome(s)	Variable	Attitudes towards mathematics: lack of confidence in the future, perceived competence, perceived utility, intrinsic motivation, achievement motivation, lack of interest in mathematics, anxiety and feelings
	Measure (instrument, scale, source)	IAM test score

The control condition in this efficacy trial was business as usual. Control schools that participated in the data collection activities were offered the same benefits as the intervention schools: including the opportunity to attend the same ED training as the intervention schools (at the end of the study) and an advance payment of £421 (€500) to cover any expenses related to the pre- and posttest processes. More details about the monitoring of control schools can be found under Early Intervention in the Implementation and Process Evaluation Results section.

Participant Selection

The sample consisted of secondary schools in Catalonia, Aragon and Andalusia, from a variety of contexts (urban/rural, public/private, etc.). Schools in Catalonia and Aragon were recruited through their respective regional Departments of Education, whereas in Andalucia the close collaboration with our partners from HelloMath and their significant connections with the Andalusian government enabled us to achieve effective diffusion in the region. Approximately 5000 schools were contacted (this is an estimation since the initial call was posted by the Departments of Education on their respective websites or published in their newsletters). In addition, there were schools from La Rioja, Castilla-La Mancha and Extremadura that asked to join. Given their proximity, those from La Rioja were included with the schools in Aragon and those from Castilla-La Mancha and Extremadura with the schools in Andalusia. There were 108 schools that expressed their intention to participate, 84 of them signed the MoU and one of them withdrew before the intervention. The recruitment was carried out by the commission and organisation team, without the participation of the evaluation team. The table below shows the final distribution of the sample schools.

Table 4: Sample distribution by region and trial arm

	Andalusia	Catalonia	Aragon	Total
Control	18	19	4	41
Intervention	15	23	4	42
Total	33	42	8	83

After being assigned to the intervention group, the schools sent their 1st-year secondary school maths teachers to attend the ED training. After training, each school had to submit a 17-week lesson plan describing the intervention. Teachers were provided with a template and instructions for the essential contents of the lesson plan, which were evaluated by the intervention team. The intervention could only begin once feedback had been received on the lesson plan.

Outcome Measures

Primary Outcome

The primary outcome measure was computational thinking skills, in particular, decomposition, patterns and generalisation, and debugging, assessed using an ad hoc instrument designed for this study, the BBACT test (see Appendix C). The BBCAT test was specifically designed for this research project to minimise the

influence of prior training in programming-related skills, such as computational thinking, computer science, robotics, or coding fields, which are offered by many Spanish schools as extracurricular activities. The test assesses core aspects of computational thinking, including decomposition, pattern recognition, abstraction, algorithms (with debugging) and modelling. It consists of 17 items derived from Bebras tasks (Dagiene & Dolgopolovas, 2022), focusing on unplugged computational thinking skills that do not require familiarity with any programming language, whether formal or visual. Each task was adapted to a multiple-choice format with four options, only one of which was correct. Scores were calculated as a percentage of correct answers.

In the protocol and SAP, the instrument was called the Mathematical Dimensions of Computational Thinking test (MCDT). We later reconsidered the name because it overfitted in terms of referring to mathematical dimensions. Considering that the instrument was primarily designed using Bebras items, we renamed it the Bebras-Based Assessment for Computational Thinking (BBACT) test (see Appendix B). The need to define an instrument stemmed from the goal of not focusing CT on programming skills, given that available instruments mainly focus on programming and debugging. While programming is a core skill in CT, it is not the only one. Therefore, in line with the Bebras initiative, we aimed to focus on problem decomposition, modelling and other relevant aspects.

The main features of the BBACT instrument are described in Santaengracia et al. (2024). It was piloted in a Spanish and Uruguayan sample of 302 students of the same age as those in the ED intervention. BBACT emerged as an effective assessment tool for evaluating CT in 1st-year secondary school students, which is aligned with Lockwood and Mooney's (2018) instrument targeting older age groups (15 to 17 years). This coincidence emphasises the importance of incorporating diverse Bebras tasks in terms of topics and CT components to ensure the instrument's utility. Contrary to Wiebe et al. (2019), developed for 1st-year students and incorporating Bebras tasks but focused on algorithms, BBACT includes a comprehensive range of CT components: decomposition, pattern recognition, abstraction, modelling, algorithms and debugging. By doing so, it assesses a wider range of CT components and, in addition, provides a more comprehensive measure of students' skills.

Table 5: Specification matrix of the BBACT instrument

Item number	CT Component	Expected difficulty	Difficulty
1	Decomposition	Average	Average
2	Decomposition	High	High
3	Decomposition/patterns	Low	Average
4	Decomposition	High	Average
5	Patterns	Low	Low
6	Patterns	Low	Low
7	Patterns	Average	Average
8	Patterns	High	Average
9	Algorithms	Low	Low
10	Debugging/algorithms	Average	Average
11	Debugging/algorithms	High	Average
12	Abstraction	High	Average
13	Abstraction	Average	Average
14	Abstraction/patterns	Low	Average
15	Abstraction/patterns	High	High

16	Modelling	Low	Average
17	Modelling	Average	High

BBACT incorporates a set of questions of varying difficulty, which resulted in significant differences in correct and incorrect responses between items. The BBACT raw score is based on the number of correct responses. In the pilot study, the average score was 8.95 and the standard deviation 3.79. According to the pilot study, three items are considered difficult, eleven items are medium difficulty and three easy items (see Table 5). Accuracy rates for the difficult items were below 30.37 %, between 30.37 % and 75.01 % for the medium-difficulty items and over 75.01 % for the easy items (two of them very easy, with accuracy rates above 90 %). The average time for completing the assessment in the pilot study was between 23 and 28 minutes.

Although some previous instruments on CT reported gender differences favouring boys, particularly in higher student ages (Román-González et al., 2017), BBACT did not report significant differences in gender during the pilot study (this is also confirmed in this project, as reported in the Results section). This finding supports the hypothesis that the use of Bebras tasks instead of Scratch-based tasks could explain this lack of gender difference (see, for example, the large-scale evaluation by Izu et al., 2017).

Regarding the content validity of BBACT, its basis in Bebras reinforces its validity, as interest among psychometricians in using Bebras tasks as an evaluation tool has increased in recent years. Lockwood and Mooney (2018) concluded that Bebras items can be a good way of assessing CT, but, for the instrument to be useful, questions that are varied in terms of themes and CT components need to be included. Wiebe et al. (2019) combined the Computational Thinking Test (CTt) by Román-González et al. (2017) with Bebras tasks, resulting in an instrument with 25 questions (19 from CTt and 6 from Bebras) for adolescents aged 11 to 13. This approach enriched the algorithmic focus, while the instrument was still focused on it. However, the dimensional structure accepted by the scientific community for describing Bebras items is still under debate, as these appear not to align with standard psychometric constructs. For instance, Palts et al. (2017) found only two factors, instead of the five Bebras components present in the considered items.

The BBACT instrument was developed after discussions with a group of experts in different research fields of mathematics and mathematics education, computer science and computer science education, which helped in determining the final version of the BBACT test and led to the name change.

As for the construct validity of BBACT, the Exploratory Factor Analysis (EFA) revealed a multi-factor structure (between 5 and 7 factors, depending on the measures considered) with low performance in terms of explained variance (30.716 % of the total). Nevertheless, commonalities were predominantly low (< 0.30), suggesting a poor representation of variables by the factors, contrary to the recommended loadings, which were greater than 0.40 (Glutting, 2002). The presence of both high and low-difficulty questions significantly complicated the factorial structure adjustment. Adjustment measures (very low KMO index 0.03, but Bartlett's Test of Sphericity significant $p < .001$) suggested that the data presented a low suitability for EFA. This is usual in school tests, which often have situations of ambiguous dimensionality (Fernández-Alonso, 2005). In other words, when the specifications are broad (as they are in the BBACT) they rarely fit a single dimension. However, for practical purposes, most (if not all) educational assessment studies (including the most robust ones: PISA, TIMSS, PIRLS, etc.) provide a total score. The difficulty of obtaining a clear factorial structure together with a minimum level of content validity in a test that combines broad specifications has led researchers to determine intermediate degrees for balancing structure and content validity. A classical study by Hattie (1985) provided different ways to measure one-dimensionality as a degree rather than a threshold. The pilot BBACT study provided a moderate internal consistency (Cronbach's alpha value 0.59), which could be acceptable given the number of items and the different components that are analysed (Taber, 2018). Additionally, Reckase (1979) concluded that when a test has

a dominant factor – explaining at least 10 % of the total variance – the skill estimates for individuals are highly correlated with the scores on this factor. Both conditions hold in the case of BBACT, given that the first factor is over 10 % of explained variance and the correlation between direct scores and factor load for the items is very high (0.93). Therefore, we have sufficient evidence, based on the mentioned criteria, to support that BBACT may be measuring a single dimension, even if this dimensionality is somewhat ambiguous in factorial terms – a situation that is common in school-based tests.

The BBACT test, both in pre- and posttests, was administered online by the University of Oviedo for security and data protection aspects. Teachers did not have an active role in the administration, they only had to provide the students with the corresponding links. They also received instructions about not guiding or helping the students with the answers. The pretest was only administered on paper at one school due to power supply issues in the building. Teachers did not keep copies of the pretest and were not aware that some items would be the same in the posttest. We chose to keep the same items in the pre- and posttests due to the nature of the items, which were not directly linked to the syllabus of the mathematics course and given that the time lapse between pre- and post- was large enough to avoid any memory effect.

Students were assigned a code by their teachers. This code was blind to the evaluation team so that only the teachers were able to match a student with a code. Students used this code when completing the online tests.

Secondary Outcome

The secondary outcome measure was attitudes towards mathematics. This outcome was examined using a subset of the Attitudes Toward Mathematics Inventory in the Spanish version, named the IAM test (see Appendix D), which was analysed in Fernández et al. (2015) and García et al. (2016). It constitutes a widely accepted scale in the Spanish context. The selected subset comprises 32 items on a 5-point Likert scale (from *strongly disagree* to *strongly agree*), covering the following dimensions: anxiety, feelings, perceived utility, perceived competence, intrinsic motivation, parents' and teachers' attitudes, extrinsic motivation and motivation towards social achievement. The names initially assigned to the selected dimensions in the SAP were reconsidered to reflect the items that were ultimately introduced (for example, having explicit items asking about parents' and teachers' attitudes was considered useful). The IAM score was obtained by adding each item score, after reversing the items which are negatively expressed (15 out of 32).

The administration of the IAM followed the same procedure as the BBACT: online (except in one pretest case), anonymously, using the same set of items in the pre- and posttest. Students were assigned a code by their teachers. This code was blind to the evaluation team so that only the teachers could match a student with a code. Students used this code when completing the online tests.

BBACT and IAM were administered consecutively on Microsoft Forms.

After piloting the selected items from IAM and in order to shorten the length of the test, we reduced the number of items and modified some of the dimensions. The final list of dimensions was as follows: lack of confidence in the future, perceived competence, perceived utility, intrinsic motivation, achievement motivation, lack of interest in mathematics, anxiety and feelings.

Sample Size

The power analysis gave estimated minimum detectable effect sizes (MDES) for the primary outcome. The sample of 83 schools provided a potential total sample size of 5262 students. Two schools dropped out (34 students) and there were 52 cases of wrong or duplicated student codes, which were not considered. Therefore, the analysis sample consisted of 81 schools (41 in the control arm and 40 in the intervention arm) that gave a total of 5176 students (2661 in the control group, 2515 in the intervention group). More details are provided in Table 8.

Calculations of MDES and correlations were obtained for the sample size using the PowerUpR package. Due to the lack of scientific literature on this specific area of study, we applied the maximum uncertainty principle, setting correlations at 0.5. Alpha values were set at 5 %, a common threshold in statistical studies, with statistical power set at 0.80.

Information about the missing data is provided in the corresponding section.

Randomisation

As previously explained, once the participant schools were determined, the selection of the two arms of the study was performed by simple blinded randomisation with stratification. The variables considered for stratification included the geographic area in terms of the population size of the village/town/city in which the school was located (< 50 000, 50 000-100 000, 100 000-250 000, > 250 000 inhabitants) and the region (Catalonia, Aragon, Andalusia), the socioeconomic status of the families (low, medium-low, medium, medium-high, high), the type of school (public, state-funded private or private) and the school's previous syllabus in computer science topics (yes/no). The randomisation process was carried out on 6 June 2023 using an Excel spreadsheet and the random number generator provided by the software, controlling the balance between both groups in all the strata. The schools were contacted immediately after the randomisation had been completed to inform them of their allocation.

Statistical Analysis

Primary Analysis

The same items of the BBACT test were included in the pretest and the posttest to ensure consistency. The pretest consisted of 17 single-choice items, each providing a stimulus and then offering four possible answers. The instrument was administered as an online version (MS Forms).

A multilevel approach was adopted considering pupils grouped into clusters by school (two-level multilevel model). Linear mixed regression models were constructed for the primary and secondary outcomes, using pretest and group and all other covariates as fixed effects. School was considered random. The statistical analysis was performed using R software, version 4.4.1. (Bates et al., 2015; R Corte Team, 2024). An intention-to-treat analysis was considered, to ensure that all students were analysed in the group they were originally assigned to, regardless of their level of participation.

Model equation:

$$Y_{ij} = \beta_0 + \beta_1 BBACT_{ij} + \beta_2 IG_j + \beta_3 IA_j + \beta_4 IB_i + \beta_5 IC_i + \beta_6 ID_i + \mu_i + \varepsilon_{ij}$$

Being:

Y_{ij} : BBACT posttest for j^{th} student in i^{th} cluster (school)
$BBACT_{ij}$: pretest for j^{th} student in i^{th} cluster (school)
IG_j : indicator variable for intervention/control group of j^{th} student
IA_j : indicator variable for gender of j^{th} student
IB_i : indicator variable for previous training in i^{th} cluster (school)
IC_i : indicator variable for population size in i^{th} cluster (school)
ID_i : indicator variable for socioeconomic status in i^{th} cluster (school)
μ_i : random effect in i^{th} cluster (school)
ε_{ij} : residual term for j^{th} student in i^{th} cluster (school)

Secondary Analysis

In terms of the secondary outcome, an intention-to-treat analysis was also considered, so that all students were analysed in the group they were originally assigned to, regardless of their level of participation.

Model equation:

$$Z_{ij} = \beta_0 + \beta_1 IAM_{ij} + \beta_2 IG_j + \beta_3 IA_j + \beta_4 IB_i + \beta_5 IC_i + \beta_6 ID_i + \mu_i + \varepsilon_{ij}$$

Being:

Z_{ij} : IAM posttest for j^{th} student in i^{th} cluster (school)
IAM_{ij} : pretest for j^{th} student in i^{th} cluster (school)
IG_j : indicator variable for intervention/control group of j^{th} student
IA_j : indicator variable for gender of j^{th} student
IB_i : indicator variable for previous training in i^{th} cluster (school)
IC_i : indicator variable for population size in i^{th} cluster (school)
ID_i : indicator variable for socioeconomic status in i^{th} cluster (school)
μ_i : random effect in i^{th} cluster (school)
ε_{ij} : residual term for j^{th} student in i^{th} cluster (school)

Analysis in the Presence of Non-Compliance

Compliance was defined at the school level. Based on the theory of change, teachers' attendance to training sessions and adherence to their individual work plans were identified as the key elements of the intervention, assigning each a weight of 35 % in the final compliance index. Three other aspects were evaluated during the intervention, each allocated a weight of 10 %: level of teachers' involvement, proportion of students actively involved in the programme and use of materials provided.

Therefore, the factors considered in the compliance indicator were:

a. Average percentage of attendance of teachers to the training sessions (obtained from school records).

b. Average score for the work plan developed by the teachers (obtained from the score awarded by the intervention team, on a 0-100 scale).

c. Average score of teachers' commitment to the intervention plan. Commitment was analysed through follow-ups in a weekly newsletter and an online meeting, with records kept of both communication channels, including 1530 email exchanges throughout the intervention,[1] assigning a value on a 0-100 scale.

d. Average percentage of students at the school who actively participated in the intervention, excluding students who sabotaged the activity (obtained through a questionnaire administered to teachers after the intervention).
e. Average percentage of use of materials developed by the school's teachers (obtained through a questionnaire administered to teachers after the intervention).

The total score consisted of an index (J) constructed at the school level, as follows:

[1] The way to measure the degree of commitment changed from the initial plan. Originally, commitment was to be measured through a questionnaire administered during the intervention, on a 0-100 scale. Later, it was decided that there was sufficient information retrieved from the process to produce a score for commitment.

$$J = 0.35 \cdot (a + b) + 0.1 \cdot (c + d + e)$$

J ranges between 0 and 100 and provides a measurement of the overall compliance. If a school failed to achieve a score of 80 for J, it was analysed individually to identify the circumstances that led to this outcome, potentially resulting in its exclusion from the sample if deemed necessary and comparing it with schools above the threshold.

Missing Data Analysis

A multilevel mixed-effect logistic regression model was run to assess any statistically significant predictors of missing primary outcome data (where 1 = missing; 0 = complete), including all available student and school-level baseline data, with group as fixed effects and school as a random effect. Missing values were identified in each variable and the pattern and type of missingness were studied. A missing rate of 5 % or less would not typically bias the primary impact estimates, regardless of the pattern of missingness (Schafer, 1999) and so, a complete-case analysis was employed. When missing data resulted in an exclusion of 5 % of data or more, a multiple imputation technique was used and a sensitivity analysis was performed.

Sub-Group Analyses

No sub-group analysis was carried out.

