

The Crucial Importance of Typical Discussion Roles of Pupils for the Effective Implementation of Peer Instruction in Teaching Elementary School Mathematics

Krucialní význam typických diskuzních rolí žáků pro efektivní implementaci peer instruction ve výuce matematiky na základní škole

Tomáš Zadražil¹

¹ Department of Mathematics Education, Charles University in Prague, Sokolovská 83, Praha, Czech Republic; tomas.zadrazil@gmail.com

Studies over the past decade have indicated that Czech pupils do not have an appropriate level of mathematics comprehension and their attitudes towards this subject deteriorate over time. The article deals with a possible solution to the described problem; namely the implementation of *peer instruction*, an active learning strategy, in elementary mathematics teaching. In order to decide whether it is possible to use peer instruction as a teaching method in an elementary school environment, action research project with the properties of a mixed empirical study was performed in one eighth grade class (thirty participants) at a Czech multi-year grammar school. The main idea was to compare the results of the class before, during, and after the implementation of peer instruction. The pupils' level of understanding was monitored by normalized learning gains calculated on the basis of pre/post-testing design. Changes in the pupils' attitudes towards mathematics were mapped using continuous and pre/post-test questionnaires. In the spirit of the action research, interim data and results were regularly discussed with a group of selected pupils or experts in the field. The results show that there is a strong relationship between normalized learning gain and one of four typical roles with which pupils identify during group discussions: *passengers*, *standard discussants*, *advisors*, and *dominant speakers*. The research has also indicated that peer instruction needs to be appropriately modified to increase the passengers' activity.

Key words:
peer instruction,
education of
mathematics, elementary
school, normalized
learning gains, attitudes
towards mathematics.

Received 7/2020
Revised 10/2020
Accepted 11/2020

Studie uplynulé dekády ukazují, že čeští žáci nerozumějí matematice na očekávané úrovni a že se jejich postoje vůči tomuto předmětu s časem zhoršují. Článek se zabývá možným řešením popsání problému, a sice implementací metody aktivního učení *peer instruction* ve výuce matematiky na základní škole. Za účelem rozhodnout, zda je pomocí peer instruction možno vyučovat i v prostředí základní školy, byl v jedné třídě českého gymnázia (třicet žáků osmého ročníku) realizován akční výzkum o vlastnostech smíšené empirické studie. Hlavní myšlenkou bylo srovnání výsledků pozorované třídy před, během a po implementaci peer instruction. Úroveň porozumění žáků byla sledována pomocí normalizovaného učebního zisku vypočteného na základě designu pre/post testování. Změny žákovských postojů vůči matematice byly mapovány pomocí průběžných a pre/post dotazníků. V duchu akční studie byla průběžná data i výsledky pravidelně diskutovány se skupinou vybraných žáků nebo s odborníky na dané problematiku. Výsledky ukazují, že existuje silný vztah mezi normalizovaným učebním ziskem a jednou ze čtyř typických rolí, se kterou se žáci identifikují během skupinových diskuzí: *pasážéři*, *standardní diskutéři*, *poradci* a *dominantní řečníci*. Rovněž se ukazuje, že pro navýšení aktivity pasažérů je potřeba peer instruction vhodně modifikovat.

Klíčová slova:
peer instruction, výuka
matematiky, základní
škola, normalizovaný
učební zisk, postoje vůči
matematice.

Zasláno 7/2020
Revidováno 10/2020
Přijato 11/2020

1 Introduction

Studies of the past decade show that the attitudes of Czech pupils towards mathematics deteriorate as their school years advance and the biggest decline occurs in the second stage of primary school from sixth to ninth grade (Pavelková & Hrabal, 1988; Chvál, 2013). In general mathematics is seen as a difficult and unpopular subject (Pavelková & Hrabal, 1988). Moreover, the results of international tests (PISA2012, TIMS2007, etc.) indicate that Czech pupils' understanding of mathematical concepts is insufficient (Vondrová et al., 2015).

Professionals, politicians, and the lay public are aware of those problems, and there is an effort to deal with them; however, there is no consensus on the optimal solution. One possible way presents *active learning*, which could be defined as any instructional technique that engages students in the learning process (Prince, 2004). Based on previous studies it could be stated that active learning has a positive impact on

students' attitudes towards the subject, their content memory, and engagement in general (Prince, 2004). It could be also claimed that active learning slightly increases students' exam performance (Freeman et al., 2014) and considerably boosts their conceptual understanding (Prince, 2004; Freeman et al., 2014).

In the following section peer instruction is proposed as a possible way of alleviating above mentioned problems.

2 Peer Instruction

In 1985, it was pointed out by Halloun and Hestenes (1985) that although a great deal of college students were able to state Newton's third law, not many of them fully understood it. Their understanding was the same before and even after introductory physics courses. In fact, the courses changed almost nothing in terms of students' preconceptions about Newtonian mechanics. This discouraging conclusion led Eric Mazur (2009), a physics professor at Harvard University, to come up with his own teaching approach called peer instruction.

Peer instruction is one of the world's most known active learning strategies (connected with voting) that mainly relies on group discussions over a difficult multiple-choice conceptual question (referred to as a *ConcepTest*). The aforementioned benefits (mentioned in Introduction for active learning) are especially prominent in peer instruction (Mazur, 1997; Vickrey et al., 2015; Chien et al., 2016) especially in terms of students' exam performance and increased conceptual understanding (Mazur, 1997; Hake, 1998; Michinov et al., 2015; Chien et al., 2016; Balta et al., 2017).

Effectiveness of peer instruction is generally demonstrated through comparison of normalized learning gains of experimental and control group. Normalized learning gains are typically significantly greater within experimental groups than in control groups as was for example shown by Eric Mazur himself (Mazur, 1997; Mazur & Crouch, 2001) and as it was highlighted in several meta-analysis (Vickrey et al., 2015; Balta et al., 2017). This approach works well in subjects like physics where is possible to use the same pre/post-test in order to compute normalized learning gains because students are familiar with taught concepts. Unfortunately, especially in the education of elementary school mathematics pupils typically are not often familiar with taught concepts and therefore this approach is quite hard to apply.

