How is two better than one? An observational study on the impact of working in pairs when solving Bebras tasks

Carlo Bellettini[1 [0000−0001−8526−4790]], Violetta Lonati[1 [0000−0002−4722−244X]], Mattia Monga[1 [0000−0003−4852−0067]], and Anna Morpurgo[1 [0000−0003−0081−914X]]

Università degli Studi di Milano, Milan, Italy
{bellettini,lonati,monga,morpurgo}@di.unimi.it

Abstract. Every year the Bebras challenge proposes small tasks to students, based on CS concepts. In Italy, in 2021 for the first time, it was possible to choose whether to participate in the challenge individually or in teams of two students. The team size was expected to affect the performance of students; in particular working in pairs was expected to increase the probability of solving the tasks correctly. We carried out an observational study on the results of the 2021 Bebras challenge in Italy, aiming at investigating and measuring the effects of team size on the performance. The findings confirm that working in pairs generally improves the team performance, but the impact is much smaller than expected. We observed that the positive effect of collaboration is greater with younger pupils and somewhat decreases when age increases. We identified and discuss the features of tasks where the impact was more relevant, and where this trend was more evident. We also propose some hypotheses, to analyze in future qualitative studies, to interpret the results.

Keywords: K12 · Bebras challenge · observational studies.

1 Introduction

In cognitive theory, many studies suggest that collaboration between peers enhances learning. This seems particularly true in STEM (Science, Technology, Engineering, and Math, including Computer Science[1]) education, where several popular methodologies, *e.g.,* collaborative and problem-based learning, are in fact based on this assumption[7]. Moreover, *pair programming* is a common practice in the so called *agile* approaches to software engineering and is often also adopted in many educational contexts [9]. Faced with a problem, and working in a small group to solve it, pupils can explore the problem and its features, and thus devise, analyse and contrast solving strategies, in a process of collaborative knowledge building. Collaborative learning is also said to increase motivation and engagement [4].

[1] See for example `https://www.ed.gov/stem`.

For these reasons, since its first edition in our country, the participation to the Bebras Challenge was organized around teams. The Bebras International Challenge on Informatics and Computational Thinking[2] is a yearly contest organized in several countries since 2004 [1,3], with almost three million participants worldwide. The contest, open to pupils of all school levels (from primary up to upper secondary), is based on tasks rooted on core informatics concepts, yet independent of specific previous knowledge such as for instance that acquired during curricular activities. According to the (informal) feedback we received from many teachers, the Bebras challenge is able to engage pupils — even those who show less motivation in usual school activities — and to often activate lively discussions and interesting exchanges within the groups.

In 2021, due to the pandemic and to adhere to the social distancing rules, the participation in teams could have been problematic. Hence, we allowed participation in "teams" of individuals ("*singles*") or teams of pairs ("*doubles*"), to meet different schools' needs and organizational constraints. We analyzed the results of overall 19'490 teams, 11'055 singles, 8'435 doubles, who participated over 5 different age categories (for a total of 27'925 students). All teams in the same age category were asked to answer the same suite of questions, regardless of their team composition, but their results were ranked distinctly. Together with the submitted answers, the Bebras platform[12] we use collects data concerning the interactions of teams with the platform itself (how much time pupils spend on each specific task, whether and when they go back and review/change their answer to an already completed task, whether they perform actions that generate feedback from the system, and so on). This offered us the chance to conduct an observational study about the effects of team size on the performance. Our main research question is:

RQ - How does the team size affect the performance of Bebras solvers?

Our initial hypothesis was that teams formed by two pupils would perform *better* than the individuals. The research question can then be articulated in two further sub-questions:

RQ1 - For which categories of pupils does working in pairs have the most positive impact?

RQ2 - For which kinds of tasks does working in pairs have the most positive impact?

Our findings confirm the initial hypothesis, but show that the effect of team composition over performance is in general less than expected. Moreover we observe that such effect occurs differently according to the age of pupils and the features of tasks. We discuss these differences and state some hypotheses that may explain them. Such hypotheses are to be explored further in a future in-depth qualitative study.