Estimation of Effect Sizes

Effect sizes for both outcomes were calculated by dividing the adjusted mean difference between the intervention and control groups by the pooled standard deviation obtained from the unconditional model. A 95 % confidence interval for the effect size was calculated by dividing the 95 % confidence limits for the adjusted mean difference by this same variance. Thus, for the primary outcome:

$$ES = \frac{\bar{Y}_I - \bar{Y}_C}{\sqrt{s^*}}$$

Where $\bar{Y}_I - \bar{Y}_C$ denoted the mean difference between the trial arms obtained from the model and *s** denoted the pooled unconditional variance from the unconditional model. The same method was used for the secondary outcome (*Z* instead of *Y* in the formula above).

Estimation of ICC

The intra-cluster correlation coefficient (ICC) associated with the school for the pretest and posttest outcomes was provided with 95 % confidence intervals (CI) and determined using the ICC function of the R performance library. The ICC represents the proportion of variance in each outcome that can be explained by the variation between clusters (i.e., schools):

$$ICC = \frac{\text{Random Intercept Variance}}{\text{Total Variance}} = \frac{\text{Random Intercept Variance}}{\text{Random Intercept Variance} + \text{Residual Variance}}$$

The school-level intra-cluster correlation coefficient for the posttest primary outcome was extracted from each mixed model, with a 95 % CI. The ICC associated with the school for the pretest primary scores were presented with a 95 % CI.

Implementation and Process Evaluation

Research Methods

The analysis of the implementation and process evaluation combined qualitative and quantitative research methods. Evidence collection comprised a set of instruments (see Table 6) including surveys, documents, classroom observations, interviews and assignments. Different data analysis methods were carried out according to each instrument: document analysis, descriptive statistics, inductive and deductive coding, observation record, etc. Census analyses (that is, considering all the participants in the intervention) were used in all instruments except for observation phases. For these phases, 34 of the 40 intervention schools were visited. Initially, the evaluation team tried to visit all the schools, therefore, no sampling was considered. There are various reasons as to why six schools could not be visited. It was logistically very difficult (almost impossible) to visit three of the schools with two located in rural areas of the region that is Andalusia and one, in Catalonia, which was very far away from the rest. In the other three cases, the schedules proposed by the schools were incompatible with the researchers' agendas. The census analysis for the observation phases can, therefore, be considered as an intentional (by convenience) non-probabilistic sampling.

For the implementation and process evaluation, the following research questions, classified by the moment they refer to, were posed:

Pre-intervention:

RQ1) Have teachers actively participated and engaged in the training?
RQ2) Did the training generate exchanges of points of view and discussion among the teachers?
RQ3) Have families been informed? What was their reaction?
RQ4) Have teachers mastered the teaching resources based on ED?
RQ5) Have teachers created teaching materials based on the resource?
RQ6) How different is Exploding Dots from 'business as usual'?

Early intervention:

RQ7) Did the intervention team solve the doubts that arose during the process?
RQ8) Has the intervention team been receptive to the contextual situations of the schools?
RQ9) Has the implementation been linked to the contents and competencies of the mathematics course?
RQ10) Has contact with the team been fluid and dynamic during the implementation?
RQ11) Have the control groups changed their methods in relation to maths and computational thinking?
RQ12) Did the control groups implement the pretest?

During the intervention:

RQ13) Have students been actively involved and motivated during the implementation?
RQ14) Have students positively assessed the experience?
RQ15) Has the implementation gone according to the schedule?
RQ16) Have the designed materials been used in the implementation?
RQ17) Have teachers solved students' doubts during the implementation?

Post-intervention:

RQ18) What is the teacher's perception of the effectiveness of the programme?
RQ19) Did the teachers identify any unexpected issues during the implementation of the programme? What did schools need to have in place for the programme to go as planned and what were the challenges they faced?
RQ20) What is the teacher's perception of the support received from the centre?
RQ21) Unintended consequences – is there any evidence of a displacement of provision?

Table 6. Overview of the implementation and process evaluation methods

Stage	Research method	Data collection methods	Participants /data sources	Data analysis methods	Who collected the data?	Research questions addressed	Implementation/ logic model relevance
Pre-intervention	Quantitative	Questionnaire/ Survey (MMACA's assessment)	All teachers	Descriptive statistics	Intervention team	1, 2	Responsiveness
	Quantitative	Questionnaire	All principals	Descriptive statistics	Intervention team	3	Reach
	Qualitative	Document revision (MMACA's assessment)	All teachers	Document analysis	Intervention team	5	Fidelity, Adaptation
	Quantitative	Questionnaire/ Survey (MMACA's assessment)	All teachers	Descriptive statistics	Intervention team	4	Quality
	Quantitative	Questionnaire	All teachers	Descriptive statistics	Intervention team	6	Programme differentiation
Early intervention	Qualitative	Document revision	All teachers		Intervention team	7, 8, 9, 10	Fidelity
	Qualitative	Document revision	All teachers		Intervention team	11, 12	Monitoring
On-going intervention	Qualitative	Interview/ observation	Sample teachers	Content analysis	Evaluation team	15, 16, 17	Responsiveness
	Qualitative	Document revision	Sample teachers	Document analysis	Evaluation team	15	Dosage, adaptation
	Qualitative	Interview/ observation	Sample students	Inductive & deductive coding / Observation record	Evaluation team	13, 14, 16	Fidelity, Adaptation
	Qualitative	Interview/ observation	Sample students	Observation record	Evaluation team	13, 14, 17	Quality
Post-intervention	Quantitative	Questionnaire	All teachers	Descriptive statistics	Evaluation team	18, 19, 20	Programme differentiation, Quality
	Quantitative /qualitative	Survey/ observation	Sample students	Inductive coding	Evaluation team	13, 14	Responsiveness

Analysis

Before the classroom intervention, teachers participated in training sessions delivered by ED programme trainers although the rest of the intervention team also attended the training sessions. Families were informed about the project via a letter from each school's director. Each school submitted its intervention plan, based on the project materials, to be assessed by the intervention team, who reviewed it and gave feedback. A Google Classroom site was established for each of the three regions, where the community of teachers and trainers centralised fluid and constant communication about the intervention.

During the intervention, observations were carried out, together with semi-structured interviews with the participating students and teachers. For the observation phases, a rubric was used to check the adherence to the programme as well as the classroom climate. For the interviews, inductive coding was applied. The Google Classroom sites were maintained for the duration of the intervention.

Additionally, an online meeting was held with the intervention team on 18 December 2023, to check that everyone was prepared and to control any issues before the posttest. All schools were represented at the meeting. A weekly newsletter was delivered to the schools on Thursdays, from 5 September 2023 to 27 September 2024. No perception of a lack of support from the school or of a displacement of provision was reported.

After the intervention, questionnaires were administered to students and teachers. These combined Likert-scale items with open-ended questions and were analysed using quantitative and qualitative techniques. Quantitative techniques consisted of obtaining descriptive statistical measures of the questionnaires' information. Qualitative analysis was conducted using a combination of inductive and deductive coding. Inductive coding allowed themes and categories to emerge from the data, in particular, the analysis focused on conceptions about mathematics and its teaching and learning, and features and issues regarding the intervention. Deductive coding involved applying existing theoretical frameworks, particularly about factors influencing students' attitudes towards mathematics, to interpret the data in a structured way.

Costs

There were three levels of technology used in the programme's implementation: computers with internet access, computers without internet, and no computers at all. Therefore, costs for technology depended solely on the technological resources already available at each school.

The primary cost of the project is associated with teacher training and includes the number of hours of in-person and online training, the number of participating trainers and trainees, and the costs per hour and person. As previously mentioned, the teacher training cost of £168 (€200) per teacher, giving an estimated cost per student of £4.20 (€5).

Timeline

Below is the timeline of the intervention.

Table 7: Timeline

Dates	Activity	Staff responsible/leading
6 June 2023	Randomisation	Evaluation team
6 and 17 June 2023 2 September 2023	Teacher training	Intervention team

Dates	Activity	Staff responsible/leading
6 June 2023	Randomisation	Evaluation team
25-29 September 2023	Pretest	Evaluation team Commission and organisation team
1 October 2023 to 15 February 2024	Intervention	Intervention team
6-20 November 2023	Visits to schools (1st round)	Evaluation team
10-20 January 2024	Visits to schools (2nd round)	Evaluation team
19-23 February 2024 26 February to 1 March 2024	Posttest	Evaluation team Commission and organisation team

Impact Evaluation Results

The figure below shows the participant flow including losses and exclusions.

Figure 2: Participant flow diagram (two arms)

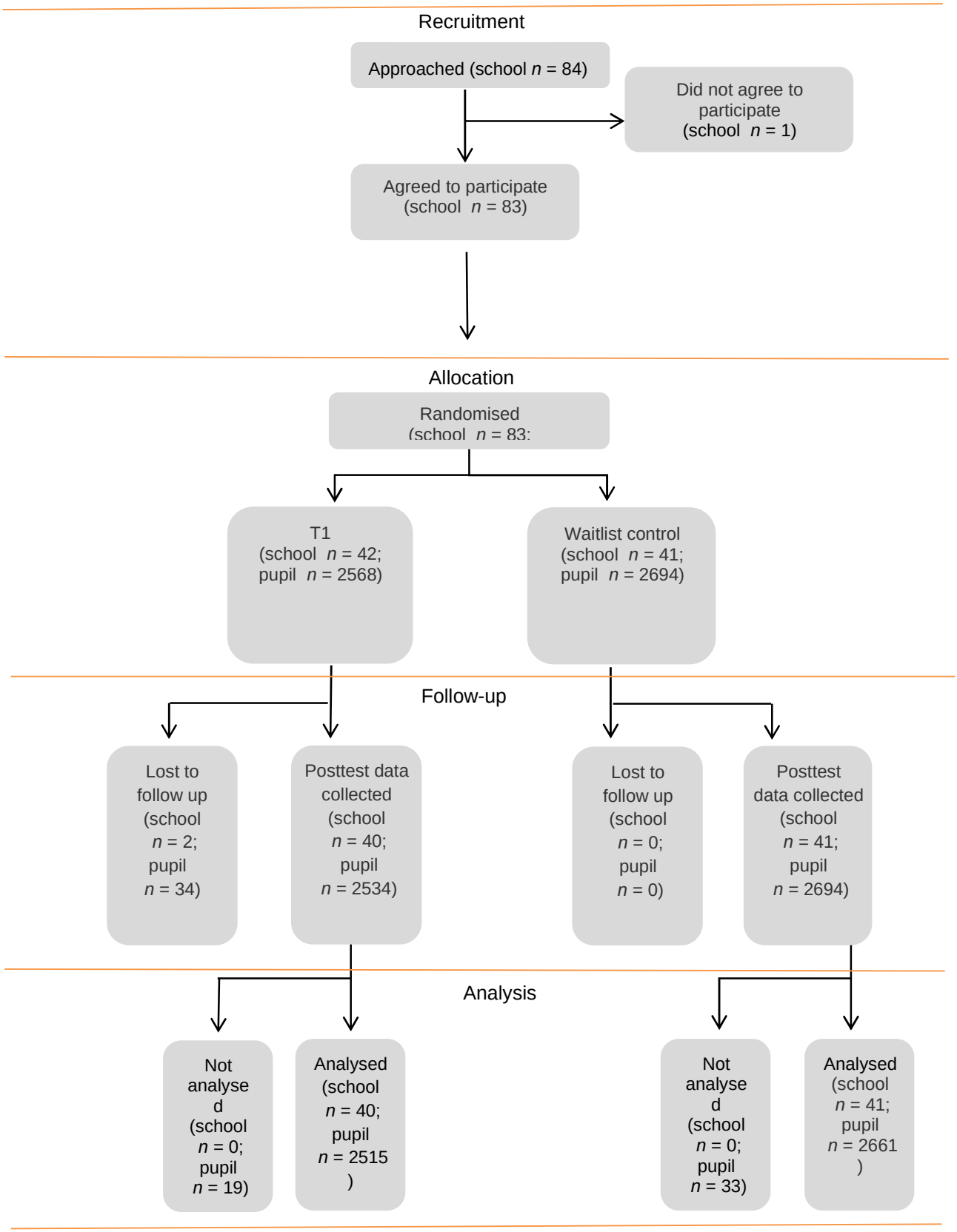


Table 8: Minimum detectable effect size at different stages

		Overall sample		
		Protocol	Randomisation	Analysis
MDES		0.151	0.11	0.192 (BBACT) 0.154 (IAM)
Pretest/posttest correlations	Level 1 (pupil)	0.5	0.5	0.376 (BBACT) 0.504 (IAM)
	Level 2 (class)			
	Level 3 (school)	0.5	0.5	0.087 (BBACT) 0.026 (IAM)
Intracluster correlations (ICCs)	Level 2 (class)			
	Level 3 (school)	0.05	0.05	0.031 (BBACT) 0.026 (IAM)
Alpha[2]		0.05	0.05	0.05
Power		0.8	0.8	0.8
One-sided or two-sided?		2	2	2
Average cluster size		15	60	64
Number of schools[3]	Intervention	40	43	40
	Control	40	41	41
	Total:	80	83	81
Number of pupils[4]	Intervention	1200	2568	2661
	Control	1200	2694	2515
	Total:	2400	5262	5176

Note: There was a mistake in the version of Table 8 published in the SAP, which does not affect the main measures. The correct total number of students and schools are those in this document.

Attrition

Table 9 shows the main data about attrition. Two schools in the control group dropped out of the project (they participated in the pretest and then abandoned the programme), there were 34 students in those two schools. Additionally (as reported in Figure 2), 19 students in the intervention group and 23 in the control group were excluded from the analysis because the codes were incomplete or wrong, in both the pretest and posttest.

Table 9: Student-level attrition from the trial (primary outcome)

		Intervention	Control	Total
Number of pupils	Randomised	2568	2964	5262
	Analysed	2515	2661	5176
Pupil attrition (from randomisation to analysis)	Number	53	33	86
	Percentage	2.06 %	1.22 %	1.63 %

Pupil and School Characteristics

Pretest Results

Table 10 shows the main descriptive statistics of the pretest results. When considering the whole sample for the primary outcome, the distribution of the BBACT score, which ranges from 0 to 17, is quite symmetric, indicating that the difficulty of the instrument was well-calibrated. Items on the test were scored as 1 if the answer was correct and 0 if the answer was incorrect. The total score was the sum of all the results. The midpoint of the score scale is 7.5 and, as shown, both the mean and the median are slightly higher than this. The maximum and minimum are also shown.

The IAM score, used for the secondary outcome, ranges from 0 to 128. Each item was scored from 0 to 4 depending on the degree of agreement with the statement and the total score was obtained by adding all the items. Table 10 shows that the distribution of the whole sample is slightly skewed to the left, although still quite symmetric.

Regarding the balance between the control and intervention groups at baseline, statistical analyses were performed for both primary and secondary outcomes. No statistically significant differences were detected between the two groups. In both cases, given the large sample size, score distribution and p-value of the F-test for equality of variances (p-value = 0.864 for BBACT, p-value = 0.215 for IAM), the Student's t-tests provided large p-values (p-value = 0.981 for BBACT, p-value = 0.215 for IAM), meaning that the control and intervention group averages can be considered equal. Histograms of scores are displayed in Figures 3 and 4, together with the posttest distributions.

Table 10: Main descriptive statistics of the pretest results

Sample	Outcomes	n	Average	SD	Min	P25	Median	P75	Max
Total	BBACT score (0-17)	5176	8.85	2.73	1	7	9	11	17
Intervention	BBACT score (0-17)	2515	8.86	2.76	1	7	9	11	16

Control	BBACT score (0-17)	2661	8.85	2.70	1	7	9	11	17
Total	IAM score (0-128)	5176	85.10	16.15	21	74	86	97	124
Intervention	IAM score (0-128)	2515	85.39	16.12	21	75	86	97	124
Control	IAM score (0-128)	2661	84.83	16.18	22	74	86	97	123

In terms of the distribution of the other analysed variables, as indicated in Table 11, the sample was quite balanced with respect to the baseline measures. Regarding the school type, the population data of students enrolled in compulsory secondary school education (ESO, in Spanish), which includes those aged 12 to 16, from the Spanish Ministry of Education. The same data source was used for obtaining information on gender at baseline (official statistics do not report non-binary or gender-neutral students), which was also very balanced in both arms.

In terms of population size, the data source used was the Spanish census, although the information published is aggregated for ≤ 50 000, 50 000-100 000, 100 000-200 000 and > 200 000 inhabitants. Using the microdata, we determined the baseline for our intervals by taking the population living in cities/towns with > 250 000 inhabitants and subtracting it from the population living in cities/towns with > 200 000 inhabitants. There is no school population data linked to population size, therefore there are differences between the baseline and the percentages, particularly between small and large cities/towns. This is because most of the young population is concentrated in large cities, while the majority of the population in smaller towns is older. Therefore, the sample is fairly balanced in terms of the student population, also in these two categories.

There is no baseline measure for prior training in CT, only indirect measures on which regions have a course related to CT on their curriculum, but not about the number of students receiving such training.

Lastly, in terms of socioeconomic status, no public data is released at a national level. We acknowledge that this scale is based on perceptions, as it was assigned by the directors of the participating schools. Therefore, apart from pointing out that both extremes are not represented in the intervention group, there is no baseline measure for comparison.

Differences in some values between the intervention and control groups are explained by the fact that two schools dropped out, also because the randomisation was done at the school level and because the schools had different numbers of students, classes and class ratios. The randomisation was therefore conducted in an attempt to balance the number of students. Nevertheless, the differences are, in general, minor.