Peer instruction lessons are divided into several blocks (see Fig. 1). Each block begins with a brief lecture on a selected concept or another introductory educational activity (1). During this step, formulas and other mnemonics that may distract the learners from conceptual understanding should be avoided. At the end of opening activity, a ConcepTest is posed (2), and students are given a short time to reflect upon it individually (3). Students then vote for their answers via flashcards, clickers, or an application on their personal smart devices (3). The following step is selected based on the distribution of answers obtained. If more than 70% of the answers are correct, the solution will be briefly explained by the teacher or one of the students (5). On the other hand, if less than 35% of the answers are correct, the students will be given appropriate help (6) or the problematic concept will be explained in another way (7). However,

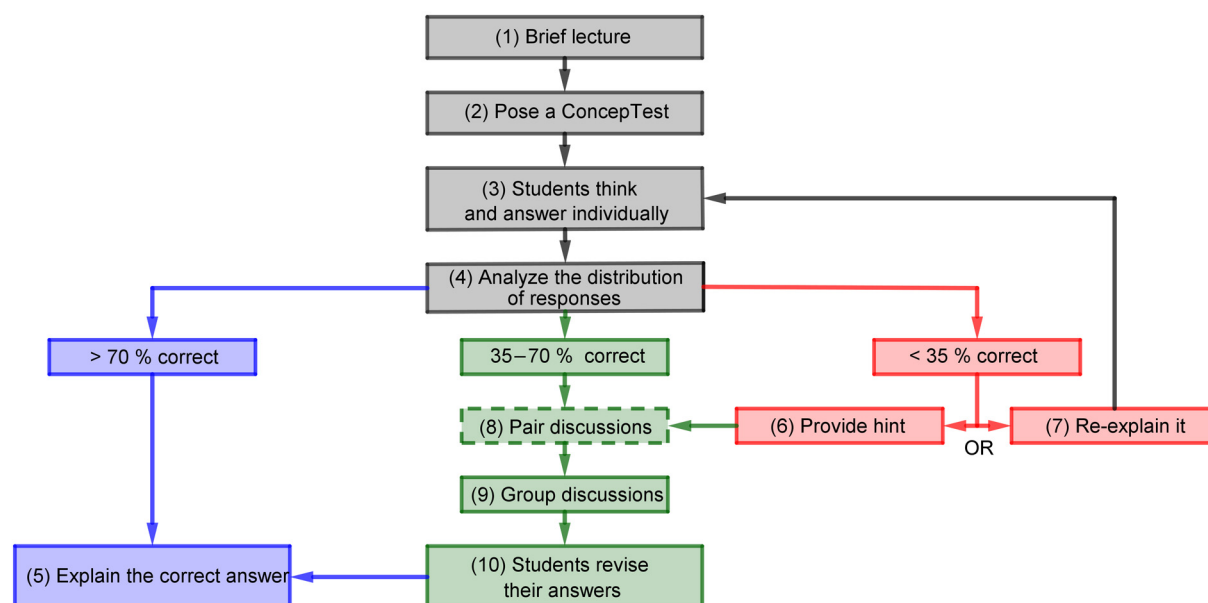


Fig. 1: Extended scheme of one block of peer instruction

the last remaining situation (35–70% of the answers are correct) is crucial for peer instruction. In this case, the students are asked to create smaller discussion groups (three to four members) to convince their peers of the validity of their answers (9). The group discussions are finished with a revised vote (10) and a subsequent explanation of the correct solution to the ConcepTest (6). In addition to the steps originally designed by Eric Mazur (1997), the extended peer instruction scheme contains a voluntary step (8), in which students are first asked to pre-discuss their answers in pairs before starting group discussions (9).

Group discussions (9) are evidently the best option here. At this step, students are encouraged by the instructor to justify – not just to merely report – their statements to peers. A student who has recently understood the concept discussed knows how to overcome its obstacles and is thus more likely to guide his peers to understanding than the instructor himself. There is also a considerably lower communication barrier among students themselves than between students and the instructor. The revised voting is therefore usually connected to a significant increase of votes in favour of the right option (Mazur, 1997; Mazur & Crouch, 2001; Pilzer, 2001; Lucas, 2009; Michinov et al., 2015; Vickrey et al., 2015). This increase has been proven as consequences of the group discussion rather than mere copies of the most common answer (Smith et al., 2009; Vickrey et al., 2015).

Even though peer instruction is now one of the most surveyed teaching strategies connected to electronic voting and there have been many studies on its implementation in STEM disciplines (Mazur & Crouch, 2001; Vickrey et al., 2015; Michinov et al., 2015; Balta et al., 2017), few examples have been realised in elementary (Chen et al., 2005) or high school education or in mathematics education (Pilzer, 2001; Lucas, 2009; Olpak et al., 2018).

Along with the benefits mentioned in the introduction, the implementation of peer instruction also poses considerable challenges which are highly relevant to this article. Peer instruction was originally designed for the purposes of college physics education and the standard physical ConcepTest deals with objects and phenomena with which students are inherently familiar. Unfortunately, this is not the case for mathematical concepts, which are usually entirely unfamiliar to students (Pilzer, 2001). Group discussions could be led by dominant group members at the expense of their peers (Lucas, 2009; Michinov et al., 2015). In such cases, the discussions will not be fruitful for everyone and this is perhaps the main reason why it is difficult to exceed a medium level of average normalized learning gain (Hake, 1998; Mazur & Crouch, 2001; Michinov et al., 2015). In addition, Chen et al.'s study (2005) pointed out that elementary school pupils could have insufficient social skills to maintain effective group discussions over ConcepTests.

3 Methodology

3.1 Purpose of the study

The main purpose of the study was to decide whether it is possible:

- a) to satisfy assumptions of peer instruction in lower secondary mathematics education (e.g., are there sufficient social skills to maintain fruitful group discussions?);
- b) to achieve similar outcomes promised by research studies on university physics peer instruction (e.g., more profound conceptual understanding, significant normalized learning gains, improvement of attitudes towards mathematics, etc.).

3.2 Research Timeline and Design

In order to answer questions a) and b), an action research project (Mertler, 2019) was carried out from the beginning of the 2018/2019 school year to the untimely end of the 2019/2020 school year (due to coronavirus), with a single class (in the following text referred as the class or our class) of thirty eighth grade participants at a Czech grammar school (ninth graders in the 2019/2020 school year). Within the class mathematics was taught by the researcher himself for the whole duration of research. The core idea of study was to compare the class before, during and after the implementation of peer instruction to itself and to global datasets (standards for tests and questionnaires from: (Pavelková & Hrabal, 1988); (Chvál, 2013); PISA2012; TIMSS2007, etc.). Fig. 2 shows a simplified scheme of our research timeline. (Numbers in curly brackets refers to Fig. 2.)