The paper is organized as follows. In Section 2 we present the collected data and the methods we used to analyze them. In Section 3 we present our findings:

[2] See `http://bebras.org`.

in Section 3.1 we compare the performances of singles versus doubles, and analyze the role of age categories on the differences between such performances; in Section 3.2 we show which tasks benefit most from collaboration and detect relevant features of these tasks. In Section 4 we acknowledge the limitations of our study. After discussing some related works in Section 5, conclusions are drawn in Section 6.

2 Methodology

This is an *observational study*. This means that the data we analyzed were not purposely gathered with a designed experiment, instead they were collected during the Bebras challenge held in November 2021. We first describe the data set and then present the methods we used for the analysis.

2.1 Dataset

The data were collected in order to manage the participation of schools, administer the contest, monitor and possibly fix issues arising during the challenge (*e.g.,* malfunctioning, cheating, loss of data), and to perform statistical analyses.

Schools participating in the challenge were informed that students' data were collected and they consented to their use for research and statistical presentation of the results. In fact no national ranking is ever published, only aggregated data (but the teachers can see the performances of all the teams of their school and the ranking within an institution). All data analyzed were anonymized by deleting most of the personal identifying data: we retained only the regional provenance of teams in order to analyze their geographical distribution (we cover all the administrative regions of our school system).

The dataset contains information about the performance of each team in the contest. Each team belongs to one category (among five) according to their components' age. Teams can have different sizes, *i.e.,* there are teams formed by a pair of students ("doubles") or just a single individual ("singles"). The number of teams considered in our analysis are reported in Table 1, grouped by category and team size. All teams in the same age category were asked to solve the same suite of 12 tasks, independently of the team's size. Some tasks appeared in more than one category. For each team, we know for which of the tasks assigned to their category they answered correctly and for which not. Table 1 presents also the average ratio of correct answers to tasks in each category. Finally, we have data concerning how the teams interacted with the contest platform while solving the task; the kind of data we can collect are described in [12]. All the anonymous data we analyzed are available at `https://doi.org/10.13130/RD_UNIMI/WT9NHU` for independent studies and cross-validation.

2.2 Analysis methods

We considered each task solution as a random event with a binary outcome: solved or not solved. To simplify the problem, we considered each task as an

Table 1. Number of participants and success ratio (average over all tasks) for each category and team size. The last column reports the increment in the average success ratio obtained with doubles compared to singles.

category	*n. of teams*	*success ratio*	*Δ doubles − singles*
IV-V grade	2740	45.4%	8.6%
singles	1092	40.2%	
doubles	1648	48.8%	
VI grade	7544	35.1%	5.7%
singles	4545	32.8%	
doubles	2999	38.5%	
VII-VIII grade	3431	32.6%	2.6%
singles	2031	31.5%	
doubles	1400	34.1%	
IX-X grade	3450	32.7%	1.7%
singles	2245	32.1%	
doubles	1205	33.8%	
XI-XIII grade	2325	39.3%	4%
singles	1142	37.3%	
doubles	1183	41.3%	

independent event. In order to estimate the probability of answering correctly, we used a Markov chain Monte Carlo approach, then we relied on this estimation to compare the performances of singles and doubles, and to contrast them with relevant combinatorial benchmarks.

Estimating the probability of a correct solution. Let us consider the probability of the event "correctly solving any Bebras task", that is the probability that covariate C (as for "correctness") is 1 (C is 0 if the team gives the wrong answer to the task). We can estimate such probability by sampling a probabilistic model in which C is a random variable with a Bernoulli likelihood with an unknown parameter p, the probability of solving any task (not a specific one); then we estimated the *a posteriori* (*i.e.,* having seen the actual data) distribution of p with a Markov chain Monte Carlo approach (to implement our model we used the probabilistic programming language Stan[3]). We used a uniform prior for p: this assumption is rough (p is certainly different from 0 and 1, for example), but it matters very little in the process since we have a lot of data and the estimation of the posterior distribution is in fact rather robust w.r.t. to the choice of the prior. One could estimate p by simply taking the average success ratio (see Table 1), but this is a *point estimation* with no information about the uncertainty of its value[4]. The method we followed[5], instead, gives the whole distribution of p that we can use to estimate uncertainty intervals (for example the range in which 99% of the distribution lies), valid under the explicit model we used (*i.e., C* is Bernoulli distributed with unknown p).

In particular we used this method to estimate the distribution of probability $p_{singles}$ of the event "correctly solving any Bebras task" for any singles, and of probability $p_{doubles}$ of the event "correctly solving any Bebras task" for any

[3] See `https://mc-stan.org/`.