Table 11: Baseline characteristics of groups as randomised

		Intervention group		Control group	
Pupil-level (categorical)	National-level mean	n/N (missing)	Count (%)	n/N (missing)	Count (%)
Public schools [1]	67 %	1610/2515 (0)	64 %	1572/2661 (0)	59.1 %
State-funded private schools [1]	29 %	905/2515 (0)	36 %	999/2661 (0)	37.5 %
Private schools [1]	4 %	0/2515 (0)	0 %	90/2661 (0)	3.4 %

Low socioeconomic level	No data	381/2515 (0)	15.1 %	93/2661 (0)	3.5 %
Medium-low socioeconomic level	No data	1204/2515 (0)	47.9 %	1530/2661 (0)	57.5 %
Medium-high socioeconomic level	No data	930/2515 (0)	37 %	948/2661 (0)	35.6 %
High socioeconomic level	No data	0/2515 (0)	0 %	90/2661 (0)	3.4 %
Population ≤ 50 000 [2]	47.10 %	759/2515 (0)	30.2 %	1098/2661 (0)	41.3 %
Population (50 000-100 000) [2]	13.04 %	366/2515 (0)	14.6 %	246/2661 (0)	9.2 %
Population (100 000-250 000) [2]	15.07 %	281/2515 (0)	11.2 %	309/2661 (0)	11.6 %
Population > 250 000 [2]	24.79 %	1109/2515 (0)	44.1 %	1008/2661 (0)	37.9 %
No previous experience in CT	No data	1971/2515 (0)	78.4 %	1925/2661 (0)	72.3 %
Previous experience in CT	No data	544/2515 (0)	21.6 %	736/2661 (0)	27.7 %
Gender (Male) [1]	51.5 %	1231/2515 (33)	49.6 %	1230/2661 (22)	46.6 %
Gender (Female) [1]	48.5 %	1123/2515 (33)	45.2 %	1282/2661 (22)	48.6 %
Gender (Other/Non-binary)	No data	128/2515 (33)	5.1 %	127/2661 (22)	4.8 %

[1]Source: "Datos y cifras. Curso Escolar 2024/2025" (**https://www.educacionfpydeportes.gob.es/servicios-al-ciudadano/estadisticas/indicadores/datos-cifras.html**)
[2]Source: Estimated from "Encuesta continua de población en España" (**https://www.ine.es/dyngs/Prensa/es/ECP2T24.htm**)

Outcomes and Analysis

Primary Analysis

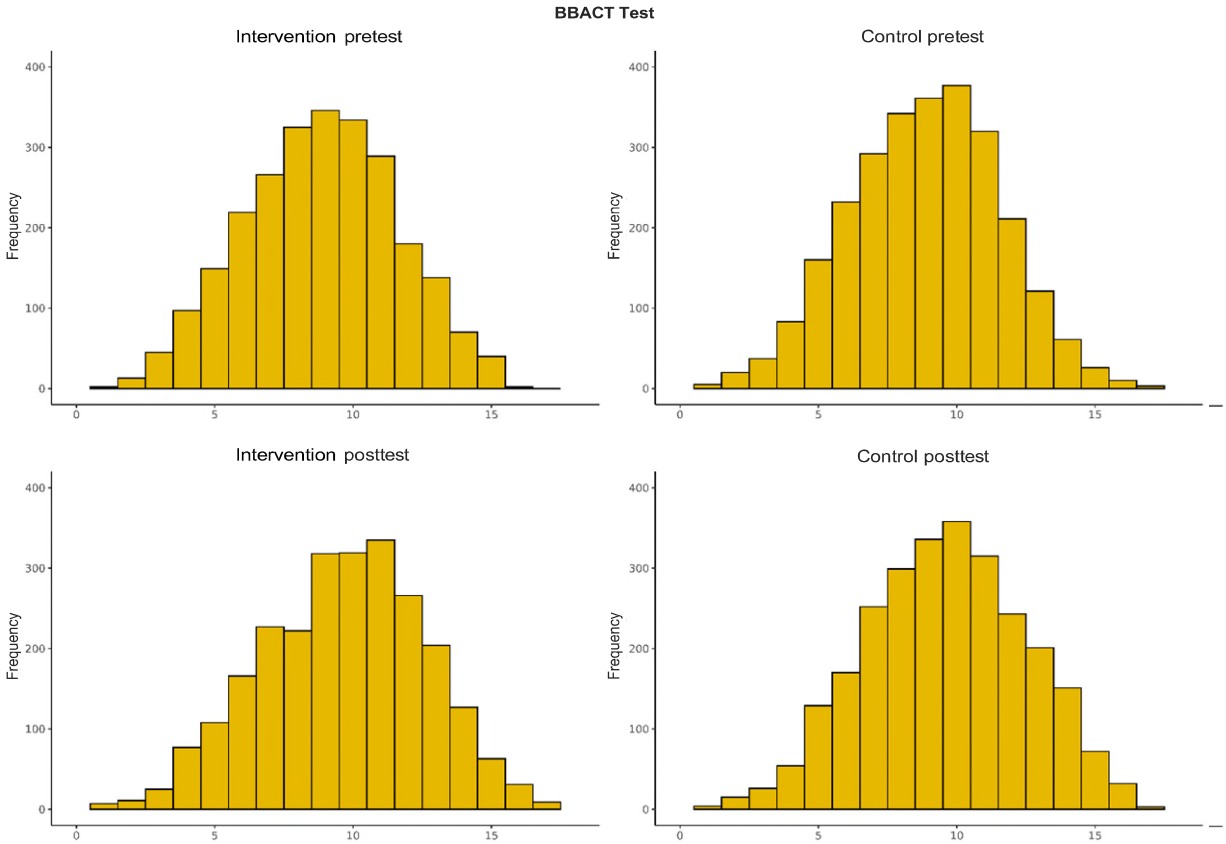
Table 12 summarises the BBACT scores in the posttest for the entire sample and in both arms of the study. No statistically significant differences were observed between the groups, even when there was a slightly better performance in the intervention group.

Table 12: Main descriptive statistics of the posttest results in the BBACT score (primary outcome)

Sample	n	Average	SD	Min	P25	Median	P75	Max
Total	5176	9.62	2.93	0	8	10	12	17
Intervention	2515	9.65	2.96	1	8	10	12	17
Control	2661	9.58	2.91	0	8	10	12	17

Figure 3 shows bar charts for the BBACT scores in both arms obtained after the pre- and posttests. Several differences and similarities can be observed. In the intervention group, from pre- to posttest the lowest scores improved and the medium-high scores increased, moving the average up to 9.65 (this was 8.86 in the pretest, see Table 9). In the control group, the number of medium-high scores moderately increased from pre- to posttest, although the modal value was lower in the posttest. Tables 22 and 23 in the **Additional Analyses** section provide more information about how each student performed in each item.

Figure 3: BBACT scores in pre- and posttest for the intervention group (left) and the control group (right).



The results of the mixed-effect model are summarised in Table 13. Of note, there is no significant influence of group (intervention or control) on posttest performance, but the positive association of being in the intervention group is low. This non-significant low association also occurs when controlling for prior CT training. In terms of population size, the lowest association comes from cities/towns between 50 000 and 100 000 inhabitants, with a low positive association in the other groups. Significant associations can be found in gender (with male and non-binary/other participants performing significantly worse than females) and socioeconomic status (the association is significant in the case of low socioeconomic status relative to high socioeconomic status and, in general, the lower the status, the worse the performance).

Table 13: Mixed-effects model for BBACT posttest score (main outcome)

Explanatory variables	Coefficient (CI 95 %, significance)
BBACT pretest score	0.42 (0.39 to 0.45, p < 0.001)*
Group (Intervention)	0.22 (–0.05 to 0.48, p = 0.109)
Gender (Male)	–0.15 (–0.30 to 0.00, p = 0.044)*

Gender (Other)	–0.55 (–0.89 to –0.22, p = 0.001)*
Previous CT Experience (Yes)	0.34 (0.04 to 0.63, p = 0.027)*
Social Economical Status (Low)	–2.18 (–3.37 to –0.99, p < 0.001)*
Social Economical Status (Medium-Low)	–1.23 (–2.38 to –0.99, p = 0.038)*
Social Economical Status (Medium-High)	–0.89 (–2.03 to 0.26, p = 0038)*
Population (< 50 000)	0.26 (–0.20 to 0.72, p = 0.264)
Population (100 000-250 000)	0..27 (–0.29 to 0.84, p = 0.345)
Population (> 250 000)	0.33 (–0.15 to 0.81, p = 0.185)

Secondary Analysis

Table 14 summarises the IAM scores in the posttest for the entire sample and both arms of the study. No significant differences were observed between the groups, even when there was a slightly better performance in the intervention group. As the items in the IAM test were rated from 0 to 4 (Likert-type with 5 possible answers), the midpoint of a 32-item scale would be 64 points. The average value obtained for the IAM in this sample was almost 20 points above the average value, meaning that attitudes were already rated high in the pretest (see Table 10), making a significant improvement harder to achieve.

Table 14: Main descriptive measures of the post-test results in the IAM score (secondary outcome)

Sample	n	Average	SD	Min	P25	Median	P75	Max
Total	5176	81.11	17.07	22	69	82	93	124
Intervention	2515	81.26	17.10	22	69	82	93	123
Control	2661	80.97	17.04	23	69	82	93	124

Figure 4 shows the histograms of the IAM test scores in both arms obtained after the pre- and posttests. The two arms clearly followed a similar trend (fairly symmetric, not very skewed) in both tests. Between the pre- and posttests there was a minor increase in the medium-low scores, which explains the decrease in the average score values.

The results of the mixed-effect model for the IAM score are shown in Table 15. Like the BBACT results, there is no significant influence of the group (intervention or control) on posttest performance, but the positive association of being in the intervention group is low. This is observed for participants without prior CT training, which suggests a small improvement in IAM scores for those without such training. In terms of population size, the associations are divergent but non-significant: cities/towns between 50 000 and 100 000 inhabitants and over 250 000 perform lower than cities/towns below 50 000 and between 100 000 and 250 000 inhabitants. The gender effect is also opposite to what happened with BBACT: females performed moderately higher than non-binary/other participants, and males performed significantly higher than the other two groups. Significant associations can also be found in socioeconomic status, in this case, in line with BBACT scores, but with an even stronger association: notably, the higher the status, the higher the levels on the IAM scale.

Figure 4: IAM scores in pre- and posttests for the intervention group (left) and the control group (right).

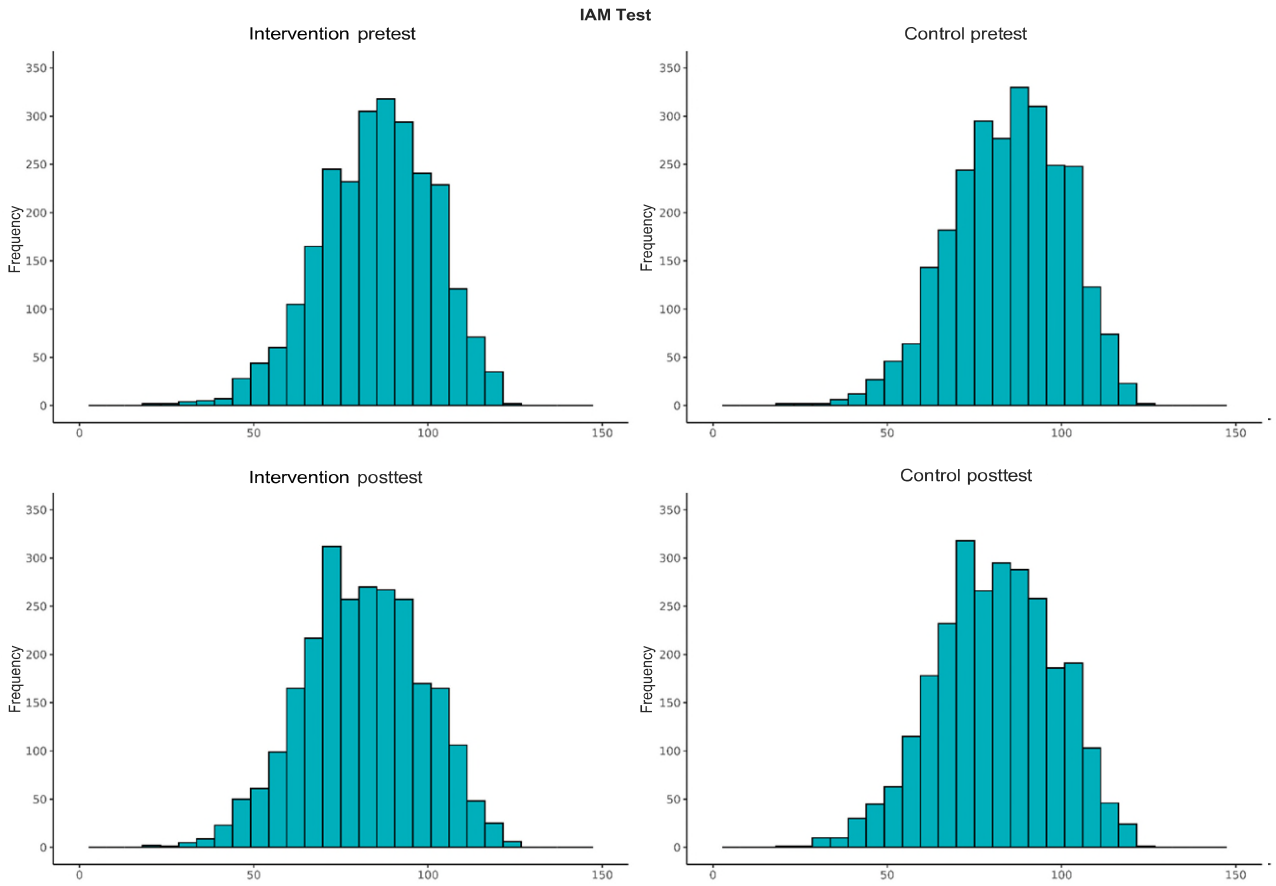


Table 15: Mixed-effects model for IAM posttest score (secondary outcome)

Explanatory variables	Coefficient (CI 95 %, significance)
IAM pre-test score	0.50 (0.48 to 0.53, p < 0.001)*
Group (Intervention)	0.07 (–1.18 to 1.32, p = 0.454)
Gender (Female)	0.35 (–1.54 to 2.25, p = 358)
Gender (Male)	3.08 (1.18 to 4.97, p = 0.001)*
Previous CT Experience (Yes)	–0.69 (–2.09 to 0.71, p = 167)
Social Economical Status (Medium-Low)	3.12 (0.91 to 5.34, p = 0.003)*
Social Economical Status (Medium-High)	3.78 (1.47 to 6.10, p = 0.001)*
Social Economical Status (High)	6.68 (0.50 to 12.86, p = 0.017)*
Population (50 000-100 000)	–0.64 (–2.47 to 1.45, p = 0.273)
Population (100 000-250 000)	0.41 (–1.70 to 2.53, p = 0.351)
Population (> 250 000)	–0.46 (–1.90 to 0.99, p = 0.268)

Analysis in the Presence of Non-Compliance

This analysis was not necessary as all the schools fulfilled the ≤ compliance criterion (J ≥ 80).

Missing Data Analysis

With a database of n = 5176, missing values of over 5 % were found in the BBACT and IAM scores. In the case of the primary outcome, Table 16 shows the percentages of missing cases for BBACT scores and we can see that these percentages are over 11 %.

Table 16: Percentage of missing cases in the BBACT score in pre- and posttests

BBACT #Item	Pretest	Posttest	BBACT #Item	Pretest	Posttest
1	11.28	11.42	10	12.98	12.11
2	12.64	12.21	11	13.91	12.54
3	11.23	11.36	12	13.16	12.29
4	13.06	12.71	13	13.31	13.02
5	11.15	11.30	14	12.25	11.69
6	11.36	11.36	15	13.62	12.37
7	11.57	11.59	16	13.89	11.77
8	11.80	11.53	17	12.54	11.79
9	11.77	11.46			

In the case of IAM scores, both in pre- and posttests, we also found a percentage of missing data exceeding 11 % (Table 17).

Table 17: Percentage of missing cases in the IAM score in pre- and posttests

IAM #Item	Pretest	Posttest	IAM #Item	Pretest	Posttest
1	10.92	11.57	17	11.57	13.43
2	11.05	11.73	18	11.77	12.93
3	11.36	13.18	19	11.22	12.33
4	11.24	12.38	20	11.55	12.17
5	11.26	12.52	21	11.34	12.33
6	11.51	12.60	22	11.50	12.44
7	11.22	12.35	23	11.69	13.18
8	11.67	12.25	24	11.53	12.69
9	11.55	13.00	25	11.55	12.58
10	11.59	12.96	26	11.69	12.46
11	11.69	12.81	27	11.71	13.00
12	11.55	13.25	28	11.79	12.85
13	11.32	13.16	29	11.63	13.27
14	11.36	12.58	30	11.73	13.23
15	11.63	12.83	31	11.48	12.96
16	11.79	12.85	32	11.50	12.29

Following, are the reasons behind these percentages. There is no pattern in the difficulty of the BBACT items, so some students simply left the answer blank. In the case of the IAM test, the score was placed at the end of the instrument, so the effects of fatigue and boredom among the students were possible, even when the instrument was prepared to be completed in a short time. Some missing observations were due to a mismatch of the student identification code with the data and other reasons included student absence on the day of the tests.

To determine the missing data pattern in the BBACT and IAM scores, we constructed a mixed-effects logistic regression using the *lme4* package in R (Bates et al., 2015), considering 1 as missing data and 0 as complete data, based on the other variables in the study. The ANOVA table for the constructed model showed that some variables reached statistical significance, indicating their association with missing data. This suggests that the missing data pattern aligns with the MAR (Missing at Random) assumption, classified by Rubin (1975), as indicated in Tables 18 to 21. It should be noted that we built the model for the total BBACT and IAM scores, not for individual items. Given that more than 5 % of the data is missing, we considered applying multiple imputation by chained equations (MICE), using the *mice* package in R.