At the beginning of the very first mathematics class, questionnaires Q1 and Q2 {1} (Pavelková & Hrabal, 1988; Chvál, 2013) were assigned with connected datasets for the Czech Republic to determine the pupils' initial attitudes and motivational structures toward mathematics. During the following lesson, pupils completed a test T1 {1} aimed to gauge their initial understanding of geometric concepts (e.g. *area, perimeter, distance, median, altitude*, etc.) and argumentation skills. The very same Q1 questionnaire was assigned to the pupils four more times – after two months of classic teaching (i.e. combination

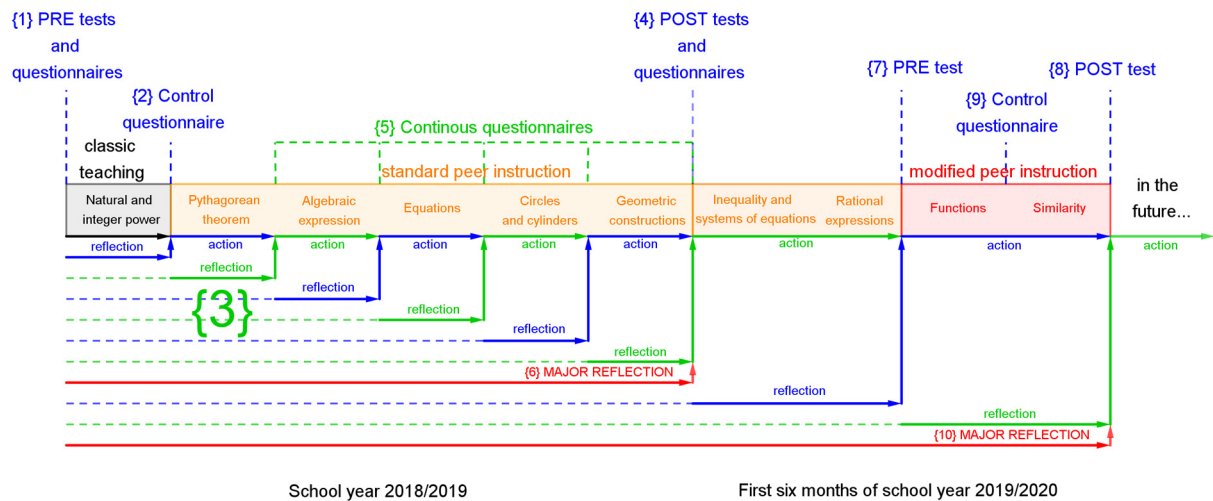


Fig. 2: Research Scheme

of lecturing and practising) {2}, after another eight months of peer instruction (in Fig. 2 referred as a period of standard peer instruction), at the end of the 2018/2019 school year {4}, and lastly, after the first sixth months of the 2019/2020 school year {9} (in Fig. 2 referred as a period of modified peer instruction, which differs from standard peer instruction by added pair discussion step as was mentioned in the description of Fig. 1). Q2 was only assigned once more at the end of the 2018/2019 school year {4}.

The argumentation-understanding T1 test was assigned to pupils once again at the end of the 2018/2019 school year {4} along with the T2 test on relevant mathematical tasks from that school year according to the Programme for International Student Assessment 2012 (PISA2012) and the Trends in International Mathematics and Science Study 2007 (TIMSS2007). Based on the results of pre-questionnaires Q1 and Q2 as well as continuous test performance during the first two months of classic teaching, the pupils were divided by cluster analysis into five characteristic groups shortly after {2}. (The initial understanding of geometric concepts and argumentation skills was quite low for the class in general; as such, the results of pre-test T1 were not included in the cluster analysis.) Resulting characteristic groups were then carefully compared to observations and experiences within the class. For example, one characteristic group consisted of pupils who had good attitudes toward mathematics and high continuous tests performance. In order to see longitudinal changes in pupils' attitudes towards mathematics, questionnaire Q3 was assigned periodically after each topic was taught {5}. To determine pupils' attitudes towards peer instruction itself, questionnaire Q4 was used from the study of Olpak et al. (2018) at the end of the 2018/2019 school year {4}. (Discussions of design, usage and results of Q1, Q2, Q3, T1 and T2 can be found later in the article.)

In the spirit of action research (Mertler, 2019), several pupils from the characteristic groups were selected to form a reflexive group {2} which met monthly from the beginning to the end of the research. The objective of the reflexive group was to discuss continuous results, experiences, and observations in order to suggest appropriate improvements to simplify and improve peer instruction implementation in lower secondary school mathematics. As can be seen in Fig. 2, the reflexive group mainly discussed the most recent topic but considered "all previous history" {3}. Meetings of the reflexive group were audiotaped, transcribed, coded, categorized and then carefully analysed through both analytic induction and comparative analysis technique. Resulting outputs of analysis together with proposals of the group were implemented immediately. One input from the reflection group was the pair discussion step (step 8 in Fig. 1). The efficacy of this modification was later checked by another pre/post-argumentation understanding test T3 (testing concept of similarity in geometry) that was assigned before {7} and after {8} the mentioned upgrade to peer instruction. Another example input from the reflexive group was preference of using own smart devices for purposes of voting over both flashcards and clickers because of sense of ownership and higher anonymity (there were being used flashcards during first month of peer instruction, clickers in first two months of the 2018/2019 school year and smart phone's application Socrative in remaining time). In addition to the periodical reflection group meetings, there were also two major reflections carefully considering everything up to that date.

Experts in the field were constantly consulted on the research design, data collected, and continuous results. The usage of the questionnaire Q2 was verified with its co-author M. Chvál. Data collected on pupils' attitudes towards mathematics and their motivational structure discussed, along with research design, with the author of the questionnaire (Q1) I. Pavelková.

3.3 Pre/post-understanding argumentation test T1

The pre/post-understanding argumentation test T1 was constructed during our pilot study in the 2017/2018 school year. At the beginning, pupils' common geometric misconceptions were carefully selected. The first version of the test addressed the concepts of *perimeter*, *area*, *distance*, *altitude*, *median*, *angle*, and *polygon*. This version was then individually assigned to ten eighth grade pupils. The obtained results were then discussed with the very same pupils who wrote the test and other faculty colleagues. Based on the feedback, the test was upgraded to its second version which was then assigned to thirty eight grade pupils (to a different class within the same school as our class). The results of the second version were once again carefully analysed and discussed with involved pupils and faculty. The test was upgraded for the second and last time. The third version was first assigned to approximately sixty eighth grade pupils (to two different classes within the same school as our class) and then as the pre/post-test to our class. Two example tasks from the T1 test were published in Zdražil (2020) and another can be seen under this paragraph.

Example task from the T1 test

Ctirad drew an ABC triangle. Subsequent measurements determined the distance of the vertex C from the side c to be 12 cm. Decide which of the following statements for the distance of the centroid T of this triangle from the vertex C is certainly true.