[4] One can estimate also the variance of p in order to have a measure of the variability, but an estimation of the error with respect to the "true value" needs inevitably some assumption on the underlying distribution.

doubles. More formally,

$$p_{singles} = p(C = 1 \mid teamsize = 1)$$
$$p_{doubles} = p(C = 1 \mid teamsize = 2)$$

Analysis of the impact of team composition on the correctness of answers. We expect that working in pairs improves the performance of teams. More formally, we expect $p_{singles} < p_{doubles}$. From a purely combinatorial viewpoint, we can say that the collaboration in a pair is *fully successful* if the pair is able to answer correctly whenever there is at least one of its member that would answer correctly alone. This means that the pair is able to recognize the correct answer even when the other member, alone, would answer incorrectly. We say that the collaboration is *fully harmful* in the opposite, worst-case, scenario, that is if the pair gives a wrong answer except when both pupils are able to answer correctly alone. This means that when only one of the pupils, alone, were able to answer correctly, the pair would not be able to recognize the correct answer and that the wrong answer always prevails. In general, we expect that the collaboration takes place at an intermediate level between fully harmful and fully successful. In probability terms, a pair is right with probability $p_{worst} = p^2_{singles}$ if the collaboration is fully harmful, and with probability $p_{best} = 1-(1-p_{singles})^2 = 2p_{singles}-p^2_{singles}$ if it is fully successful. We will compare this combinatorial benchmarks with the actual performance of doubles.

Teams with unusual team composition. The composition of teams is decided by teachers. Organizational issues (*e.g.,* the availability of a sufficient number of computers) probably had a relevant role in this choice. Moreover, constraints on the size of teams were due to the pandemic special regulations, which varied among regions and school levels (*e.g.,* remote attendance was avoided in primary school, whereas hybrid attendance was very common in high school); during the contest, some classes were attending in person, others remotely, and others used hybrid attendance. Besides these external factors, teachers were free to choose between singles and doubles. For instance they may have built teams randomly, or may have let their students choose how and with whom to participate, but they may also have considered students' prior ability to form balanced teams; in particular, they may have decided to pair students with special educational needs with a mate, or to let excellent students compete alone in a single team. We do not have any direct information about the criteria each teacher used to compose their teams. However, we know the number of double and single teams for each teacher, and this allows us to distinguish the cases where the choice of a different size is dictated by situations like the class having an odd number of pupils from special cases where the composition of a team turns out to be unusual for that teacher, and hence might be related to the ability of its components. We focus on the set of teams that have a typical composition among those of the same teacher: these are the singles of teachers who have at least 75% of singles and the doubles of teachers who have at least 75% of doubles. For these teams ("typical teams") we have reasons to believe the composition type is not biased by the

members' prior ability, whereas the others could have been formed according to some specific ability-related criterium. In fact, we found 17'871 teams with a typical composition type, and only a small proportion of all teams (8%) with an untypical composition type. It is still possible that some criterium to compose teams was adopted at the school level, but we believe this is improbable in the general case, since mixing people from different classes is normally quite difficult in our school system and unlikely for a non competitive contest like Bebras.

The role of content and task features. The difference in performances between singles and doubles varies from task to task, and we identified the tasks where the impact of team composition on correctness was higher. We analyzed the specific content and features of those tasks and formulated some hypotheses that would explain the higher impact for those tasks. We provided some support for these hypotheses by analyzing the data concerning the interaction of teams with the contest platform when solving those tasks.

3 Findings

3.1 Comparing performances of singles and doubles

Figure 1 shows the distribution of probabilities $p_{singles}$ and $p_{doubles}$ together with the combinatorial benchmarks corresponding to fully successful and fully harmful collaborations.

The probability of solving a task (any task) for singles is on average 33%. Our model estimated that 99% of the probability mass (High Density Interval, HDI) lies between 0.33 and 0.34. Doubles have a higher probability (the mean of $p_{doubles}$ is 39%, HDI: 0.39–0.40) and the difference is on average +6% (HDI: 0.054–0.065). However, 39% is much *lower* than 56%, the value one would have with fully successful collaborations ($p_{best} = 1 - (1 - p_{singles})^2$).