Table 18: Analysis of deviance for BBACT pretest scores (Type III Wald χ^2 tests). Significance codes: 0 ***, 0.001 **, 0.01 *

	χ^2	df	P(> χ^2)	Significant
(Intercept)	17.4689	1	2.92e– 05	***
Intervention group	1.6435	1	0.1998	
Type of school	1.3428	2	0.5110	
Socioeconomic index	8.8152	1	0.0030	**
Population	3.8136	3	0.2823	
Previous experience in CT	0.2151	1	0.6428	
Gender	0.1266	2	0.9386	

Table 19: Analysis of deviance for BBACT posttest scores (Type III Wald χ^2 tests). Significance codes: 0 ***, 0.001 **, 0.01 *

	χ^2	df	P(> χ^2)	Significant
(Intercept)	25.0456	1	5.599e– 05	***
Intervention group	0.7841	1	0.3759	
Type of school	2.6881	2	0.2608	
Socioeconomic index	0.1080	1	0.7424	
Population	2.8403	3	0.4169	
Previous experience in CT	0.3592	1	0.5490	
Gender	20.5828	2	3.392e– 05	***

Table 20: Analysis of deviance for IAM pretest scores (Type III Wald χ^2 tests). Significance codes: 0 ***, 0.001 **, 0.01 *

	χ^2	df	P(> χ^2)	Significant
(Intercept)	18.2504	1	1.937e– 05	***
Intervention group	1.3064	1	0.2530	
Type of school	1.3840	2	0.5006	
Socioeconomic index	8.7115	1	0.0031	**
Population	3.8076	3	0.2830	
Previous experience in CT	0.1172	1	0.7321	
Gender	0.0870	2	0.9574	

Table 21: Analysis of deviance for IAM posttest scores (Type III Wald χ^2 tests). Significance codes: 0 ***, 0.001 **, 0.01 *

	χ^2	df	P(> χ^2)	Significant
(Intercept)	25.7723	1	3.842e– 07	***
Intervention group	0.7299	1	0.3929	
Type of school	2.5521	2	0.2791	
Socioeconomic index	0.1872	1	0.6653	
Population	3.0153	3	0.3893	
Previous experience in CT	0.3397	1	0.5600	
Gender	18.6397	2	8.963e-05	***

Tables 18 to 21 show that missing data is not significantly influenced by group assignment, school type or population size. However, in Tables 17 and 19 the socioeconomic index was relevant in the missing data in the pretest on both scales (BBACT and IAM) and gender impacted missing data in the posttest on both scales. By performing a post hoc logit regression, we determined that in the pretest on both scales, students with a low socioeconomic index appeared more prone to missing data and in the posttest on both scales, male and non-binary/unreported students were more likely to have missing data than females.

Additional Analyses

In addition to the mixed-effects model study, we analysed students' performance item by item on both scales (BBACT and IAM). In terms of BBACT, Table 22 shows the percentages of correct answers per item in the pre- and posttests in both groups. Of note is the comparison of Table 22 with Table 5, which reveals a strong consistency in the levels of difficulty, supporting the content validity of the instrument. We also observed that, in both arms, the percentages of correct answers increased for every item, but to varying degrees. Thus, while in the intervention group, the items with the greatest increase from pre- to posttest were 2, 10 and 1, in the control group the greatest increase was found in items 1, 2 and 11; the smallest increases were observed in items 5, 11 and 4 in the intervention group and items 5, 6 and 17 in the control group. It was more difficult to increase the percentage of correct answers on the less difficult items. Item 15 remained the most difficult and item 5 the easiest, consistently in pre- and posttests in both arms.

Table 22: Percentages of correct answers in BBACT scale in pre- and posttests, in both arms

Intervention				Control			
# Item	Pre	Post	% Variation	# Item	Pre	Post	% Variation
1	52.4	61.6	17.56 %	1	55.5	67.2	21.08 %
2	24.5	30.4	24.08 %	2	25.0	29.1	16.40 %
3	59.5	66.3	11.43 %	3	60.4	63.5	5.13 %
4	38.4	40.0	4.17 %	4	39.1	44.6	14.07 %
5	90.9	93.1	2.42 %	5	92.3	94.2	2.06 %
6	82.5	86.8	5.21 %	6	80.7	84.5	4.71 %
7	58.8	63.9	8.67 %	7	58.5	63.1	7.86 %
8	51.2	57.3	11.91 %	8	50.4	56.7	12.50 %
9	78.6	84.9	8.02 %	9	76.4	83.4	9.16 %
10	37.1	45.5	22.64 %	10	37.1	39.6	6.74 %
11	26.5	27.3	3.02 %	11	24.9	28.6	14.86 %
12	54.8	64.3	17.34 %	12	56.6	63.5	12.19 %

13	66.3	70.3	6.03 %	13	62.8	68.6	9.24 %
14	56.2	61.6	9.61 %	14	53.3	61.0	14.45 %
15	15.5	17.1	10.32 %	15	16.7	17.9	7.19 %
16	55.6	63.9	14.93 %	16	56.6	63.5	12.19 %
17	31.0	36.1	16.45 %	17	30.8	32.2	4.55 %

An analysis per item and participant was also carried out. The results are shown in Table 23. For each item, the percentage of incorrect and correct answers is given. For the posttest, these are divided according to whether the answers remain incorrect or become correct. McNemar's (Pembury Smith & Ruxton, 2020) test for paired categorical values was used and the p-value is included in Table 23. Thus, for example, in item 1, 52.9 % of those who answered incorrectly in the pretest, continued to do so in the posttest, while the remaining 47.1 % improved by selecting the correct option. However, 72.3 % of those who answered correctly in the pretest, continued to do so in the posttest.

Table 23: Analysis of individuals' answers to BBACT per item

# Item	Pre-test answer	% Post incorrect	% Post correct	p-value
1	Incorrect	52.9	47.1	< 0.001
	Correct	27.7	72.3	
2	Incorrect	77.1	22.9	< 0.001
	Correct	62.9	37.1	
3	Incorrect	48.4	51.6	< 0.001
	Correct	38.4	61.6	
4	Incorrect	67.9	32.1	< 0.001
	Correct	60.8	39.2	
5	Incorrect	19.7	80.3	0.037
	Correct	16.5	85.3	
6	Incorrect	36.2	63.8	< 0.001
	Correct	20.1	79.9	
7	Incorrect	55.0	45.0	< 0.001
	Correct	34.7	65.3	
8	Incorrect	56.9	43.1	< 0.001
	Correct	41.7	58.3	
9	Incorrect	39.6	60.4	< 0.001
	Correct	20.1	79.9	
10	Incorrect	68.1	31.9	< 0.001
	Correct	46.6	53.4	
11	Incorrect	76.0	24.0	0.013
	Correct	72.1	27.9	
12	Incorrect	53.0	47.0	< 0.001
	Correct	34.5	65.5	
13	Incorrect	49.4	50.6	< 0.001
	Correct	31.4	68.6	

14	Incorrect	55.1	44.9	< 0.001
	Correct	37.9	62.1	
15	Incorrect	86.4	13.6	0.176
	Correct	73.5	26.5	
16	Incorrect	48.5	51.5	< 0.001
	Correct	39.4	60.6	
17	Incorrect	72.2	27.8	0.001
	Correct	63.8	36.2	

Of note is that the most difficult items (such as 2, 11 or 15) obtained the lowest percentages of correct answers in both pre- and posttests, with items 5, 6 and 9 obtaining the highest percentages of correct answers in both tests.

A similar analysis was performed for IAM. Table 24 shows the average score per IAM item in the pre- and posttests in both groups. It should be noted that the IAM items are presented on a Likert scale with five levels coded from 0 (*Strongly disagree*) to 4 (*Strongly agree*). Therefore, 2 is the midpoint of the scale. In contrast to the BBACT test, the IAM showed that there were many items where the average score decreased. In both groups, item 4 presented the largest decrease (over 10 % in both cases), whereas item 13 increased its average score by more than 6 %.

Table 24: Average scores of IAM answers for pre- and posttests, in both arms

Intervention				Control			
# Item	Pre	Post	% Variation	# Item	Pre	Post	% Variation
1	3.21	3.12	–2.80 %	1	3.22	3.14	–2.48 %
2	2.99	2.84	–5.02 %	2	3	2.81	–6.33 %
3	3.33	3.14	–5.71 %	3	3.33	3.17	–4.80 %
4	2.62	2.31	–11.83 %	4	2.57	2.31	–10.12 %
5	2.09	2.01	–3.83 %	5	2.15	2.05	–4.65 %
6	2.68	2.45	–8.58 %	6	2.7	2.48	–8.15 %
7	2.46	2.36	–4.07 %	7	2.51	2.47	–1.59 %
8	2.81	2.63	–6.41 %	8	2.86	2.68	–6.29 %
9	3.08	2.87	–6.82 %	9	3.08	2.93	–4.87 %
10	3.4	3.15	–7.35 %	10	3.45	3.19	–7.54 %
11	2.97	2.69	–9.43 %	11	3.03	2.78	–8.25 %
12	2.91	2.71	–6.87 %	12	2.98	2.76	–7.38 %
13	2.03	2.17	6.90 %	13	2.12	2.26	6.60 %
14	2.17	2.08	–4.15 %	14	2.13	2.09	–1.88 %
15	2.18	2.03	–6.88 %	15	2.22	2.05	–7.66 %
16	3.31	3.15	–4.83 %	16	3.33	3.17	–4.80 %
17	3.26	3.09	–5.21 %	17	3.24	3.08	–4.94 %
18	2.29	2.15	–6.11 %	18	2.33	2.21	–5.15 %
19	2.69	2.5	–7.06 %	19	2.72	2.52	–7.35 %
20	1.32	1.31	–0.76 %	20	1.33	1.34	0.75 %

21	3.2	3.2	0.00 %	21	3.18	3.21	0.94 %
22	2.82	2.85	1.06 %	22	2.87	2.85	–0.70 %
23	2.25	2.25	0.00 %	23	2.21	2.19	–0.90 %
24	1.44	1.46	1.39 %	24	1.43	1.49	4.20 %
25	2.6	2.49	–4.23 %	25	2.7	2.54	–5.93 %
26	2.54	2.44	–3.94 %	26	2.52	2.46	–2.38 %
27	2.77	2.54	–8.30 %	27	2.85	2.57	–9.82 %
28	2.6	2.53	–2.69 %	28	2.69	2.58	–4.09 %
29	3.15	3.02	–4.13 %	29	3.19	3.04	–4.70 %
30	2.45	2.35	–4.08 %	30	2.52	2.35	–6.75 %
31	2.51	2.39	–4.78 %	31	2.58	2.38	–7.75 %
32	3.03	2.9	–4.29 %	32	3.02	2.86	–5.30 %

IAM items were grouped into dimensions, so analysing the results in terms of this aspect was also of interest. In general, all the dimensions scored lower in the posttest than in the pretest. The results are shown in Table 25 where the lowest average scores are found in the dimension of *Social valuation*, while the highest scores are found in *Parents' attitudes*, consistently in both arms and both tests.

Table 25: Average scores of IAM dimensions in pre- and posttests, in both arms

	Intervention		Control	
Dimension	Pre	Post	Pre	Post
Social valuation	1.77	1.77	1.79	1.78
Anxiety	2.21	2.18	2.19	2.14
Intrinsic motivation	2.49	2.33	2.46	2.3
Feelings	2.58	2.38	2.51	2.39
Perceived incapability	2.73	2.57	2.66	2.51
Perceived competence	2.73	2.57	2.69	2.54
Perceived utility	2.95	2.74	2.96	2.82
Teachers' attitudes	2.98	2.84	2.98	2.73
Extrinsic motivation	3.03	2.95	2.99	2.94
Parents' attitudes	3.2	3.18	3.21	3.16

We also analysed the possible associations between the scores on the two instruments considered (BBACT and IAM), obtaining Spearman's correlation coefficient between BBACT and IAM for all the students in the control and intervention groups, both at pretest and posttest. The association is significant in all four cases, meaning, there is a positive association between the BBACT score and the IAM score. Table 26 summarises the results. As we can see, the intervention group increased the correlation more moderately than the control group. However, causal attributions should not be assumed, given that in all cases there is a significant correlation.

Table 26: Spearman's correlation coefficient (ρ) between BBACT and IAM scores in both arms

	Intervention		Control	
	Pre	Post	Pre	Post
ρ coefficient	0.236	0.275	0.208	0.286
p-value	$p < .001$	$p < .001$	$p < .001$	$p < .001$

Estimation of Effect Sizes

The effect size for both outcomes was calculated by dividing the adjusted mean difference between the intervention and the control group by the pooled variance obtained from running the unconditional mixed model, adjusting only for group and clustering at the school level. The pooled variance was obtained by summing the between- and within-cluster variance. A 95 % CI for the effect size was calculated by dividing the 95 % confidence limits of the adjusted mean difference by the same standard deviation (see Table 27).

Table 27: Effect sizes for both analyses. (Values in the table have been calculated based on complete cases and the raw means are for the posttest score. P-values correspond to the group coefficient in the mixed model.)

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	*n* (missing)	Mean (95 % CI)	*n* (missing)	Mean (95 % CI)	Total *n* (intervention; control)	Hedges *g* (95 % CI)	*p*-value
BBACT (main)	2515 (283)	9.65 (9.35, 9.76)	2661 (287)	9.59 (9.47, 9.69)	5121 (2482; 2639)	0.074 (–0.011, 0.160)	0.109
IAM (secondary)	2515 (288)	79.87 (79.11, 80.62)	2661 (290)	79.90 (79.18, 80.61)	3984 (1956; 2028)	0.007 (–0.069, 0.084)	0.875

Estimation of ICC

The intraclass correlation coefficient (ICC) from each multilevel analysis model, with the 95 % CI, was determined for the primary and secondary outcomes (see Table 28) by using the performance R package. As we can see, the ICC is relatively low in both cases.

Table 28: Intraclass correlation coefficient (ICC) of both analyses

Outcome	ICC (95 % CI)
BBACT (main)	0.031 (0.014, 0.048)
IAM (secondary)	0.026 (0.011, 0.054)

Implementation and Process Evaluation Results

Pre-intervention

Family information and participation: Families were informed about the study in a detailed letter from the school directors. The reception was highly positive, with fewer than 3 % of families opting out of the study (RQ3).

Teacher training and participation: Teachers from each participating school attended an initial training session, as described in the Methods section. Each school had at least one teacher in attendance, with an overall participation rate averaging 75 % across schools. The teachers demonstrated high levels of engagement during the training sessions, actively participating in the proposed activities and discussions (RQ1). The quality, adequacy and relevance of the training were rated highly by the teachers, with an average score of 4.58 out of 5, with 5 being the modal value (RQ2).

Teacher questionnaire results: Teachers completed a questionnaire between 5 and 29 September 2023. Key results included:

- Engagement in discussions: Teachers rated their engagement at an average of 4.80 out of 5, with 5 being the modal value (RQ2).
- Perception of differences in mathematics lessons: The difference between regular maths lessons and those proposed in the intervention was rated at an average of 4.23 out of 5, with 5 being the mode (RQ6).
- Rethinking their practice: Teachers indicated that the training encouraged them to rethink their teaching practices, with an average score of 4.13 out of 5, with 4 being the modal value (RQ6).

Online skills survey: An online survey evaluated the teachers' proficiency in educational design. On a 10-point scale, the average score was 8/10 (RQ4).

Intervention plans: Each school submitted an intervention plan based on the programme materials. The intervention team reviewed the submitted plans, provided feedback, and evaluated them. Scores for the adaptation and use of the materials averaged 7/10 (IPE RQs 4 and 5).

Early Intervention

Evaluation of the intervention plans (RQ9): A template was provided to teachers as a planning tool for the educational intervention. It asked them to plan the entire 17-week programme by breaking it down into sessions, specifying the duration in weeks for each one. For each session, teachers were required to outline the specific learning objectives, define the Exploding Dots activities to be used to meet those objectives and describe the assessment methods. Each planned intervention (one per school) was shared with the intervention team through the Google Classroom platform for review. Each plan received feedback and scores, with an average rating across submissions of 8/10.

Follow-up of early implementation (RQs 7, 8, 10): The Google Classroom platform allowed teachers to ask questions and raise any issues with the intervention team. These conversations were monitored by the commission and organisation team to ensure that all concerns were addressed. Additionally, this team maintained weekly contact by email with the schools participating in the intervention, encouraging teachers to submit their questions and provide updates on their progress.

The commission and organisation team was also regularly in contact with the control schools via email to keep them engaged, providing key project updates, such as participation in the pretest and the number of students impacted. The importance of their role was consistently emphasised. No changes in practice (R11)

were observed, though no formal monitoring was conducted beyond the statement included in the randomisation results letter.

The pretest of both the control and intervention groups (RQ12) was done online, so participation was automatically recorded.

During the Intervention

As follow-up tools during the intervention, consistent communication was maintained with all teachers and the teams involved via a weekly newsletter and on the Google Classroom platform.

In-class observations were carried out in 34 of the 40 schools in the intervention group, in two periods: November 2023 and January 2024 (see the Methods section for further details about school sampling). Semi-structured interviews were held before, during or after the lessons (depending on the availability and schedules of each school) with pupils and teachers.

Observations (RQs 13, 14, 15, 16 and 17) were conducted following a two-indicator rubric in which the evaluators checked the use of the designed materials and the classroom climate (subdivided into respect, socioemotional support, motivation and persistence). Without exception, all the schools used the designed materials. The vast majority of the observed schools developed their classes in a climate of respect and warmth, engagement and opportunities for discussion, in which teachers promoted students' risk-taking by posing questions or asking them to answer questions posed by classmates. However, in one school, participation was notably low due to significant absenteeism following a cultural festival the day before the observation. While the classroom climate in this case was not negative, the low participation impacted the dynamic.

In terms of student interviews (RQs 13 and 14), these were semi-structured and based on four flexible questions, so that, depending on the students' answers or on the situation, the evaluators could pose different questions or vary the initial ones. These questions were the following: (1) Do you like ED? Why? (2) Do you like it better than the other maths classes? Why? (3) Do you like maths more now? (4) Do you like maths? When allowed by the school (and the parents), the interviews were recorded and then transcribed. When this was not possible, notes were taken during the interviews. More than 30 hours of recordings were collected from the student interviews. These were analysed using inductive (to obtain themes) and deductive (to analyse the presence of IAM instrument dimensions) content analyses.