- a) $|CT| = 4$ cm b) $|CT| = 6$ cm c) $|CT| = 8$ cm d) another option

The option we are looking for is... because...

Normalized learning gain of individuals was then computed by the Richard Hakes (1998) formula as the ratio of actual learning gain ($\text{post}[\%] - \text{pre}[\%]$) to the maximum possible learning gain ($100\% - \text{pre}[\%]$):

$$g = \frac{\text{post}[\%] - \text{pre}[\%]}{100\% - \text{pre}[\%]}$$

The average normalized learning gain ($\langle g \rangle$) of the class itself, or other specific subgroups, was computed in the same way but with the average pre/post-test scores $[\%]$.

There are three levels of normalized learning gain:

- low normalized learning gain $g < 0.3$,
- medium normalized learning gain $0.3 < g < 0.7$,
- high normalized learning gain $g > 0.7$ (Hake, 1998).

3.4 Pre/post-understanding argumentation test T2

The pre/post-understanding argumentation test T2 was constructed in a slightly different way from the T1 test. The tasks included were primarily selected from ConcepTests that had been used with another peer instruction class during the previous 2018/2019 school year and were then modified to be open questions. The first version of this test addressed the concepts of *similarity* in geometry: the *similarity coefficient* and *theorems about the similarity of triangles* and was assigned to thirty ninth grade pupils. Once carefully analysed, the results were discussed with involved pupils and faculty colleagues. Based on feedback obtained, the test was upgraded to its second version. The second version was then assigned as the pre/post-test to our class (a month before and a month after *similarity* was taught in the class, respectively). Two example tasks can be seen under this paragraph.

Example task 01 from the T2 test

To cover a large pizza, exactly twice as many ingredients are needed as to cover a small pizza at the same coverage density. What is the ratio of the diameter of the large pizza to the diameter of the small pizza (in that order)? Explain your answer briefly.

Example task 02 from the T2 test

If we magnify a trapezoid four times (meaning each side), how many times does its area increase? Support your answer with calculations or at least briefly explain it.

3.5 Questionnaire Q1

The questionnaire Q1 from Pavelková and Hrabal (1988) consists of 5-points Likert scales (specified towards mathematics): *popularity, difficulty, importance, my talent, my motivation, my diligence, my interest, boredom*. It also contains a short survey on motivational structure.

3.6 Questionnaire Q2

The Q2 questionnaire is based on Osgood’s semantic differential technique and works with 13 evaluative bipolar 7-point scales and 16 evaluated concepts: game, future, counting, technic, truth, formulas, life, peers, mathematics, duty, me, geometry, theory, discussion, school, world, and teacher (of mathematics). An example of evaluation form for concept mathematics can be seen in Fig. 3.

		<i>very</i>				<i>slightly</i>				<i>slightly</i>				<i>very</i>			
1.	usefull	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	useless
2.	slow	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	fast
3.	strong	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	weak
4.	monotone	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	various
5.	young	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	old
6.	faraway	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	nearby
7.	beautiful	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	ugly
8.	passive	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	active
9.	deep	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	airy
10.	boring	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	interesting
11.	flexible	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	firm
12.	narrow	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	wide
13.	simple	---	:	---	:	---	:	---	:	---	:	---	:	---	:	---	difficult

Fig. 3: Example of evaluation form for concept mathematics from the questionnaire Q2

In addition to the usual descriptive analysis, the Euclidean distances of the evaluated concepts were calculated in the case of the Q2 questionnaire. Moreover, four dimensions for each of the evaluated concepts were also computed as mean values of ratings on corresponding scales (in terms of their negative – positive orientation):

- Dimension of evaluation: useful–useless, monotone–various, beautiful–ugly, boring–interesting:
 “Within this dimension, we look at the assessed concepts from the point of view of a certain subjective evaluation, it is an assessment of the concept at first based on our primary evaluation (the so-called ‘good’ or ‘evil’ concept). This factor captures our first emotional attitude to the concept under consideration (‘as we like it’).” (Pöschl, 2005, p. 13)
- Dimension of activity: slow–fast, young–old, passive–active, firm–flexible:
 “This factor is characterized by adjectives associated with movement, active change, dynamics, and variability over time.” (Pöschl, 2005, p. 13)
- Dimension of potency: strong–weak, nearby–faraway, airy–deep, narrow–wide
 “Within the dimension of potency, we look at the assessed concepts from the point of view of a certain static stress (force, measure, distance, weight, hardness, . . .). This factor could, in a certain approximation, be compared to the parallels of energy that must be expended to change a certain state).” (Pöschl, 2005, p. 13)
- Dimension of difficulty: simple–difficult

For example, if a pupil rated mathematics: 7 on useful–useless, 4 on monotone–various, 2 on beautiful–ugly, 5 on boring–interesting; the rating was then recounted into 1 on useless–useful, 4 on monotone–various, 5 on ugly–beautiful, 5 on boring–interesting and the dimension of evaluation was stated as $\frac{1+4+5+5}{4} = 3.75$.

3.7 Questionnaire Q3

Fig. 4 contains the first part of the continuously assigned questionnaire Q3, and the evaluation axes for the educational activities used. Pupils were asked to place each of the ten items on both axes after each of the topics were taught.

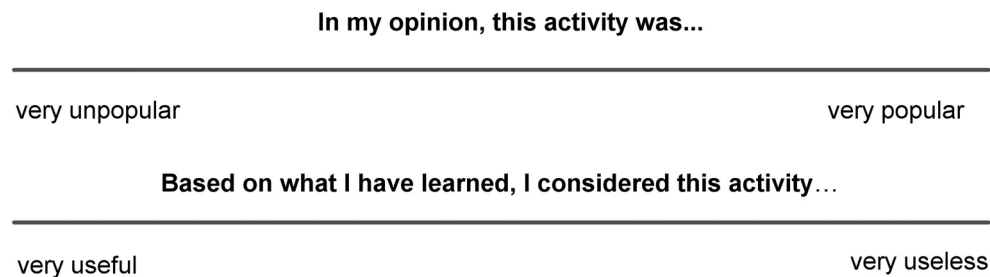


Fig. 4: Dominance–Effort model

In the second part of this questionnaire, pupils had to evaluate the topic in the same way as the concepts in questionnaire Q2 (see Fig. 3). Dimensions of evaluation, activity, potency, and difficulty were computed in the same way as questionnaire Q2 in order to follow changes in pupils' perception of mathematics and their attitudes over time and the topics addressed.