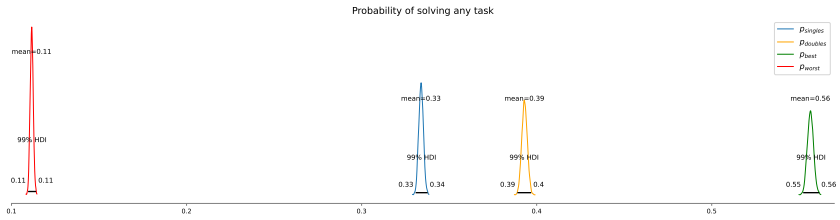

Fig. 1. The distribution of $p_{singles}$ (blue) is less than the distribution of $p_{doubles}$ (orange). The figure shows also the benchmarks for fully harmful collaboration (red) and fully successful (green) collaboration.

Notice that the diagram represents distributions of probability for p_{worst}, $p_{singles}$, $p_{doubles}$, p_{best}, from left to right. According to our model, the estimated

values (their distributions) of $p_{singles}$ and $p_{doubles}$ do not overlap (in particular their HDI do not overlap), thus the difference (and its measure) is supported by a clear evidence, if our statistical model is a sensible abstraction of our domain.

In summary, it is clear that the overall effect of collaboration is positive, however it is limited with respect to the combinatorial benchmark p_{best} of the fully successful collaboration.

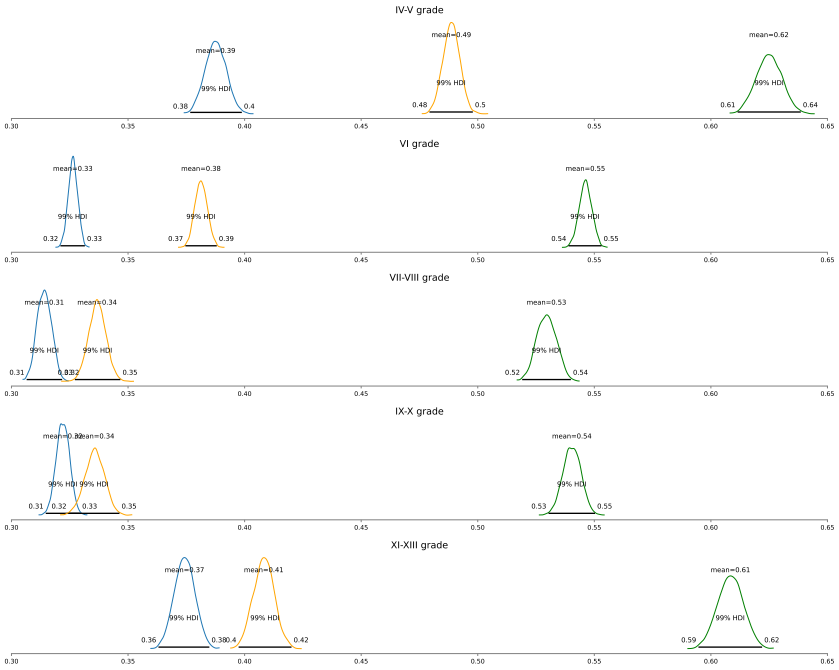

Fig. 2. Diagrams showing the differences between actual data ($p_{singles}$ blue, $p_{doubles}$ orange) and fully successful collaboration (p_{best} green); worst-case collaborations are not shown.

Figure 2 shows the distribution of probabilities for the five age categories for the typical teams (see 2.2). We can observe that the improvement from $p_{singles}$ to $p_{doubles}$ is greater for the youngest and decreases with age. Such an improvement is less evident in the categories where $p_{singles}$ is already low; for IX-X grade, there is even some uncertainty about the actual improvement. The gap between $p_{doubles}$ and the best-case benchmark is large for most categories, except for IV-V grade.

3.2 Impact of tasks content and features

For each task t of the 60 tasks used in the contest, we computed the probability distribution of the event "correctly solving task t" for singles and doubles, and the

delta between them. The diagram in Figure 3 positions all 60 tasks considering the mean probability for singles ($p_{singles}$) and the delta $p_{doubles} - p_{singles}$. The more a task is on the right (high $p_{singles}$), the simpler it resulted (high success probability by a single solver); the more a task is in the upper part of the diagram (high delta), the higher was the increase of success probability from single to double teams.