With regard to the inductive content analysis, the following themes were identified (all excerpts come from different, anonymous, students):

- Positive engagement and interest in ED: Students expressed enjoyment and interest in ED, appreciating its originality and interactive nature. They found it entertaining, creative and a refreshing alternative to traditional maths, even when they acknowledged intrinsic difficulty with the subject. "I like ED, it's fun. It gives you a lot of freedom… and I think it's a good way to learn maths", "ED is fun although it's quite difficult to work with the numbers; having the dots as support makes it easier".

- Preference for ED and benefits from using it: Most students recognised ED as a helpful tool for understanding maths, they found it helpful for concepts like powers and visualising maths processes. "ED is easier, it's the same as in primary school but now it's easier", "ED is another way of adding and it's easier to learn", "With ED I learn more and it helps me with normal maths lessons". There is a mixture of arguments on ease of use, but also on offering a different perspective that enhances understanding.

- ED as an engine for understanding: Many students valued ED's uniqueness, finding it innovative and different from rote memorisation in traditional maths. They felt it encouraged critical thinking and a new approach to learning. "ED is like maths but more dynamic, without so much theory", "ED is more about observing and understanding the process and thinking".

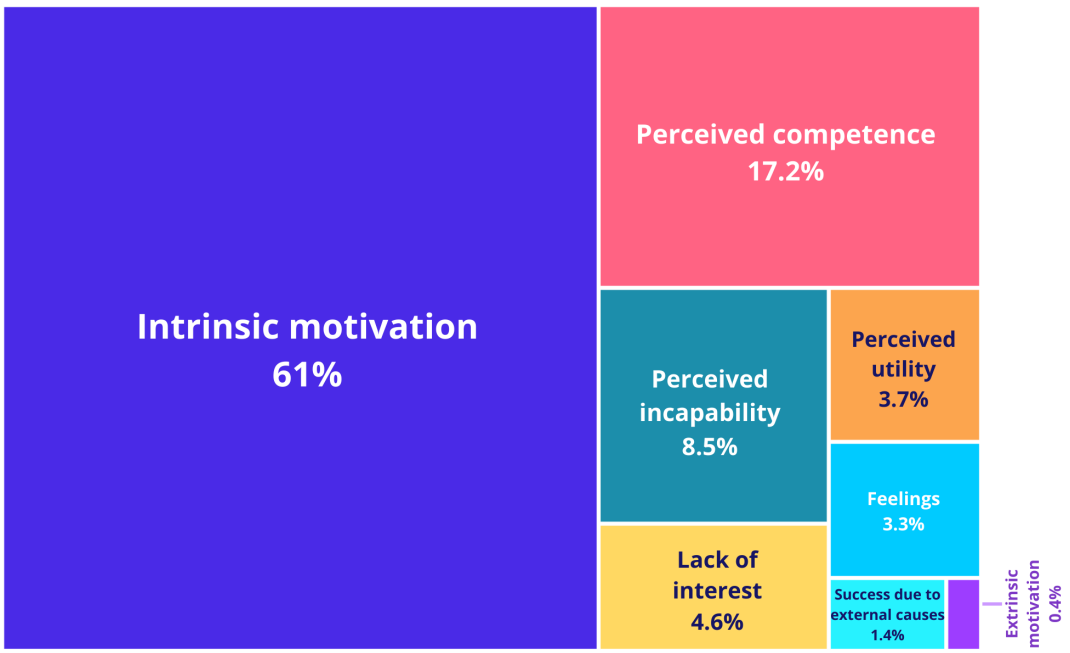
- Mixed feelings and preference for 'other maths': Some students preferred traditional maths, citing familiarity, ease or perceived importance. Others appreciated ED but still favoured aspects of traditional maths: "I don't like ED because it is more difficult. Maths is more about theory, you simply memorise it", "Normal maths is easier than ED because we have been doing it all our lives", "I don't like ED; I don't like maths because I'm bad at it", "I think other maths lessons teach me more productive things". There is a mixture of conceptions of maths as an instrumental set of knowledge, but also a low self-esteem in maths.

- Bored of using ED: A small portion of students expressed that ED was boring. "I don't understand ED. It's a little bit boring", "ED is boring because it's very easy", "It's different, but once you know the topic, you have to keep doing it and it gets a bit boring".

- Teacher's role: The teacher's attitude, teaching style and classroom dynamics impacted students' perception of ED, with many enjoying ED more due to positive teacher interactions. "I like ED because E. [the teacher] is nice and explains it well", "I like ED much more. Also, the teacher motivates me".

The deductive content analysis complemented the inductive one, providing the evaluators with a whole panorama of how the students' feelings, likes, conceptions and interests influenced their perception of the project. Figure 5 shows that the most frequent IAM dimension present in their answers was intrinsic motivation (in more than 60 % of excerpts), followed by perceived competence (over 17 %) and finally, the remaining dimensions with percentages below 9 %.

Figure 5: Results on deductive coding (IAM based) from students' answers to the semi-structured interviews during the intervention



Below are excerpts from some examples of the recordings. Overall, positive statements were much more frequent than negative ones:

- Intrinsic motivation: "I prefer ED because I like learning through play and it awakens my curiosity", "I like ED, it's a fun way to learn", "I like ED because it is about mathematical reasoning, you think with your brain", "ED is fun and creative. You learn maths differently", "I like ED because it is an innovative, complex and fun way to learn maths".
- Perceived competence: "I like ED, it's easy, I understand it pretty well", "ED is easier than other maths classes, you can use the drawings to understand it better", "ED is easier, it's like playing a game".
- Perceived incapability: "I don't like ED; I don't like maths because I'm bad at it", "I don't like ED because it is more difficult", "I find it complicated and I'm very bad at maths".
- Lack of interest: "ED is too easy and I get bored", "I like ED, although I still don't really understand why we are doing it", "I prefer ED because we don't do anything".
- Perceived utility: "With ED you learn maths without realising it", "I like maths because it is a subject that prepares you for the future", "I only like maths when it works out; otherwise, I get frustrated".
- Feelings: "I love it, it's really fun. It brightens up my day", "I don't like ED, I don't like doing subtraction exercises", "I like ED".
- Success due to external causes: "If I get a good grade, I like it better", "I got a 10 in the [ED] test and I don't always get a 10 in maths".
- Extrinsic motivation: "I like ED because E. [the teacher] is nice and explains it well", "In maths, you have the pressure of having to understand it to get a good mark in the exam".

The teacher interviews (RQs 15, 16 and 17) were also recorded, when possible, except in some cases where they were conducted in places with too much noise (school corridors or canteens). The interview was semi-structured and open to any observations or suggestions from the teacher. The themes proposed were the problems, obstacles, unforeseen events or changes in the implementation of ED, the relationship with the intervention team, the perception of self-efficacy during the implementation, the connection of ED with the maths syllabus, the impact of the implementation on teaching practices and the perception of the impact on students. As these interviews were much longer than those of the students, below is a summary of the main results. The evaluation team still needs to conduct further analysis, probably using grounded theory, to determine if any theory can be inferred from these data.

In general terms, teachers stressed that they felt comfortable and confident with the implementation and that the relationship with the intervention team was fluid. Any doubts that had arisen were solved immediately. The following excerpts illustrate this point.

- "I feel comfortable, sure of myself."
- "We have exchanged several messages and they [the intervention team] have always been very attentive, really good."
- "In fact, I am in permanent contact with Eulàlia [from the intervention team]. I am just about to start with multiples and factors and I asked her [how to do it with Exploding Dots]."

Regarding the teachers' view of ED, most of them stated that ED is aligned with their ideas about teaching and learning mathematics, hence the intervention was not a big change in terms of the didactic perspective

they already had, nor to 'business as usual'. The change was more to do with content than methodologies. The first excerpts show the alignment between the teachers' conceptions of maths and the ED approach, whereas the second illustrates a moderate change, assuming that their conception of teaching and learning mathematics was already updated and aligned with ED.

- "Perhaps it hasn't changed my approach, but it has given me an additional tool that aligns somewhat with the ideas I pursue. I mean, I really like it – I always tell them [the students] that mathematicians are lazy, slackers because we always try to do things in the most efficient way possible, right? And this method encourages exactly that. In many situations, it forces you to think about how to do things in the most efficient way possible. That's why I like it – that is its strength. It reinforces, so to speak, my way of teaching because it's a tool that allows me to reinforce those ideas."

- "Well, it helps to do certain things, but no, not in particular (the view on learning-teaching in mathematics)."

Many teachers agreed that students did not properly see the usefulness of the method and saw it more as a game. For instance, in this excerpt, the teacher stressed that students did not explicitly perceive the usefulness but, indirectly, made sense of it in further lessons.

- "If you ask me whether my students have found it useful, I'd say no – it was just an anecdote. However, they are finding meaning in the lessons we are doing."

In the following excerpt, the teacher points out that abstraction as a feature makes the perception of ED's usefulness more difficult.

- "You can say it's a mathematical game, but one that has no real-world application; it is very abstract in that sense and purely mechanical if you prefer."

The lack of perception of usefulness led some students to get bored with the programme.

- "And well, yes, it's true that perhaps we had mentioned it – in the first sessions there was a certain level of curiosity, right? The expectation, wondering what would happen, how it would go... Then, it stagnated a bit, and we also noticed it ourselves."

Often, more able students found ED too easy and then boring, but those students who were able to extrapolate and reflect further (e.g. applications to polynomials) had a much richer experience. In contrast, if the students were not able to make connections, the experience became less productive.

- "For some, I do believe it enriches them and they are able to see beyond."

- [Speaking as a student] "I am doing it because I'm required to, but no, I don't connect with it." [Speaking as the teacher] "It has that abstract aspect that makes it difficult for them to engage with. With the videos and everything available, they do try to interact a little and the sounds and effects aim to make it more engaging, but it's still challenging."

As for the others, teachers pointed out that many students enjoyed it because it was a dynamic and new method, as illustrated in the following excerpt.

- "Students who tell you at the beginning of the course that they don't like maths end up changing their mindset, realising: 'Wow, it's actually cooler than I thought it was back in primary school'."

- "There are students who aren't really present in class, their minds are elsewhere and you can't reach everyone. But for those who are receptive, everything improves – from their perception of maths to opening up to new ways of thinking and learning. They absorb knowledge like sponges."

In almost all interviews, teachers pointed out that it was still too early to see results in terms of improvements in numeracy or problem-solving, but they had already observed that students seemed to be reacting to the subject more positively with respect to the rest of the maths lessons. Teachers saw this as an immediate benefit, arguing that it may reduce some of the blocking of the students most frustrated with the subject.

- "I believe this is indeed a long-term process. But yes, I do think a seed is being planted and sooner or later, it will grow. [...] In two or three years, once these 1st-year students have progressed through year 2 and year 3, having explored machines inside out, mastering the foundations in every possible way, I believe it will bear fruit."

- "Even the students who struggle the most are not completely disconnected."

Many teachers reported individual cases of students (with or without special needs) for whom the implementation significantly changed their attitudes towards mathematics. The first excerpt refers to a student with dyslexia and the second is a general comment about students who are good at concentrating.

- "I've noticed a change in a student with dyslexia – she struggles with worksheets, but when it comes to doing the activities, she feels much more comfortable. With maths, dealing with numbers and everything can be overwhelming. But with this, since there are just boxes, it's much easier for her than a worksheet full of information. I've really seen a difference – she feels much more at ease with ED compared to other methods."

- "I like it because I think that, in some way, this kind of strategy can reach some children who would otherwise be harder to engage."

Many teachers pointed out that ED is a good way to refresh the knowledge of students who did not develop adequate number sense in primary school or who simply memorised algorithms and procedures without conceptual understanding. The novelty of ED allows knowledge to be rebuilt on a solid basis. This is illustrated in the following excerpt.

- "The starting point, in the end, is common for everyone because we introduce the machines from scratch. So, you don't rely on previous knowledge and, therefore, everyone can get on board, so to speak."

Some teachers stressed that, while the intervention adequately catered for diversity in terms of learning difficulties, the materials may have been too easy for more able students and they felt that there were no high-level activities or tasks for these students or those who were faster and able to make connections easily.

- "Some [students] grasp everything very quickly in 30 seconds, while others, of course, need much more time. And, as always, those who get bored are the ones who have done it straight away."

There were also some responses about the difficulties some students had in understanding the statements because their reading comprehension was poor.

- "Because nowadays, children don't read. The materials available need to be read carefully to understand what is being asked. But now, everything for children is visual."

As for the relationship with the curriculum, although all the teachers implemented ED as a separate lesson (following indications of the intervention team), they recognised the clear relationship to the curricular content of number systems, arithmetic and computational thinking. They also detected unexpected connections and sometimes made them explicit in the classroom (e.g., powers and roots).

- "Well, in fact, it has been useful for me because I start like everyone else with the numeration block and begin with Topic 2: powers."
- "That connection between mathematics and ED is wonderful and doesn't always happen. So when it does, it must be taken advantage of. So, what I try to do is interrelate what we do in the mathematics subject with what we do in ED, helping them understand that what we do in maths is connected to what we do in ED – that ED is not just a separate world where we simply explode dots and that's it."
- "Regarding the curriculum, well, as an extension of numerical sense, I think it does make sense here."
- "In my case, I happened to be explaining powers using the worksheet, which was also about powers – so we took advantage of that connection."

However, many teachers did not believe that ED could be extrapolated across the curriculum. In addition, there was disagreement about the level of implementation of ED. Some teachers stated that they would discard all previous experiences and would use ED much more in years 2 and 3 of secondary school, than in the first year.

- "Maybe it could work [extrapolating ED], but I can't quite picture it. For example, with fractions, how to add them, or topics related to measurement or geometry. Not proper geometry."
- "It would have been more appropriate to apply this in year 3 because that's when students start working with integers, as integers are introduced in the first topic."

On the other hand, some teachers argued that the initial number-system activities using ED are an appropriate method that should be introduced in primary school to create a stronger foundation for beginning secondary school.

- "I don't know what will happen with multiplication and subtraction, but I think addition could be introduced at an earlier age."

Finally, there were some differences about the possible manipulative use of the materials, while several teachers mentioned that they would have liked to have ready-to-use materials to develop the tasks manipulatively, others were happy to build their own materials, adapting them to the needs of their class.

- "It's exactly what we were missing – something more… more constructive, more manipulative."
- "Some students need to physically place the cube or put it in the box to understand it better, and I think any hands-on approach we can incorporate is beneficial for them."
- "[One student] instead of solving problems with pen and paper, she made a box out of modelling clay – and honestly, she showed initiative, which is something you don't always see in class."

Post-intervention

After the intervention (in February 2024), two questionnaires (Spanish and Catalan versions) were administered to the intervention group: one for teachers and one for students (RQs 13, 14, 18, 19 and 20). A total of 1291 students answered the anonymous questionnaire, consisting of six statements to which they had to indicate their degree of agreement, according to a four-level Likert scale: *Disagree*, *Somewhat agree, Mostly agree, Totally agree.* Table 29 shows the results of each question.

Table 29: Results of the questionnaire administered to students after the intervention

		Number of answers (percentage)			
#	Statement	Disagree	Somewhat agree	Mostly agree	Totally agree
1	I like Exploding Dots	313 (24.24 %)	358 (27.73 %)	400 (30.98 %)	220 (17.04 %)
2	I think Exploding Dots is interesting	247 (19.13 %)	361 (27.96 %)	434 (33.62 %)	249 (19.29 %)
3	When it's time for maths, I prefer to do Exploding Dots	417 (32.30 %)	312 (24.17 %)	269 (20.84 %)	293 (22.70 %)
4	I like maths	245 (18.98 %)	331 (25.64 %)	351 (27.19 %)	364 (28.20 %)
5	I think Exploding Dots is easy	255 (19.75 %)	408 (31.60 %)	391 (30.36 %)	236 (18.28 %)
6	I understand maths better thanks to Exploding Dots	580 (44.93 %)	381 (29.51 %)	205 (15.88 %)	125 (9.68 %)

The ED programme is mostly perceived as interesting by the students (statement 2), although they prefer maths when asked about preference (statements 1 and 4). Almost one-third of students do not prefer doing ED in their maths lessons (statement 3), but over 44 % prefer it mostly or totally. The perception of ease is quite divided, less than 20 % of students do not consider ED at all easy (statement 5), but there is still some degree of agreement among the rest. There is a relevant perception that ED does not make maths easier to understand, 44.93 % do not think ED contributes to a better understanding (statement 6).

When analysing the results in Table 29 together with those from the secondary outcome (the mixed model for the IAM score), several relationships can be detected: attitudes towards maths did not improve in the posttest (in either group), as they were initially quite good (pretest). The effect of the intervention was not significant, but this is better understood by the fact that more than half the intervention students agreed with the statement 'I like maths'. This means that the initial level of positive attitudes was high and even though ED obtained very satisfactory results in terms of being liked, it was not perceived as a driver of mathematical understanding.

The teacher questionnaire was also administered in February 2024 and 47 teachers from the intervention group schools participated. The questionnaire consisted of eight items on a Likert scale from 1 to 4, in which teachers had to show their degree of agreement (1 being the lowest and 4 the highest) with the statements; two items in which teachers had to select (from a closed list of options) the statement that best represented their opinion; three items to indicate the percentage of compliance and, finally, one open question for further comments.

As Table 30 shows, for items 1 to 8, the degree of agreement with the statements was generally very high, with distributions markedly skewed to the right in all the cases. Furthermore, for 6 of the 8 statements, less than 25 % of respondents scored 3 out of 4 or less on the agreement scale, indicating a very high degree of agreement. Average scores were higher than 3 for all statements, with small standard deviations. In three cases the median was 4 (the maximum rating on the scale). These statements referred to the usability of the implemented materials (statement 4), the possibility of answering students' questions (statement 5) and the responsiveness of the intervention team (statement 6).