3.8 Questionnaire Q4

The questionnaire Q4 was originally connected to the article by Olpak et al. (2018) and was used without changes at the end of the 2018/2019 school year in order to state pupils' attitudes towards peer instruction itself. It consists of 25 questions grouped into three categories of pupils' evaluations regarding the peer instruction method, the ConcepTests, and the group discussion step.

4 Results

4.1 Relationship between normalized learning gain of individuals and their typical roles in group discussions with peers

Results presented in section 4.1 are based mainly on data collected during the 2018/2019 school year and were already presented in Tab. 1 and Fig. 5, 6 and 7 (Zadrazil, 2020).

4.1.1 Average normalized learning gain of the class

There was on average a medium normalized learning gain (Hake, 1998) for the class $\langle g \rangle = 0.49 \pm 0.22$. This result was in agreement with other studies, because the medium value of the average normalized learning gain is typical for peer instruction courses (Hake, 1998; Mazur & Crouch, 2001; Michinov et al., 2015).

4.1.2 Relationship between normalized learning gain and pupils' typical roles during group discussions

The prime assumption was that there should be a relationship between which characteristics group a pupil belonged to and his or her normalized learning gain, but such a relationship was not found. However, it was clear that five out of eight participants with the lowest normalized learning gain were pupils who had low test performance and bad attitudes toward mathematics and that the remaining three pupils were opposite to the other five in all of the characteristics (interviews with both pupils and their parents showed that these three pupils had preferred to learn by memorizing and mastering problem solving algorithms rather than by thinking or truly understanding. In other words, not everyone was willingly to switch from simply memorizing to a deeper understanding, which demands more energy to do so). Based on the results of the pre-test T1, questionnaires Q1 and Q2, and continuous test performance during first two months of classic teaching, the participants were divided into five characteristic groups, as was mentioned in the methodology section.

Another assumption raised from relevant literature was that group discussions could be led by dominant group members at the expense of their peers (Lucas, 2009; Michinov et al., 2015). A relationship was therefore expected between pupils' typical roles during group discussions and the normalized learning

gains obtained. An opportunity to classify typical roles was granted by the nature of peer instruction itself. Pupils group discussions could be carefully observed by an instructor (Mazur, 1997; 2009). During approximately six months of observation (within the class there were discussed 63 ConcepTests which corresponds to approximately 350–400 minutes of careful overt participant structured observation) four characteristic roles were identified:

- R1: Pupils who were initially called *statistics* (Zadrazil, 2020) were renamed to *passengers* (Garcia-Souto, 2020). Members of this role usually changed their answer after the group discussion to the most common one or that of the most dominant peer within their group (usually in role R3). After the end of the group discussion, these pupils typically started to cross the class (if free movement was allowed) in order to get an idea of the most common answer to the actual ConcepTest. The reflexive group itself suggested that amongst *passengers* also could be found pupils who only pretended to make an effort to solve the problem without an actual need to do so. There were nine R1 members in total.
- R2: These pupils attended group discussions and behaved in an expected manner for which they were called *standard discussers*. They listened to their peers and participated actively in group discussions. There were twelve R2 members in total.
- R3: This final group consists of two subgroups – pupils who could be called *overwhelming speakers* (R3+) and those who could be called *advisors* (R3–). Overwhelming speakers usually launched the discussion and issued minor interruptions the whole time. Other group members, with the exception of advisors or other overwhelming speakers, usually remained quiet and listened silently. Advisors acted quite differently from overwhelming speakers. They listened carefully to their peers and then came up with their own questions, comments, or clarifications. It was clearly desirable to have an overwhelming speaker and an advisor complement each other in the same group. The reason why overwhelming speakers and advisors formed the R3 role together was because both were typically good mathematical thinkers who actively worked in order to solve the ConcepTests posed during group discussions, although there was a difference in their need to dominate the discussion (overwhelming speakers practically always dominated over the discussion whereas advisors did not feel such a need). Moreover, there were situations forcing advisors to change their role to overwhelming speakers and vice versa. For example, in a group discussion with more overwhelming speakers, it was typical that some of them switched to advisors in favour of more dominant peers. On the other hand, in a group discussion of only passengers and advisors some advisors were forced to become more dominant overwhelming speakers. There were nine R3 members: four overwhelming speakers (R3+) and five advisors (R3–).

After first four months of observation there were only 7 pupils with unclear membership to the concrete role but there was still a plenty of time (i.e. around two hours of careful structured observation targeting mainly these seven pupils during twenty one remaining group discussions) to classify them properly. Observation was aimed to selected pupils' behaviour, their activity and both their verbal and nonverbal communication during group discussions. Brief records out of observation were transcribed into researcher's diary and continuously compared with other records.

Membership of concrete individuals to the concrete role was presented to the reflexive group, carefully discussed then confirmed, specified or changed based on given feedback (there were only two pupils out of thirty with unclear membership, who will be discussed later). Tab. 1 shows the dominance – effort model for typical pupil roles during group discussions. Standard discussers (R2) could be found somewhere between R3– and R3+ (in the third column) because they switched between these two roles during group discussions in terms of asserting dominance.

Tab. 1: Dominance – Effort model

	Weak effort to solve ConcepTests	Strong effort to solve ConcepTests
Weak need to dominate discussions	Passengers (R1)	Advisors (R3–)
Strong need to dominate discussions	–	Overwhelming speakers (R3+)

Average normalized learning gains $\langle g \rangle$ were once again computed for each group from R1 to R3 separately. Passengers achieved a small normalized learning gain $\langle g \rangle = 0.22 \pm 0.11$, standards discussers achieved a medium normalized learning gain $\langle g \rangle = 0.49 \pm 0.11$ and overwhelming speakerstogether with advisors achieved a high normalized leaning gain $\langle g \rangle = 0.75 \pm 0.12$. The difference between each pair of roles R1–R3 was found statistically significant ($p < 0.01$) by the Mann–Whitney U-test ($U = 0$ against $U_{001} = 11$ for R1 against R3, $U = 4$ against $U_{001} = 14$ for R1 against R2, $U = 3$ against $U_{001} = 21$ for R2 against R3).

4.1.3 Relationship between normalized learning gain, pupils' typical discussion roles and test performances

There was a medium correlation ($r = 0.53$) between normalized learning gain and continuous test performance during eight months of peer instruction as can be seen in Fig. 5. It is also clearly visible that almost everyone in the R3 role had a better total score than the test median (indicated by horizontal orange line). On the other hand, almost every member in the R1 role scored below the test median. Standard discussers (R2) achieved a total score somewhere between R1 and R3. Continuous tests always consisted of three classic and two conceptual questions; therefore, it can be stated that there was a relationship between degree of pupils' cognitive activity during group discussions and their continuous test performance.