The red points are those where the delta is positive with high certainty (99% HDI of $p_{singles}$ and $p_{doubles}$ do not overlap), whereas the gray ones are those where the evidence of an increase is not certain enough (99% HDI of $p_{singles}$ and $p_{doubles}$ partially overlap, even if sometimes just slightly).

It is worth noticing that the top left part of the diagram is empty; this means that for hard tasks there is no benefit from working in pairs. Similarly, the most improvement is measured more often on tasks of medium difficulty. We analyzed some of the tasks where the improvement was more evident and we made some hypotheses about the content and features of those task, that would explain such improvements. Due to space constraints, here we discuss only one of them.

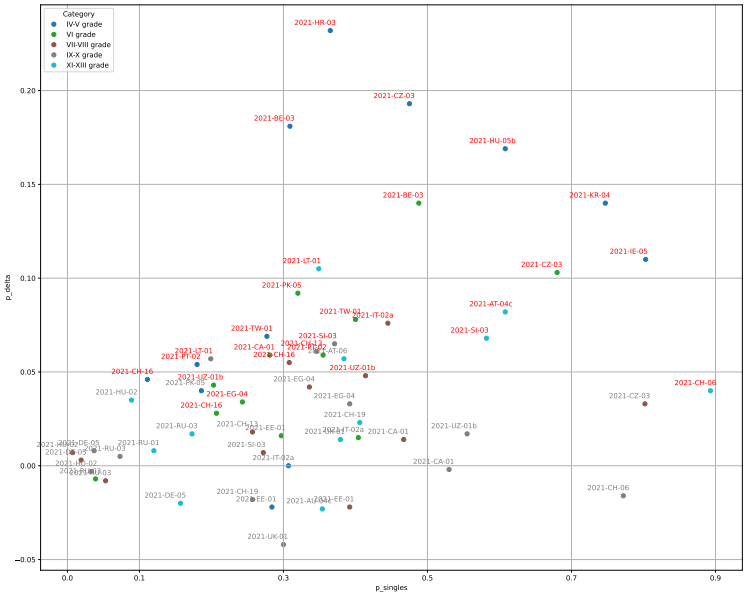

Fig. 3. Scatterplot of the tasks according to the success probability for the task and the size of improvement on the task moving from singles to doubles. For red tasks the increase has a solid evidence in the estimation (99% HDIs of $p_{singles}$ and $p_{doubles}$ do not overlap); for gray tasks 99% HDIs of $p_{singles}$ and $p_{doubles}$ overlap. The color of the dot gives the age group.

Task 2021-BE-03 (Necklaces instruction). This task proposes a programming exercise requiring to write a sequence of characters complying with a given syn-

tax. We make the hypothesis that doubles are better able to note and correct syntax errors that could go unnoticed to an individual. This hypothesis is in line with the literature, where the fact that syntax is a hurdle in learning to program is often discussed and where this fact has motivated the development of block-based programming languages, the use of Parsons problems, *etc.*. We analyzed the data for the "IV-V grade" age group, where the collaboration was more effective. We collected the following measures relevant for the solving process:

Total number of modifications The times the solvers modified the string of characters: it happened on average 26.7 (std. dev.: 23.5) times for singles, 27.1 (std. dev.: 20.7) times for doubles.

Number of corrections The times the solvers modified the string of characters excluding appends (these insertions or changes are likely to be corrections): it happened on average 5.6 (std. dev.: 9.5) times for singles, 5.7 (std. dev.: 10.5) times for doubles.

Percent number of corrections The times the solvers modified the string of characters excluding appends w.r.t. total modifications: it is on average 13.7% (std. dev.: 0.14) for singles, 14.1% (std. dev.: 0.13) for doubles.

Number of resets The times the solvers deleted the string of characters: it happened on average 0.8 (std. dev.: 1.5) times for singles, 0.7 (std. dev.: 1.3) times for doubles.

Percent times the solution was changed into wrong How frequently was a correct solution then changed in a wrong one: it happened for 0.8% of the singles and for 0.4% of the doubles. Overall it happened only for 0.5% of the teams.