Table 30: Answers to the Likert-type items on the questionnaire administered to teachers after the intervention

#	Statement	Mean	SD	Min	P25	Median	P75	Max
1	The students have been involved during the implementation of MAPS	3.38	0.68	2	3	3	4	4

2	The students have maintained positive attitudes regarding the implementation	3.15	0.72	2	3	3	4	4
3	The implementation has been on schedule	3.19	0.73	1	3	3	4	4
4	The materials designed for the implementation have been used	3.60	0.61	2	3	4	4	4
5	All students' doubts have been resolved	3.62	0.60	2	3	4	4	4
6	The intervention team has been receptive regarding the school's reality	3.57	0.61	2	3	4	4	4
7	The implementation has been linked to the mathematics curriculum	3.09	0.92	1	2.5	3	4	4
8	The students have been motivated during the implementation of ED	3.13	0.84	1	2.5	3	4	4

For items 9 and 10, teachers had to choose (single choice) the sentence (out of 4 and 3 alternatives, respectively) that best represented their opinion. Table 31 shows the results for item 9 and, as we can see, more responses acknowledge the connection and relationship between the curriculum and the ED project. However, some teachers stated that the connection was weak or that they were unconnected.

Table 31: Answers to item 9 of teachers' post-intervention questionnaire

	Choose from the following sentences the one that best represents your opinion.			
Sentences	The implementation and the curriculum have no connection, and I have had to do very different sessions.	The implementation has helped me to better understand some elements of the curriculum, but it does tie in with the education law.	The implementation and curriculum are loosely connected and, occasionally I have been able to advance curriculum content through the implementation.	The implementation and curriculum are fully connected; and, through ED I have covered the legislative requirements.
Number of answers	7	12	15	13

Table 32 reflects the teachers' self-analysis of the impact of ED on their practice. As can be seen, most of the 47 teachers acknowledged a real impact of the project on the way they teach mathematics (the impact can be moderate, n = 31, or total, n = 2). Regarding the 14 teachers that do not consider there to be any impact on their teaching, it should be noted that, as ED was a voluntary programme, a relevant innovative attitude is assumed, together with a concern for the improvement of teaching practice among the participating teachers. Therefore, not changing does not necessarily mean maintaining an attitude of resistance towards innovation, change and positive impact projects in mathematics education.

Table 32: Answers to item 10 of teachers' post-intervention questionnaire

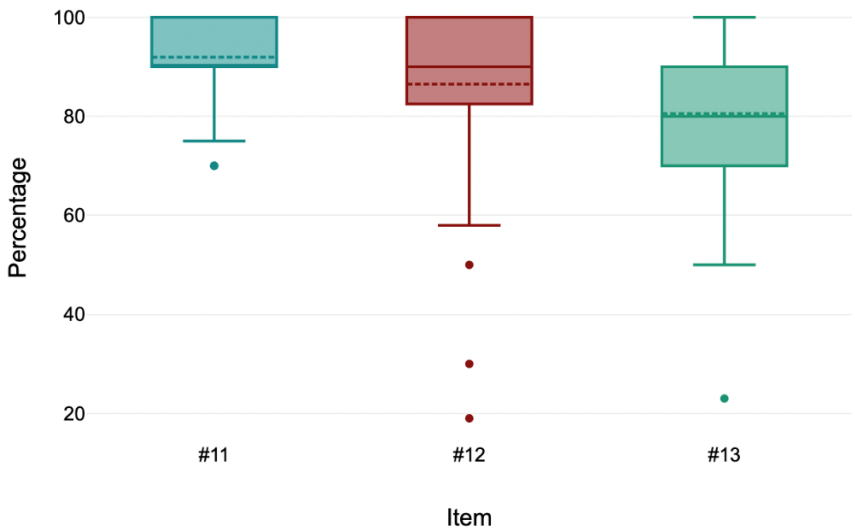
	Choose from the following sentences the one that best represents your opinion.		
Sentences	After the ED experience, I have not changed the way I teach mathematics.	After the ED experience, I have slightly changed the way I teach mathematics.	After the ED experience, I have completely changed the way I teach mathematics.
Number of answers	14	31	2

As for the teachers' responses to the items asking for a percentage, Table 33 shows the main parameters and Figure 6 shows the distribution of responses. A very high percentage of commitment is observed (item 11), with the majority of responses concentrated at 90 % to 100 % of commitment. The distribution is slightly less skewed to the right in item 12 but still shows a very high percentage of active students. Finally, item 13 also shows a high degree of use of the designed materials, although with moderately lower values than the other two items.

Table 33: Statistical description of items 11, 12 and 13 of teachers' post-intervention questionnaire

#	Statement	Mean	SD	Median
11	Rate (from 0 to 100) your level of commitment to the intervention plan	91.94	8.80	90
12	Indicate the average percentage of students who have actively participated in the intervention out of the total who initially started (excluding those who sabotaged or refused to participate in the activity)	86.51	18.43	90
13	Indicate what percentage of the designed materials you have used in your classes	80.55	15.22	80

Figure 6: Distribution of answers to items 11, 12 and 13 of teachers' post-intervention questionnaire



The last item on the teachers' questionnaire asked them to add any other comments on the intervention that had not been covered in the questions. There were 21 responses (out of 47 participants) from which, through inductive content analysis, different themes were identified, which are particularly relevant in terms of learnings from this project for future interventions:

- Student participation and motivation: Teachers pointed out that, despite the concern with maintaining student motivation, this decreased over time, after a very high initial point. "At the beginning, they were very motivated and curious, but as the sessions progressed, they started to lose initiative", "At the beginning, all the students were very interested, but by the end, some had lost interest", "Towards the end, students seemed a bit tired and the last sessions didn't reinforce the work".

- Curriculum alignment and flexibility: Teachers were concerned about the alignment of ED with the existing curriculum, covering the necessary content and adapting it to the students' learning levels. "It would have been helpful to have more activities (both reinforcement and extension) to better adapt to the range of learning levels and paces within the same class", "I didn't focus on aligning

these sessions with the official standards, although they were closely related to the maths curriculum and served to support and aid more traditional teaching".

- Clarity of materials: Some responses highlighted the need for clearer and more flexible teaching materials to help students understand the ED methodology. "The student worksheets need improvement and should be made clearer", "We designed customised materials for the intervention", "I also prepared extra materials for working with integers and the transition from fractions to decimals to better align with […] the curriculum".

- Implementation challenges: Some teachers experienced logistical difficulties. "The number of sessions was too high", "To teach an ED class and make it hands-on, it would have been helpful to have two teachers".

- Project continuation and extension: Teachers suggested the need to continue the project and also suggested possible extensions. "It would be beneficial to explore ways of continuing the project. Also, any guidance on adapting the project to the curriculum would be appreciated", "Wouldn't it be better to start this programme in primary school?", "More examples that cover all possible cases are needed, keeping in mind the objectives pursued".

- Didactic benefits and insights: Some teachers highlighted the pedagogical insights gained, such as the differentiation between algorithms and mathematical reasoning and the promotion of critical thinking. "Personally, it has helped me a lot to differentiate the algorithm that students learn from the mathematical reasoning behind each operation", "In year 1 of secondary school, using dots and anti-dots to teach the addition of whole numbers has been a fantastic discovery", "I also developed critical thinking, as students continuously questioned why they needed to know ED when they couldn't see its application".

Compliance

Table 33 shows the descriptive parameters of all the components of the compliance index J, which was constructed as $J = 0.35 \cdot (a + b) + 0.1 \cdot (c + d + e)$, where:

a. Average percentage of attendance of teachers at the training sessions (obtained from school records).

b. Average score for the work plan developed by the teachers (obtained from the score awarded by the intervention team, on a 0-100 scale). The intervention team reviewed, gave feedback and rated the final plans, including use of materials.

c. Average score of teachers' commitment to the intervention plan. The way to measure the degree of commitment changed from the initial plan. Ultimately, it was analysed through follow-ups in a weekly newsletter and an online meeting, with records kept of both communication channels, including 1530 email exchanges throughout the intervention and a value on a 0-100 scale assigned.

d. Average percentage of students at the school who actively participated in the intervention. That is, excluding students who sabotaged the activity (obtained through a questionnaire administered to teachers after the intervention).

e. Average percentage of use of materials developed by the school's teachers (obtained through a questionnaire administered after the intervention).

The table below shows that all the schools reached the minimum required value (80) for the J index.

Table 34: Descriptive statistics of index J and its components

	a	b	c	d	e	J
Average	84.68	96.45	93.87	88.06	8516	90.10
SD	20.75	3.64	9.39	18.45	13.17	6.34
Min	50	90	70	30	50	80.00
P25	75	95	90	90	80	83.13
Median	100	95	100	95	90	92.25
P75	100	100	100	100	95	95.50
Max	100	100	100	100	100	98.25

Fidelity

All information gathered directly and indirectly in the implementation and process evaluation indicated that, unanimously, the intervention was implemented exactly as planned in all the schools. They followed the schedule and methodological approach and the materials were used with a high degree of fidelity, with any adaptations to the school context closely monitored by the intervention team.

Usual Practice

As the teachers in the control schools did not receive training on ED until after the intervention ended, it is unlikely that the intervention influenced the usual practices in these schools. No other concurrent interventions were conducted in the schools during the trial. The usual practice in the intervention schools varied. According to information obtained from the initial questionnaire (IPE section, Pre-intervention subsection) and from the interviews (IPE section, During Intervention subsection), we can determine that the impact of the intervention was based more on contents than on methods. It is important to consider that, as this was a voluntary programme, the vast majority of participating teachers already demonstrated a high baseline level of motivation, often linked to an interest in educational innovation. Consequently, group work and discovery-based methodologies proposed for ED were not particularly novel.

Conclusions

Below are the main findings of the impact evaluation.

Table 37: Key conclusions
Students in the intervention group performed moderately better in the posttest on computational thinking skills than those in the control group. However, the result is not statistically significant, indicating that the observed improvement could be due to chance. Both groups improved their performance in computational skills when comparing their pre- and posttest scores.
The attitudes towards mathematics of students in the intervention group were moderately better than those in the control group in the posttests. However, the result is not statistically significant, indicating that the observed improvement could be due to chance. Both groups worsened in attitudes towards mathematics when comparing their pre- and posttests.

During the observation phases and interviews, most students expressed a positive sentiment towards the ED intervention, emphasising that they found it entertaining, interesting and, in general, liked it more than the rest of their maths classes. However, some students reported feelings of boredom and fatigue at the end of the intervention.

In the observation phases and interviews held during the intervention, most teachers pointed out that the use of ED made their classes more inclusive, motivating students who normally did not dare to participate in maths class. They also pointed out that ED can be easily linked with the maths curriculum and suggested it would be more effective if incorporated into a more integrated year-long plan.

The fact that the ED intervention was carried out as an hour separate from the rest of the mathematics classes (which followed the normal curriculum) hindered it from being considered part of the subject, causing a positive attitude among the students towards the experience (as shown by the qualitative information) but not improving their overall attitude towards mathematics.

For the intervention to positively influence attitudes towards mathematics, ED should be completely integrated into the subject of mathematics.

Impact Evaluation and Integration

Evidence to support the logic model

Evidence obtained from the project largely supports the original logic model (see Figure 1). Questionnaires, interviews and observations demonstrate that the ED intervention enabled teachers to master a new tool for teaching maths, expanding both their mathematical and didactic knowledge. To varying degrees, teachers adapted the ED materials to fit their educational contexts. The results also indicate that some students acquired new resources for mathematical work, enhancing their number sense and making high-level connections with other mathematical topics. However, this impact was not universal, as others experienced demotivation, likely due to boredom and fatigue. Another logical implication in the original model concerned the relationship between the ED intervention and the identification of CT dimensions, though the results did not support this connection. The absence of a significant impact of the intervention on students' attitudes towards mathematics, combined with the mostly positive evidence from interviews and questionnaires, shows that there is little evidence of the input, established in the logic model, of motivation and follow-up by teachers leading to an increase in students' tolerance for errors. In contrast, questionnaires and interviews with teachers endorsed the logic model outcomes regarding the impact of their professional development.

The impact on CT and students' attitudes needs further research. To start with, researchers must discuss how to develop students' CT skills, however results from this project illustrate that we may need to determine beforehand the type of CT dimensions to be developed and how they are connected with mathematical thinking dimensions. The ED intervention only slightly contributed to a better performance in computational thinking skills, leaving programming out of the assessed dimensions. Therefore, the results indicate the need to further explore the nature and relationship of CT to mathematics. In terms of students' attitudes, the responses about boredom and not perceiving real progress in the activities, as well as students' conceptions of what is (and what is not) mathematics, suggest the need for further research on motivation and monitoring of the activity to engage all students: the novelty of the ED activity alone was insufficient and may even have represented an obstacle, clashing with a more traditional conception of maths.

Interpretation

The findings did not show a statistically significant impact of the ED-based intervention on students' computational thinking skills or attitudes towards mathematics. The results highlight the complexity of establishing a direct link between ED and the CT skills assessed in the instrument. Notably, socioeconomic status and prior experience in computer science or robotics, influenced computational thinking scores, as did gender, favouring girls. This gender association is especially relevant, as literature often indicates a male bias in computational thinking assessments. However, qualitative evidence provides insight into various aspects. For teachers, the intervention was a catalyst for their professional development, empowering them with a new mathematical tool that enabled intra- and extra-mathematical connections. ED was also perceived by teachers as a powerful tool to help students with learning or motivational difficulties. For students, the intervention split them into two unequal groups: most of the students liked ED and perceived it as easier than 'normal maths lessons' and it helped them to make mathematical connections, while a smaller group found ED boring or too far from what they conceive to be maths.

In terms of attitudes towards mathematics, quantitative and qualitative results were mixed. While observations and interviews suggest positive associations with the ED intervention, standardised questionnaires did not show an overall improvement in attitudes toward maths as a subject. The data also revealed an impact on fatigue, typical of maths courses. A more positive effect was found in the attitudes of boys and students from higher socioeconomic backgrounds. In addition, observations and interviews highlighted benefits for students with learning difficulties and those from challenging socioeconomic backgrounds, who were more engaged and motivated by ED activities. More able students, however, responded variably, depending on the teacher's ability to enrich the tasks. There is an evident effect of asking about maths and not specifically about ED in the instrument, with students perceiving the intervention as separate from maths lessons (due to the organisation of the intervention). These results suggest that a key factor for success is the integration of ED activities into the broader mathematics curriculum to avoid these tasks being perceived as separate from core maths learning.

Limitations and Lessons Learned

One limitation that needs to be outlined is that neither of the test instruments used (BBACT and IAM) made responding to all items mandatory. Therefore, we cannot conjecture about the percentage of missing values. We do not know if they were intentionally left blank (due to difficulty or other causes) or if it was an unintentional error. As noted in the Missing Data Analysis section, no pattern was observed. In addition, the BBACT scoring system could be a critical point: correct answers were awarded 1 point, while incorrect answers received 0 points. A variation could be introduced to assign different scores based on the difficulty of each question. The use of the same instrument in the pre- and posttest could also be viewed as a limitation. The reasons for ruling out a test-retest effect have already been discussed.

Socioeconomic information was relevant in the mixed models, but we did not have access to objective measures, only to subjective approaches made by the teachers. Nevertheless, the impact of this factor is consistent with relevant literature, even when based on teacher perception: belonging to a higher socioeconomic status favours better CT skills and more positive attitudes towards mathematics.

Even when the intervention focused on attention to diversity, this fact could not be controlled and depended on teachers' perceptions about the suitability of the adaptation. Some schools decided to exclude students with extreme special needs from the study, while others decided to include them. The total number of students with these characteristics who were excluded from the project was 26, distributed in different schools. Therefore, due to the small number, we do not believe that, in general terms, this had a significant effect on overall results. However, for future interventions or projects, this aspect should be addressed more specifically.

Another problem that could limit the study is the consideration of a separate environment for the intervention. As mentioned, the intervention consisted of one hour per week, within maths classes and it was mainly conceived as a stand-alone intervention (obviously linked to the curriculum, but not completely integrated). This design ensured control of the intervention, the pace of its implementation and the measurement of the number of hours and use of materials. However, it is possible that it generated a perception among students that 'it was not a maths lesson'. This perception may have influenced the results of the IAM instrument. However, in the pretest, the students scored quite high on the IAM instrument, also making it harder to improve their positive attitudes towards maths. The role of intrinsic motivation, perceived usefulness and perceived competence turned out to be very relevant: many students enjoyed the programme because they understood it and knew how to use ED. Furthermore, less positive attitudes seemed to be linked to the automation process of ED: some students began to perceive it as a mechanical procedure and got bored. Therefore, another learning from the project is the need to design and implement challenging tasks for all sessions, allowing some differentiation for students who learn faster. Teachers also expressed the need to address the diverse learning needs of students (not necessarily those with special educational needs), expressing a desire for more resources tailored to different levels and paces of student progress.

The intervention design could also influence the change in students' attitudes towards mathematics, as this was a dispersed intervention (one hour per week) during a limited time (total 17 weeks). This pace could be making an impact on attitudes harder to achieve. There is one growing hypothesis as to whether a more integrated ED intervention would have a greater impact on attitudes, as literature shows more examples of significant changes when interventions are more concentrated and intensive.

The teacher's role seemed to be crucial, according to observations and responses from both teachers and students. Of course, this is not new, the same occurs in any school intervention. The teachers were extremely engaged with the project (as revealed by the newsletter, Google Classroom and observations) and highly motivated, volunteering to participate in a demanding programme. They also fully adhered to the intervention plan. They pointed out the need for strategies to maintain the students' enthusiasm, requiring additional strategies or adjustments in pacing. Another lesson learnt is to align ED activities with the established curriculum (not only during the intervention year but also before and after) to achieve a better integration with the curriculum.