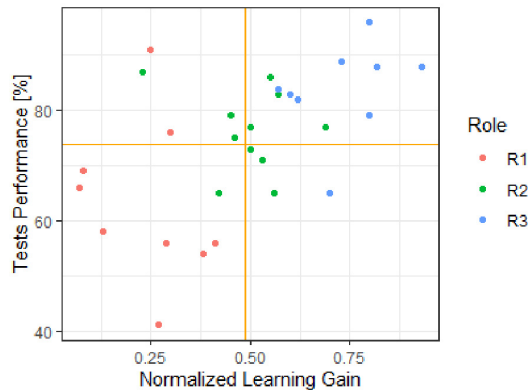


Fig. 5: Relationship between normalized learning gains and continuous test performance

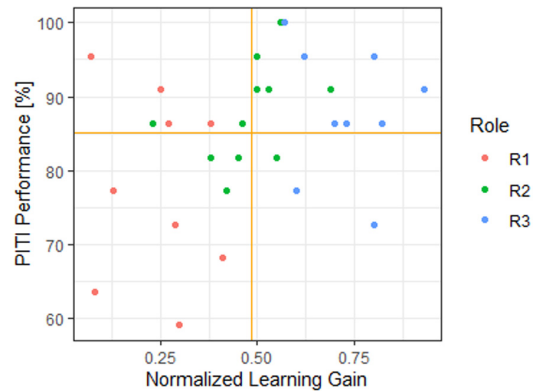


Fig. 6: Relationship between PITI performance and normalized learning gain

As was mentioned in the methodology section (the T2 test), the selection of tasks (from the PISA2012 and the TIMSS2007 relevant to the subjects taught) was assigned in two waves to the participants by the end of the 2018/2019 school year. The total performance from the PISA2012 and the TIMSS2007 (called PITI performance in Fig. 6) was then computed as the total success rate from both of those tests. As can be seen in Fig. 6, there was a relationship between membership of roles R1–R3 and PITI performance, although it was not as clear as it was for the previous case of continuous test performance.

A greater number of pupils in the R3 role scored better than the median PITI rate (indicated by the orange horizontal line). Most of the R2 group performed close to the median. Although there were four *passengers* who scored better than the median, the remaining five obtained the worst scores.

4.1.4 Relationship between normalized learning gain, pupils' typical roles during group discussions and willingness to change answers

In general, it was nearly impossible to determine pupils' true performance from the step of individual thinking over ConcepTests or during group discussions. For this reason, *pupils' willingness to change their answers* (simply put, *changeability*) was defined as the ratio of the total sum of answers that the individual changed after group discussions to the total number of group discussions that he or she attended.

Fig. 7 shows the relationship between a membership to a concrete role and changeability c of individuals. The greatest changeability $c = 0.48 \pm 0.09$ involved passengers followed by standard discussers with

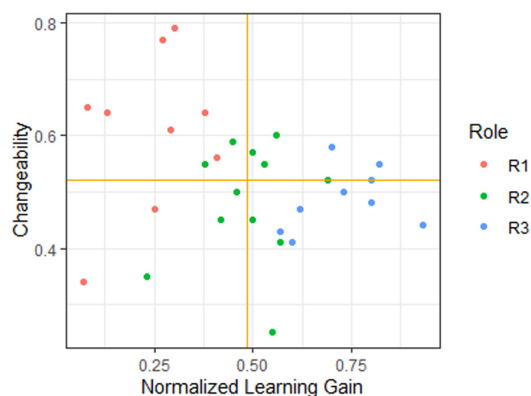


Fig. 7: Relationship between changeability and normalized learning gain

$c = 0.34 \pm 0.08$, then advisors together with overwhelming speakers (R3) with $c = 0.35 \pm 0.07$. There was a statistically significant difference between passengers and R3 ($p < 0.004$) and between passengers and standard discussers ($p < 0.027$) confirmed by the sequence of t-tests. The difference between R2 and R3 ($p > 0.592$) was clearly not significant. All combined, high willingness to change answers is one of the characteristic attributes of passengers.

Together with findings from 4.1.2 and 4.1.3, the high changeability of passengers confirms the fact that there could be pupils (passengers) who follow their dominant peers (overwhelming speakers) through group discussions (Lucas, 2009; Michinov et al., 2015). This is not fruitful for the passengers (because of low cognitive activity), therefore these pupils achieved a low normalized learning gain (Hake, 1998; Crouch & Mazur, 2001; Michinov et al., 2015).

4.1.5 Roles R1–R3 from the viewpoint of pupils' activity during group discussions and their mathematical skills perceived by the class

Although membership of individuals to the concrete roles R1–R3 was discussed with the reflection group, there was a need to confirm it once more within the entire class, as there were two pupils, marked in Fig. 8 as OS01 (overwhelming speaker 01) and PA05 (passenger 05), who may have belonged to another role in the opinion of many pupils in the reflection group (this two pupils may acted differently if group discussions were observed by the researcher). Concretely, OS01 should belong to standard discussers and PA05 to overwhelming speakers. (Please note that their initially indicated roles remain unchanged within this article because of their medium learning gains and researcher's opinion to their role, although the latter titles are probably more suitable.)

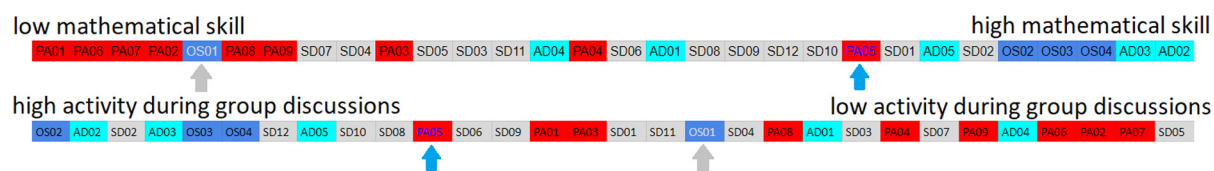


Fig. 8: Pupils mean marks of mathematical skill and activity during group discussions

In order to obtain the opinion of the whole class, the pupils were asked to evaluate their peers in terms of their mathematical skill (on a 7-point scale from low to high) and activity during group discussions (another 7-point scale from high to low). These two scales were included in the extended Q1 questionnaire that was assigned to the class after the first six months of the 2019/2020 school year (i.e., after fourteen months of peer instruction).

Please note that in the Czech Republic it is not standard for pupils to be asked by their teacher to evaluate each other in any way. Therefore, sixteen months was established as the “waiting period” to let the pupils and teacher get to know each other better in order for more pupils to feel comfortable sharing opinions of their peers with the teacher.