The two populations differ somewhat and the doubles seem to be slightly more active with the platform, but no macroscopic differences were found. This task however shows a remarkable property: solvers are in general very stable on a correct solution, when it is found. In other words, a correct solution is easy to recognize as such. This could explain why the doubles improved so much (the success ratio is 31% for singles and 49% for doubles): it is enough that one of the two solvers identifies the correct solution, the other will accept it easily; in fact $p_{double} = 0.49$ is very close to $p_{best} = 0.52$. In order to check the validity of this last observation we analyzed also the data for 2021-EE-01, a task in which, in the same "IV-V grade" age group the increment for doubles is dubious. The "Percent times the solution was changed into wrong" for 2021-EE-01 is much higher than for 2021-BE-03: 22% for singles and 26% for doubles, 25% overall.

4 Limitations and threats to validity

Indirect measures of collaboration. The main limitation of this study is that we do not have any direct data about how teams solve tasks and collaborate. We only have the measure of their performance in the Bebras challenge, and some indirect data provided by log data related to their interaction with the Bebras platform during the contest. Thus, our findings cannot be considered definitive, and need

to be further checked, *e.g,* possibly with in-depth qualitative study based on observing students interacting with a mate when solving tasks. However, the size of the data set and the rigorous methods used to analyze it supports the validity of these preliminary findings, which suggest promising directions for future investigations.

Independence of team size from team ability. If doubles were formed by pupils with lower prior ability, this would provide some explanation for the limited improvement observed between the performance of doubles w.r.t. singles. We addressed this possible bias in two ways. On the one hand we excluded from the analysis the 8'706 teams whose composition type resulted atypical w.r.t. the rest of teams of their teachers. The average increment from singles to doubles computed on the original dataset results to be slightly lower than the one showed in Table 1. The inclusion of the small proportion of teams (8%) that are possibly biased w.r.t. ability decreases slightly the advantage of having a second person in the team, supporting the hypothesis that their teachers assigned the best students to the single teams. On the other hand, we studied the geographic provenance of teams and used this as a proxy for their ability; more precisely, we used the result of standardized tests conducted every year in all schools of our country[5]. We found neither evident trends nor correlations between the proportion of singles in a region and the results in standardized tests in that region, which suggests that there is no correlation between the prior ability of teams and their composition. Even though the test results are available also with finer definition (*e.g.,* by individual school), we conducted our analysis only at the regional level, since it is not mandatory for registered teams to enter details about their school. Moreover, an analysis at the school level would pose several legal problems since we did not ask in advance for an explicit consent and the data about the standardized tests are not available as open data.

Source for team type data. The team type for each team is entered by teachers when they register their teams, and we have no direct control on the fact that the actual composition of a team corresponds to the declared one. In particular many situations may occur (*e.g.,* absence of a mate the day of the contest, odd number of pupils in a class, ...) that yield to a team registered as doubles actually being formed by a single student only. However, in order to produce certificates for their teams after the challenge, teachers had the possibility to enter in the system additional information on the teams' members. Most teachers used this feature. In order to address the possible bias of false doubles, we did not used the declared team type but adjusted the team type value in our dataset as follows: i) we excluded from the analysis all teams without explicit information on their members, since it is dubious whether they should be indeed considered as doubles or singles; ii) similarly, we set the team size type according to the number of filled in members (in some cases this meant to change the composition w.r.t. the one declared upon teams registration). As a result we ended up considering a dataset of 19'490 teams, among the larger number of 28'196 teams who participated in

[5] We used the data provided by INVALSI for the school year 2021, taken from `https://invalsi-serviziostatistico.cineca.it/`; see also [11] for a previous study.

the challenge. The remaining 8'706 teams are not invalid, but their size was uncertain and we preferred to restrict the analysis to data with some guarantees to have been curated by the teachers.

Contest aggregation. One could also take into account that the tasks come packed together in a suite of twelve. We carried out the same analysis by starting with a model with contest data aggregated by suites, but we did not find any visible difference. In principle the data observed on suites could fit the model worse (note that the two models are mathematically equivalent). The difference in the uncertainty is negligible, therefore considering the tasks independent one from each other seems to be a viable hypothesis.