Regarding the generalisability, as highlighted in the summary, given that the sampling was not random, the results are not generalisable in a strictly inferential sense. However, the results can be considered quasi-representative of the analysed context, given the composition and size of the sample. In other words, the results provide a good approximation to similar contexts, but cannot be considered generalisable.

Future Research and Publications

The Exploding Dots intervention had no statistically significant impact, only a slight one, on improved performance in students' CT skills. The instrument to measure these skills excluded programming from its dimensions and the results indicate that even when there is an obvious relationship between ED and abstraction, explicitly working on ED did not have a statistically significant effect. Therefore, more research is needed on the conceptualisation of CT: its complex nature, the description of its dimensions, how these dimensions are related and what activities help foster it inside and outside the maths classroom.

The results show that working explicitly on ED did not have a statistically significant impact, only a slight one, on the CT skills of students assessed. As mentioned in the introduction, there is no consensus on the definition of computational thinking, which is further complicated by the diversity of components, dimensions, aspects and skills identified in scientific literature. The results of this study, while still within a general framework, highlight the need for more research on these areas and their relationships within the field of CT.

Some data from the project will be considered for future publications: studies on the validation of the BBACT tool, qualitative analysis of student and teacher interviews and school-specific reports. Furthermore, the conjectures that emerged from the project provide very interesting areas of research for the future: the dimensions of CT and how they are related, the definition of instruments for measuring CT, the role of student diversity in this type of study, how conceptions about mathematics and its learning can be an obstacle for exploring new mathematics, the use of ED in primary education as a parallel support for developing number systems and in secondary education for enhancing the learning of polynomials, etc.

References

Adamuz-Povedano, N., Fernández-Ahumada, E., García-Pérez, M. T., & Montejo-Gámez, J. (2021). Developing Number Sense: An approach to Initiate Algebraic Thinking in Primary Education. *Mathematics*, *9*(5). **https://doi.org/10.3390/math9050518**

Angeli, C., Voogt, J., Fluck, A., Webb, M., Cox, M., Malyn-Smith, J., & Zagami, J. (2016). A K-6 Computational Thinking Curriculum Framework: Implications for Teacher Knowledge. *Journal of Educational Technology and Society*, *19*(3), 47-57.

Barr, V., & Stephenson, C. (2011). Bringing Computational Thinking to K-12: What is Involved and What is the Role of the Computer Science Education Community? *ACM Inroads*, *2*(1), 48-54. **https://doi.org/10.1145/1929887.1929905**

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. **https://doi.org/10.18637/jss.v067.i01**

Bråting, K., & Kilhamn, C. (2021). Exploring the Intersection of Algebraic and Computational Thinking. *Mathematical Thinking and Learning*, *23*(2), 170-185. **https://doi.org/10.1080/10986065.2020.1779012**

Brennan, K., & Resnick, M. (2012). New Frameworks for Studying and Assessing the Development of Computational Thinking. *Proceedings of the Annual Meeting of the American Educational Research Association*, p. 1-25.

Caswell, C. J., & LaBrie, D. J. (2017). Inquiry Based Learning from the Learner's Point of View: A Teacher Candidate's Success Story. *Journal of Humanistic Mathematics*, *7*(2), 161-186. **https://doi.org/10.5642/jhummath.201702.08**

Cuny, J., Snyder, L., & Wing, J. M. (2010). Demystifying Computational Thinking for Non-Computer Scientists [Unpublished paper].

Dagiene, V., & Dolgopolovas, V. (2022). Short Tasks for Scaffolding Computational Thinking by the Global Bebras Challenge. *Mathematics*, *10*(17). **https://doi.org/10.3390/math10173194**

Drijvers, P. (2013). Digital Technology in Mathematics Education: Why it Works (or doesn't). *PNA*, *8*(1), 1-20. **https://doi.org/10.30827/pna.v8i1.6120**

Fernández Alonso, R. (2005). *Evaluación del rendimiento matemático* [Unpublished doctoral dissertation]. Universidad de Oviedo.

Freudenthal, H. (1991). *Revisiting mathematics education*. Kluwer.

García Fernández, T., Kroesbergen, E., Rodríguez Pérez, C., González-Castro, P., & González-Pienda, J. A. (2015). Factors Involved in Making Post-performance Judgments in Mathematics Problem-Solving. *Psicothema*, *27*(4), 374-380. **https://doi.org/10.7334/psicothema2015.25**

García, T., Rodríguez, C., Betts, L., Areces, D., & González-Castro, P. (2016). How Affective-Motivational Variables and Approaches to Learning Predict Mathematics Achievement in Upper Elementary Levels. *Learning and Individual Differences*, *49*, 25-31. **https://doi.org/10.1016/j.lindif.2016.05.021**

Glutting, J. (2002). Some Psychometric Properties of a System to Measure ADHD Among College Students': Factor Pattern, Reliability, and One-Year Predictive Validity. *Measurement and Evaluation in Counseling and Development*, *34*, 194-209.

Grover, S., Fisler, K., Lee, I., & Yadav, A. (2020). Integrating Computing and Computational Thinking into K-12 STEM Learning. In *SIGCSE '20: Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 481-482). Association for Computing Machinery. **https://doi.org/10.1145/3328778.3366970**

Hattie, J. A. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, *9*(2), 139-164. **https://doi.org/10.1177/014662168500900204**

Izu, C., Mirolo, C., Settle, A., Mannila, L., & Stupuriene, G. (2017). Exploring Bebras Tasks Content and Performance: A Multinational Study. *Informatics in Education*, *16*(1), 39-59. **https://doi.org/10.15388/infedu.2017.03**

Lazonder, A. W., & Harmsen, R. (2016). Meta-Analysis of Inquiry-Based Learning: Effects of Guidance. *Review of Educational Research*, *86*(3), 681-718. **https://doi.org/10.3102/0034654315627366**

Lockwood, J., & Mooney, A. (2018). Developing a Computational Thinking Test using Bebras Problems. In Piotrkowicz, A. Dent-Spargo, R., Dennerlein, S., Koren, I., Antoniou, P., Bailey, P., Treasure-Jones, T., Fronza, I., Pahl, C. (Eds.), ***Joint Proceedings of the CC-TEL 2018 and TACKLE 2018 Workshops***.

Palop, B., Díaz, I., Rodríguez-Muñiz, L. J., y Santaengracia, J. J. (2025). Redefining Computational Thinking: A Holistic Framework and its implications for K-12 Education. *Education and Information Technologies*. **https://doi.org/10.1007/s10639-024-13297-4**

Papert, S. (1980). Personal Computing and Its Impact on Education. In R. Taylor (Ed.), *The Computer in the School: Tutor, Tool, Tutee*, (pp. 197-202). Teachers College Press.

Pembury Smith, M. Q. R., & Ruxton, G. D. (2020). Effective Use of the McNemar Test. *Behavioral Ecology and Sociobiology*, *74*. **https://doi.org/10.1007/s00265-020-02916-y**

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics*, *4*(3), 207-230. **https://doi.org/10.3102/10769986004003207**

R Core Team. (2024). ***R: A Language and Environment for Statistical Computing***. R Foundation for Statistical Computing.

Román-González, M., Pérez-González, J. C., & Jiménez-Fernández, C. (2017). Which Cognitive Abilities Underlie Computational Thinking? Criterion Validity of the Computational Thinking Test. *Computers in Human Behavior*, *72*, 678-691. **https://doi.org/10.1016/j.chb.2016.08.047**

The Royal Society. (2012). *Shut down or restart? The way forward for computing in UK schools*.

Sarama, J., & Clements, D. H. (2009). "Concrete" Computer Manipulatives in Mathematics Education. *Child Development Perspectives*, *3*(3), 145-150. https://doi.org/10.1111/j.1750-8606.2009.00095.x

Santaengracia, J. J., Palop, B., García, T., Rodríguez Pérez, C., & Rodríguez-Muñiz, L. J. (2025). Bebras-Based Assessment for Computational Thinking: Performance and Gender Analysis. *Education Sciences*, *15*(7), 899. **https://doi.org/10.3390/educsci15070899**

Santaengracia, J. J., Rodríguez-Muñiz, L. J., García, T., Rodríguez Pérez, C., & Palop, B. (2024). *Performance and Gender Analysis of a Bebras-Based Assessment for CT* [Manuscript submitted for publication].

Schafer, J. L. (1999). Multiple Imputation: A Primer. *Statistical Methods in Medical Research*, *8*(1), 3-15. **https://doi.org/10.1177/096228029900800102**

Selby, C. C., & Woollard, J. (2014). Computational Thinking: The Developing Definition. In *SIGCSE '14: Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, p. 356481. Association for Computing Machinery. **https://doi.org/10.1145/2538862.2538900**

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying Computational Thinking. *Educational Research Review*, *22*, 142-158. **https://doi.org/10.1016/j.edurev.2017.09.003**

Taber, K. S. (2018). The Use of Cronbach's Alpha when Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*), 1273-1296. **https://doi.org/10.1007/s11165-016-9602-2**

Wiebe, E., London, J., Aksit, O., Mott, B. W., Boyer, K. E., & Lester, J. C. (2019). Development of a Lean Computational Thinking Abilities Assessment for Middle Grades Students. *SIGCSE '19: Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (pp. 456-461). Association for Computing Machinery. **https://doi.org/10.1145/3287324.3287390**

Wing, J. M. (2006). Computational Thinking. *Communications of the ACM*, *49*(3), 33-35. **https://doi.org/10.1145/1118178.1118215**

Wing, J. M. (2008). Computational Thinking and Thinking About Computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*, 3717-3725. **https://doi.org/10.1098/rsta.2008.0118**

Zafra-Gómez, J. L., Román-Martínez, I., & Gómez-Miranda, M. E. (2014). Measuring the impact of inquiry-based learning on outcomes and student satisfaction. *Assessment & Evaluation in Higher Education*, 40(8), 1050-1069. **https://doi.org/10.1080/02602938.2014.963836**

Appendix A: Security Classification of Trial Findings

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	MDES	Attrition					
5	Randomised design	< = 0.2	0-10 %	5				5
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20 %					
3	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30 %			Adjustment for threats to internal validity 0		
2	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40 %					
1	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50 %					
0	No comparator	> = 0.6	> 50 %					

Threats to validity	Risk rating	Comments
Threat 1: Confounding	Low	The report describes a properly stratified cluster randomised design and randomisation was preserved.
Threat 2: Concurrent Interventions	Low	There is no indication that the control schools were exposed to similar computational thinking interventions during the study. Control schools continued business as usual. To incentivise participation, control schools were offered the intervention after the end of the project.

Threat 3: Experimental Effects	Low	No evidence of novelty or Hawthorne effects influencing outcomes. Although student enthusiasm for ED was observed, this did not translate into a measurable academic impact.
Threat 4: Implementation Fidelity	Low	Implementation fidelity was high, with strong adherence to plans (92 % commitment, 81 % use of materials) and consistent monitoring.
Threat 5: Missing Data	Low	Although attrition was low (~ 2%), the report acknowledges outcome-level missing data > 11 %. This was due to a mix of causes and checks suggest data is Missing at Random (MAR). Multiple imputation was employed.
Threat 6: Measurement of Outcomes	Moderate	The primary outcome instrument (BBACT) was updated and piloted to better capture computational thinking skills. The instrument had strong content relevance. However, its psychometric properties were mixedThe risk is moderate because the measurement error could affect the precision of the estimated impact..
Threat 7: Selective Reporting	Low	The trial was registered and analysis follows the pre-published protocol with minor deviations.

Appendix B: Changes Since the Previous Evaluation

	Feature	Efficacy to effectiveness stage
Evaluation	Randomisation	There was a typo in the version of Table 7 published in the SAP, which does not affect the main measures. The correct digits are those in this document.
	IPE	In the IPE methods and instruments (Table 5) in the Protocol some minor changes were introduced, resulting from the inapplicability of the planned instruments, the pursuit of a lighter evaluation schedule for participants and improvements in the planned analysis methods. Specifically, they affected the analysis method for assessing RQ4, the methodology for RQ 11 and 12, the instruments for RQ 13 and 14, the initially planned post-intervention questionnaire for the school directors (cancelled) and the planned during-intervention questionnaires for students and teachers (included in the observations due to the large number of visited schools).
	Outcomes and baseline	The original name for the instrument for measuring CT skills (MDCT) was changed to BBACT, as this was considered a better description of the instrument's features.

Appendix C: BBACT Instrument

English version with images. One question per page. Original version at: **https://hdl.handle.net/10651/75100**

Question 1: Rings

Sara is playing ring toss, and when a ring goes onto the post, she scores the following points:

On the first toss: 5 points

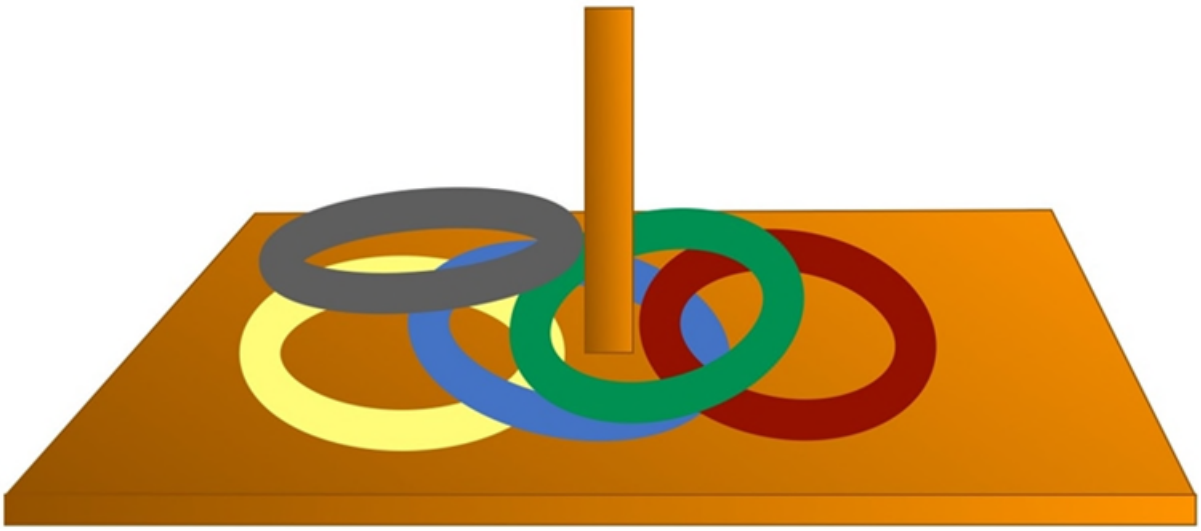
On the second toss: 4 points

On the third toss: 3 points

On the fourth toss: 2 points

On the fifth toss: 1 point

How many points did she score in this round?

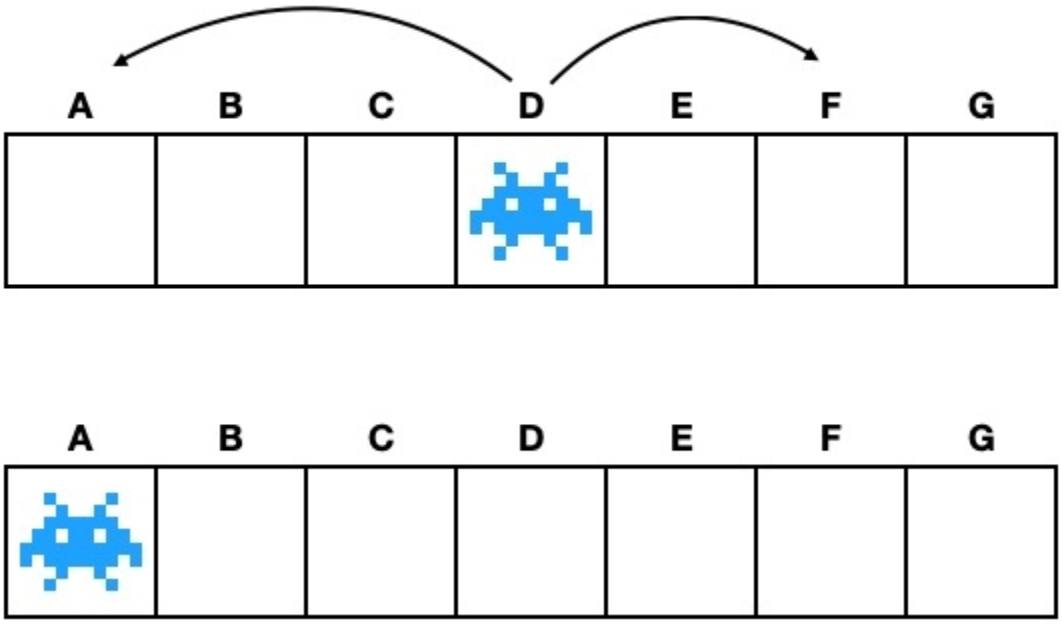


- **15 points**
- **2 points**
- **6 points**
- **5 points**

Question 2: Jumps

In this video game, there are two buttons. Pressing "right" moves the character two squares to the right. Pressing "left" moves the character three squares to the left.

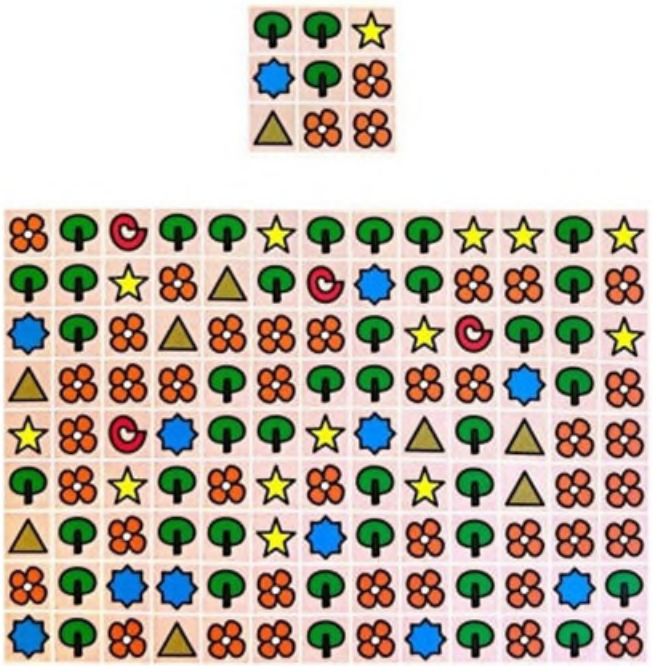
If the character starts at position A and the button is pressed three times, **in which square or squares could it end up?**



- **Square E**
- **Square G**
- **Squares C and H**
- **Squares G and B**

Question 3: Frieze

How many times does the pattern above appear in the figure below?