Even after sixteen months, four pupils refused to evaluate their peers and many others used only the positive half of both evaluation scales or 3 to 5 points from the scales. For this reason, each pupil's evaluation of his or her all peers was then recounted individually in sense of the following three marks:

- -1 for the worst rated third (by concrete individual),
- 0 for the medium rated third (by concrete individual),
- +1 for the highest rated third (by concrete individual).

All marks for perceived mathematical skill and perceived activity during group discussion were then assigned to the corresponding individuals (two mean marks for every pupil). The first line of Fig. 8 shows a sequence of pupils by mean marks of perceived mathematical skill from lowest to highest. Similarly, the second line of Fig. 8 shows another sequence of pupils ranked by mean marks of perceived activity during group discussions from highest to lowest. Passengers (red in Fig. 8) were perceived by their peers as the ones with the lowest mathematical skill and lowest activity during group discussions. Overwhelming speakers and advisors were exactly opposite to passengers with the highest perceived mathematical skill and activity during group discussions. Standard discussers were evaluated as medium on both scales.

It could be said that individual memberships to roles R1–R3 were confirmed by the class, although there were two pupils who were evaluated differently from other members who shared their role. The overall evaluation of these pupils agreed more with the opinion of doubtful members of the reflection group than with the earlier membership ascribed by the researcher himself.

4.2 Effectiveness of added pair discussion step in context of roles R1–R3

It was already mentioned in the methodology section that after eight months of teaching by classic *peer instruction*, the pair discussion step was added before the group discussion step. This modification was designed mainly for two reasons: to solve the problem of the inactivity of passengers during group discussions over ConcepTests, and to give pupils additional preparation time for group discussions. Modified peer instruction was then implemented at the beginning of the 2019/2020 year and was used until the school year unexpectedly ended in March because of coronavirus.

After six months of modified peer instruction the added pair discussion step was discussed within the reflection group. Based on the obtained feedback, the pair discussion step was generally perceived as positive and useful. Maybe popular amongst the pupils, however, was there any difference in effectiveness? To answer this question, the pre/post-test T3 was assigned a month before and a month after similarity was taught in the class as was described in the methodology section.

4.2.1 Average normalized learning gain of the whole class for the second time

Once again, on average there was a medium, nearly high, normalized learning gain (Hake, 1998) for the whole class $\langle g \rangle = 0.65 \pm 0.32$. This learning gain was higher than the previous one for the T1 test ($\langle g \rangle = 0.49 \pm 0.22$) and the difference was statistically significant ($p < 0.025$).

4.2.2 Relationship between normalized learning gain and concrete role membership

For the T2 test, passengers achieved a medium normalized learning gain $\langle g \rangle = 0.59 \pm 0.32$ which was valued as statistically greater than the one previously obtained $\langle g \rangle = 0.22 \pm 0.11$ ($U = 11$ against $U_{001} = 11$). Standards discussers achieved again a medium normalized learning gain $\langle g \rangle = 0.64 \pm 0.36$, but this value was not significantly greater than before $\langle g \rangle = 0.49 \pm 0.11$ ($U = 57$ against $U_{05} = 37$). Finally, overwhelming speakers with advisors achieved once more a high normalized learning gain $\langle g \rangle = 0.71 \pm 0.27$ which was quite close to the previous one $\langle g \rangle = 0.75 \pm 0.12$ ($U = 40$ against $U_{05} = 17$). The difference between each pair of roles (R1–R3) was now found statistically insignificant ($p > 0.05$; $U = 31$ against $U_{05} = 17$ for R1 against R3, $U = 49$ against $U_{05} = 26$ for R1 against R2, $U = 51$ against $U_{05} = 26$ for R2 against R3). In other words, there was a great improvement in normalized learning gains and passengers were the most improved in this category.

Results in sections 4.2.1 and 4.2.2 imply that the added pair discussion step may raise the effectiveness of peer instruction in sense of normalized learning gains because it grants pupils an additional opportunity to speak and therefore raises cognitive activity, especially in the case of passengers.

4.3 Changes in pupils' attitudes towards mathematics

The data and the conclusions contained in section 4.3 were carefully discussed with experts in the field and the creator of the Q1 questionnaire, I. Pavelková, during the second major reflection.

4.3.1 Perceived popularity and difficulty of mathematics across time

To track the development of pupils' attitudes across time, the extended questionnaire Q1 was assigned to the class at the beginning of the 2018/2019 school year (EC00), after two months of classic teaching (EC01), at the end of the 2018/2019 school year (EC02), and in the first half of the 2019/2020 school year (EC03). The questionnaire was composed of eight 5-point Likert scales as described in section 3.1.

The obtained results for items: *popularity*, *difficulty* and *my motivation* can be viewed in Tab. 2, 3, 4, in that order. In addition to the data of the class, reference standards for Q1 from (Pavelková & Hrabal, 1988) for the eighth grade and grammar schools are also included in each of the tables.

Tab. 2: Perceived popularity of mathematics

	Mean	Lower quartile	Upper quartile
Standards for eighth grade	2.9	2.6	3.3
Standards for grammar schools	2.8	2.5	3.0
EC00	2.1	2.0	3.0
EC01	2.4	2.0	3.0
EC02	2.5	2.0	3.0
EC03	2.3	2.0	3.0

In Tab. 2, where 1 means very popular and 5 is very unpopular, it can be seen that popularity of mathematics within the class remained constant across time.

In Tab. 3, where 1 means “very difficult” and 5 on the other hand is “quite easy”, the difficulty of mathematics within the class rose steeply during the first year of peer instruction. It rose even beyond the standards for both eighth graders and grammar schools. However, it decreased to standard level after another six months of modified peer instruction. This trend can be explained by several reasons. In the middle of the 2019/2020 school year the class was more used to peer instruction. Secondly, because of the obvious pressure, there were fewer ConcepTests in continuous tests during the last six months of modified peer instruction based on the first major reflection. Concretely, there were two ConcepTests out of five tasks in every continuous test during the 2018/2019 school year, and then there was only one ConcepTest out of five tasks in the 2019/2020 school year.

Tab. 3: Perceived difficulty of mathematics

	Mean	Lower quartile	Upper quartile
Standards for eighth grade	2.5	2.2	3.7
Standards for grammar schools	2.4	2.1	3.8
EC00	3.1	3.0	3.0
EC01	2.7	2.0	3.0
EC02	1.8	1.0	2.0
EC03	2.4	2.0	3.0

In Tab. 4, where 1 means that a pupil is very motivated and 5 means he or she is quite unmotivated, it is clear that pupils’ motivation toward mathematics remained constant.