5 Related work

Group work is often proposed as a way for improving learning, and many studied the social and emotional advantages children can gain from working together [2]. In particular, collaborative learning is an educational approach to teaching and learning that involves groups of learners working together to solve a problem, complete a task, or create a product [4]. However, while collaboration in pairs or small groups can facilitate pupils' learning and development, many observations of classroom practice show that group work does not realise the potential promised by research [10]. Sometimes peer interaction can even result in poorer learning outcomes [6]. In fact, although collaboration is often considered a beneficial learning strategy, identifying the key factors which make a collaboration successful or not is still an open issue. [7] studied important features for educators to consider when deciding when and how to include collaboration in instructional activities. Our study tries to understand in which context or task the collaboration is more effective. In 2015 the Programme for International Student Assessment (PISA[6]) launched the first large-scale, international assessment to evaluate students' competency in collaborative problem solving. It required students to interact in order to solve problems. It included group decision-making tasks (requiring argumentation, debate, negotiation or consensus to arrive at a decision), group co-ordination tasks (including collaborative work), and group-production tasks (where a product must be created by a team, including designs for new products or written reports). Collaborative problem-solving performance is positively related to performance in the core PISA subjects (science, reading, and mathematics), but the relationship is weaker than that observed among those other domains. Girls perform significantly better than boys in collaborative problem solving in every country and economy that participated in the assessment; students have a generally positive attitude towards collaboration [8].

6 Conclusions

Our observational study confirms that the effect of collaboration is positive, but it is rather limited compared to what one could expect from a fully successful

[6] See `https://www.oecd.org/pisa`.

collaboration. The positive effect of collaboration seems somewhat to decrease when the grade increases: this certainly needs further in-depth analysis, it could be related to some specificity of the age groups, but also to task features, often rather different for older students. For example, when a correct solution is easy to recognize, the collaboration seems to work more efficiently. The main limitation of this study is that we did not directly observe how teams solved tasks and collaborated. Even the interaction data we analyzed are indirect and can be interpreted in different ways. Our findings, although promising, should be considered preliminary and we intend to design a follow up qualitative study, in which we will observe students interacting to solve tasks.

References

1. Dagienė, V.: Sustaining informatics education by contests. In: Proceedings of ISSEP 2010. Lecture Notes in Computer Science, vol. 5941, pp. 1–12. Springer, Zurich, Switzerland (2010)
2. Galton, M., Williamson, J.: Groupwork in the primary classroom. Routledge (2003)
3. Haberman, B., Cohen, A., Dagienė, V.: The beaver contest: Attracting youngsters to study computing. In: Proceedings of ITiCSE 2011. pp. 378–378. ACM, Darmstadt, Germany (2011)
4. Laal, M., Ghodsi, S.M.: Benefits of collaborative learning. Procedia-social and behavioral sciences **31**, 486–490 (2012)
5. McElreath, R.: Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC (2020)
6. Messer, D.J., Joiner, R., Loveridge, N., Light, P., Littleton, K.: Influences on the effectiveness of peer interaction: Children's level of cognitive development and the relative ability of partners. Social development **2**(3), 279–294 (1993)
7. Nokes-Malach, T.J., Richey, J.E., Gadgil, S.: When is it better to learn together? insights from research on collaborative learning. Educational Psychology Review **27**(4), 645–656 (2015)
8. OECD: PISA 2015 Results (Volume V) (2017)
9. Williams, L.: Integrating pair programming into a software development process. In: Proc. 14th Conference on Software Engineering Education and Training. pp. 27–36 (2001). https://doi.org/10.1109/CSEE.2001.913816
10. Wood, D., O'Malley, C.: Collaborative learning between peers. Educational Psychology in Practice **11**(4), 4–9 (1996). https://doi.org/10.1080/0266736960110402
11. Bellettini, Carlo and Lonati, Violetta and Monga, Mattia and Morpurgo, Anna: An analysis of the performance of Italian schools in Bebras and in the national student assessment INVALSI. In: Fronza, Ilenia and Pahl, Claus (ed.) Proceedings of the 2nd Systems of Assessments for Computational Thinking Learning Workshop (TACKLE 2019). CEUR Workshop Proceedings, vol. 2434 (September 2019)
12. Bellettini, Carlo and Lonati, Violetta and Monga, Mattia and Morpurgo, Anna: Behind the Shoulders of Bebras Teams: Analyzing How They Interact with the Platform to Solve Tasks. In: Lane, H. Chad and Zvacek, Susan and Uhomoibhi, James (ed.) Computer Supported Education. pp. 191–210. Springer International Publishing, Cham (November 2020)