- **1**
- **2**
- **3**
- **4**

Question 4: Nim

For each turn of this game, you can:

Remove 1 or 2 black stones

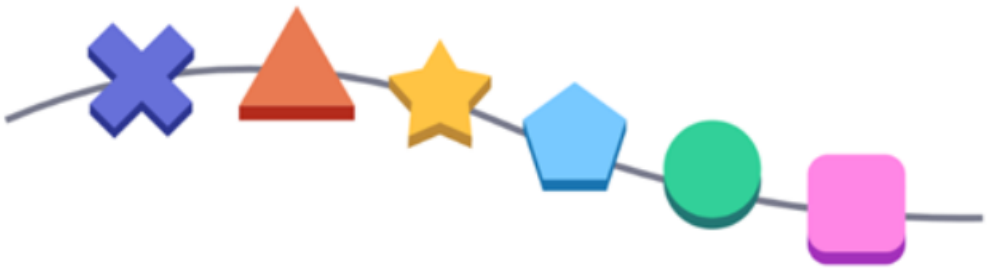
Remove 1, 2, or 3 white stones.

Two people play, and the person who removes the last stone of either colour wins the game. If it's your turn to play, **which move do you win with?**

- **1 black stone**
- **1 white stone**
- **2 white stones**
- **3 white stones**

Question 5: Bracelet

Which of the following four images shows what this bracelet looked like before it broke?



- 
- 
- 
- 

Question 6: Footprints

Four footprints have been found. The police are looking for a thief who was wearing shoes with striped soles and a thin heel.

Which of the following footprints belongs to the thief?

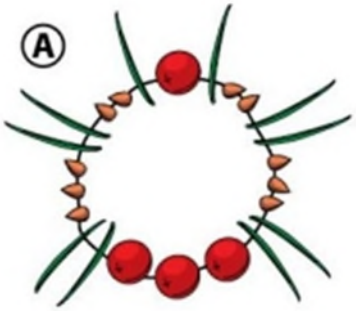
- 
- 
- 
- 

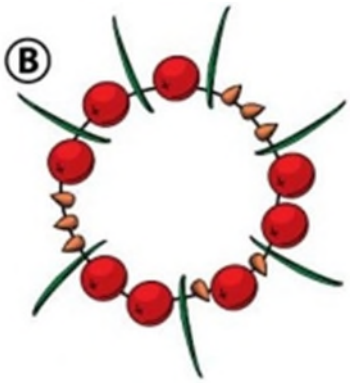
Question 7: Necklace

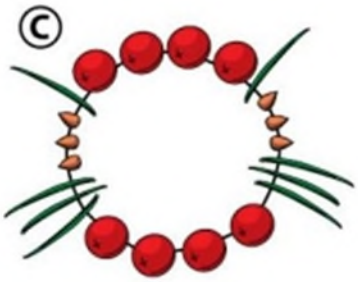
A necklace is to be made with red fruits, pine needles and brown seeds, meeting the following conditions:

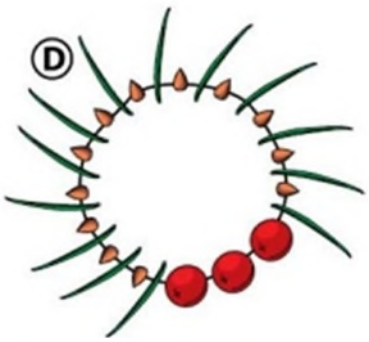
- Each group of red fruits must have a pine needle on either side.
- The number of brown seeds must be equal to the number of pine needles.

Which of the following necklaces meets these conditions?

- 

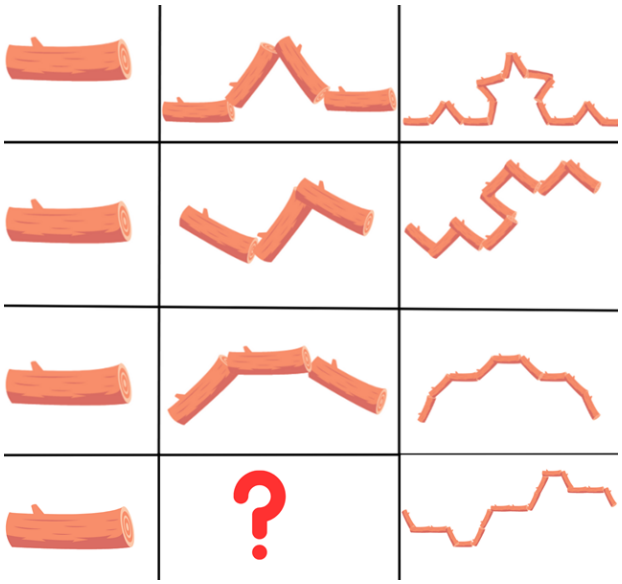
- 

- 

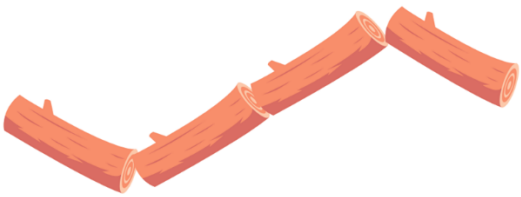
- 

Question 8: Logs

Beavers use logs to form structures that follow patterns. This table shows some of these patterns.

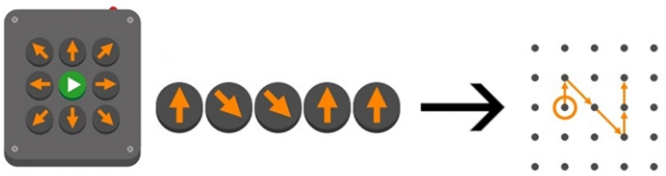


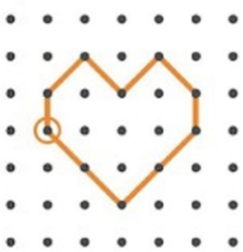
Which of the following options is the missing one?

- 
- 
- 
- 

Question 9: Robot heart

Emma is playing with a robot that draws lines between dots, as seen in the example.





Given that it starts from the circled dot, **which of the following button combinations draws the heart shape?**

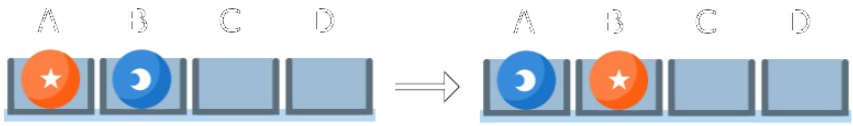
- 
- 
- 
- 

<

Question 10: Ball swap

We have 4 trays: A, B, C and D. In tray A there is an orange ball; in tray B, a blue ball; and trays C and D are empty.

If we want to swap the balls in A and B, **which of the following sequences is incorrect?**



- **1**

Step 1: pick the ball in B and place it in C. Step 2: pick the ball in A and place it in B. Step 3: pick the ball in C and place it in A.

- **2**

Step 1: pick the ball in A and place it in D. Step 2: pick the ball in B and place it in A. Step 3: pick the ball in D and place it in B.

- **3**

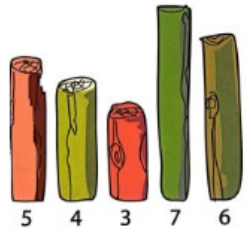
Step 1: pick the ball in B and place it in C. Step 2: pick the ball in A and place it in D. Step 3: pick the ball in C and place it in A.

- **4**

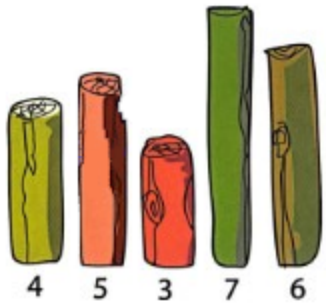
Step 1: pick the ball in A and place it in C. Step 2: pick the ball in B and place it in A. Step 3: pick the ball in C and place it in B.

Question 11: Log order

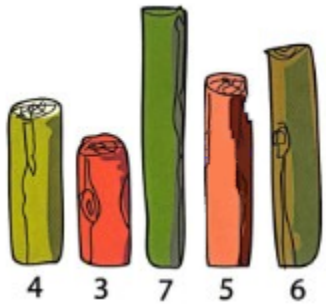
You have logs of different heights. You start from the leftmost log. If it is taller than the log on its right, you swap them. Then you compare the logs that are now in the second and third position and repeat the process until you get to the last pair.



From the initial arrangement of the logs in the image, **which of these situations is impossible to reach?**

- 

- 

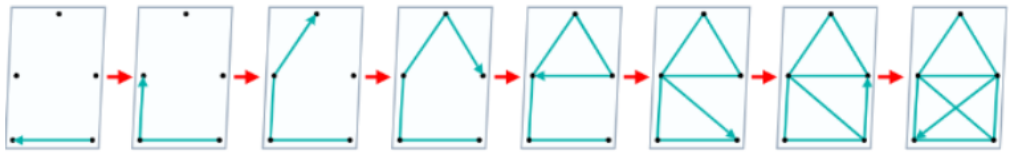
- 

- 

Question 12: Graphs

We were able to draw this figure without lifting the pencil from the paper or passing over the same line twice (you can pass through the same point twice).

Which of the other four drawings can be done in the same way?



- 
- 
- 
- 

Question 13: Flip-flop

A flip-flop is a component that has two possible states. Each time a ball passes through a flip-flop, its state changes, as shown in the animation.

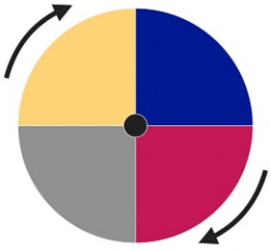
https://youtu.be/y0k3NQltUXQ

If three balls are thrown, in which tube will the third (yellow) ball fall?

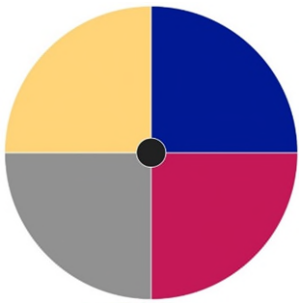
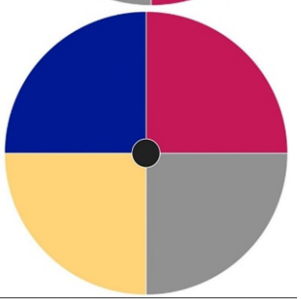
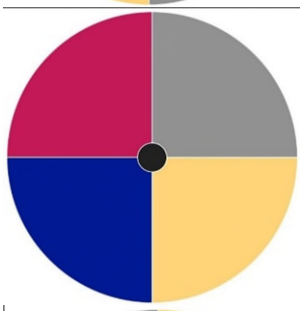
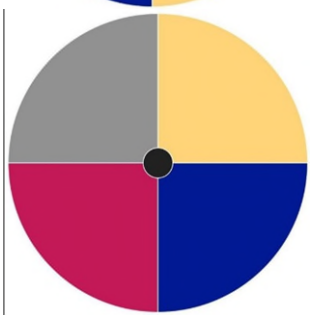
- **1**
- **2**
- **3**
- **4**

Question 14: Roulette

We have a roulette that turns 90 degrees to the right (clockwise) each time a button is pressed.



If we press the button seven times from this position, **what will its final position be?**

- 
- 
- 
- 

Question 15: Strip

This strip followed a pattern of three colours, but we have cut off a piece.

Which of the following can be the length of the piece we cut?

- **3 or 7 segments**
- **3 or 4 segments**
- **Only 3 segments**
- **Only 4 segments**

Question 16: Houses

On this street, there are five houses on each side. To paint the houses, the following rules must be followed:

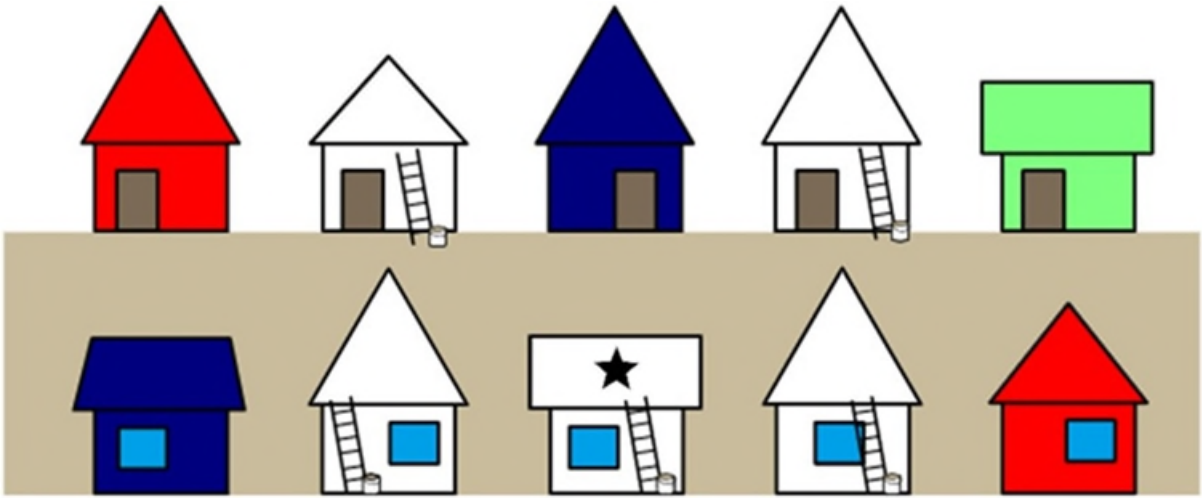
- All houses must be painted red, green or blue.

- A house cannot be the same colour as the house to its left and right.

- Two houses directly opposite each other cannot be the same colour.

As seen in the image, some of the houses have already been painted.

What colour will the house with a star on the roof be painted?

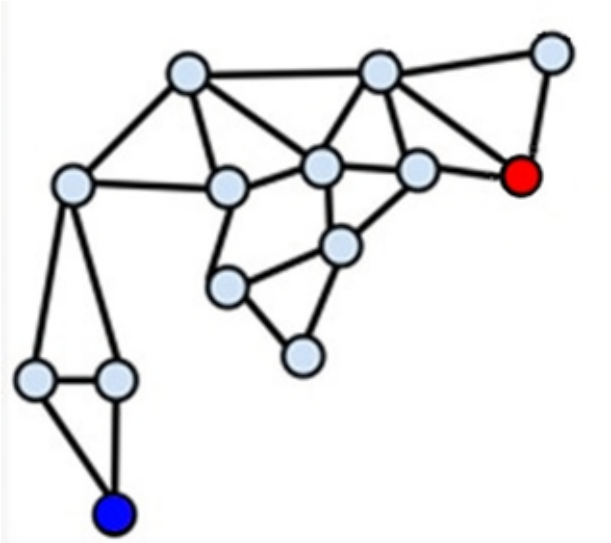


- **Red**
- **Green**
- **Blue**
- **It cannot be determined**

Question 17: Watchtower

In the image, each circle represents a watchtower. One minute after a tower is lit, the towers connected to it by a line also light up.

If we light the tower marked with a red dot, **how long does it take for the tower with a dark blue dot to light up?**



- **4 minutes**
- **5 minutes**
- **6 minutes**
- **7 minutes**

Appendix D: IAM Instrument

English version of the selected items from the original IAM instrument used in the project.

Note: items 3, 4, 7, 8, 9, 10, 12, 13, 15, 17, 23, 25, 26, 27, 28 and 31 were inverted for the analysis.

IAM

Indicate your degree of agreement or disagreement with the following statements on a 5-point Likert scale (*Strongly disagree, Disagree, Slightly agree/Not sure, Agree, Strongly agree*):

1. My parents think I need maths for whatever I want to do in the future.
2. I think I will need mathematics in my future work.
3. Maths teachers have made me feel incapable of mathematics.
4. My teachers encourage me to focus my future studies on mathematics.
5. I think I could master even the most difficult mathematics.
6. I have great confidence in myself when doing mathematical tasks.
7. For some reason, even though I study, I find mathematics extraordinarily difficult.
8. I really don't think I'm good at mathematics.
9. Maths will not be important in my future working life.
10. Maths is a waste of time.
11. If it were up to me, mathematics would cease to exist.
12. Maths is of no interest to me.
13. I see mathematics as a subject that I will rarely use in my adult life.
14. When I leave maths class, I keep thinking about the things I don't quite understand.
15. Solving mathematical problems holds little attraction for me.
16. I am sure I can learn mathematics.
17. Mathematics is a valuable and necessary subject.
18. I like mathematical puzzles.
19. When I come across a maths problem that I can't solve immediately, I keep working on it until I solve it.
20. How much respect others have for you largely depends on the marks you get in maths.
21. My parents believe that mathematics is one of the most important subjects.
22. I would very much like to be one of the best in mathematics.
23. If I got the highest marks in maths, I would rather no one knew about it.
24. I have hardly ever felt nervous about a maths exam.
25. When I work on mathematics, I go blank and am unable to think clearly.
26. Maths exams scare me.
27. I despair about maths, my grades don't change, no matter what I do.
28. Maths usually makes me feel uncomfortable and nervous.
29. I am or would be happy to get the highest marks in mathematics.
30. In maths classes, I feel very good and I am happy.
31. When I can't solve a problem in mathematics, I get angry and furious.
32. Knowing maths will help me earn a living.