Tab. 4: Perceived motivation within mathematics

	Mean	Lower quartile	Upper quartile
Standards for eighth grade	2.5	2.2	2.8
Standards for grammar schools	2.8	2.5	3.0
EC00	2.5	2.0	3.0
EC01	2.8	2.0	3.0
EC02	2.8	2.0	3.0
EC03	2.6	2.0	3.0

These results indicate that peer instruction may improve in good constellation pupils’ attitudes towards mathematics because of following reasons:

- “In principle, the medium degree of difficulty of the school subject (compared to the reference standard) can be considered optimal. Its increase is positive only if it does not reduce the motivation of pupils and the popularity of the subject and if the requirements of the teacher are valuable in terms of pupils’ development.” (Pavelková & Hrabal, 1988, p. 27)
- “Our pupils’ relationship to mathematics deteriorates during schooling. A more significant decline occurs at the beginning of the second stage of primary school. In secondary school, this relationship will not change on average, but this trend continues.” (Chvál, 2013, p. 68)
- According to oral statement of I. Pavelková during the second major reflection, difficulty plays a key role in the issue of pupils’ attitudes towards the subject.

In case of the class, the detected significant decrease mentioned by Chvál (2013) was not detected; pupils’ motivation and the perceived popularity of mathematics remained constant within the class although the perceived difficulty of the subjects was significantly rising.

4.3.2 Pupils’ attitudes towards mathematics across time in terms of the semantic differential

It was difficult to track changing attitudes towards mathematics (using only questionnaire Q1) and simultaneously distinguish the roles (R1–R3) because of the low sensitivity of Q1. To solve this problem, questionnaire Q3 was used periodically during the 2018/2020 school year along with the pre/post-questionnaire Q2. In these questionnaires, the perceived good or evil of mathematics was given an average rating on 4 seven-point scales, as described in Section 3.6. The higher the value of the dimension of evaluation, the better mathematics is perceived.

Although mathematics was evaluated in the pre/post-questionnaire Q2 in the context of another 15 concepts and in the Q3 questionnaire separately, it is possible to at least observe a trend in the evaluation dimension. Based on the pre/post-Q2 questionnaire the dimension of evaluation from the class equalled 5.59 and then 5.33. Although there was a slight decrease in this dimension, both obtained values were still higher than the standard for elementary schools (4.76) and grammar schools (4.12).

Fig. 9 illustrates trends in the dimensions of evaluation for the whole class and for the roles R1–R3 from the pre-questionnaire Q2 at the beginning of the 2018/2019 school year to the post-questionnaire Q2 at the end of the 2018/2019 school year. It is clear that passengers perceived mathematics as worse the more there was abstraction in it and that their attitudes towards mathematics was worse than any other of the other pupils. Once again, passengers were shown to be the biggest obstacle to overcome in order to effectively implement peer instruction and improve attitudes towards mathematics.

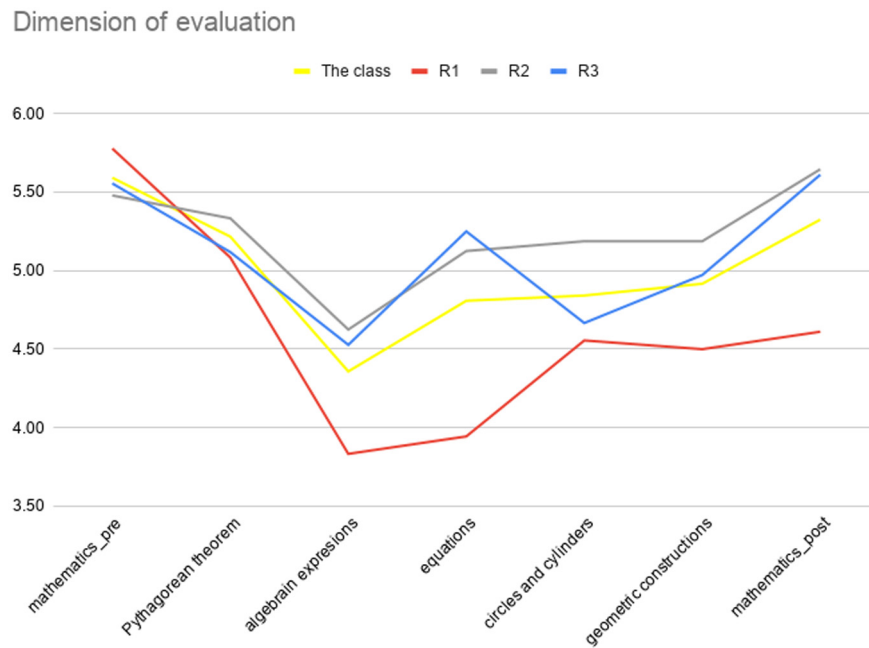


Fig. 9: Trends in pupils' attitudes

4.4 The easiest way to improve social skills is to practise them

It has been pointed out that peer instruction effectiveness relies mainly on group discussions over ConcepTests. The scheme of classic peer instruction in Fig. 1 shows that it is good to discuss ConcepTests as the success rate climbed from 35% to 70%. The following ConcepTest was included in the T1 test and was already published by Zdražil (2020).

Example task 03 from the T1 test

Choose the correct statement from a)–d) about pentagons (1) and (2) on the connected picture. Then briefly justify your answer.

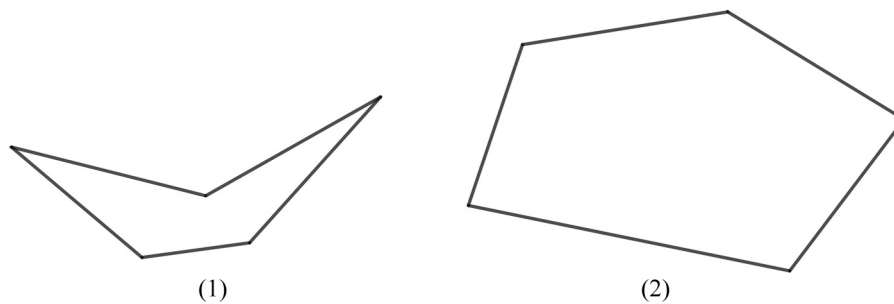


Fig. 10: Connected picture

- a) The sum of the interior angles of pentagon (1) is greater than the sum of the interior angles of pentagon (2).
- b) **The sum of the interior angles of pentagon (1) is equal to the sum of the interior angles of pentagon (2).**
- c) The sum of the interior angles of pentagon (1) is less than the sum of the interior angles of pentagon (2).
- d) It is impossible to decide which option from (a)–(c) is correct without concrete measuring.

Tab. 5 (which was not published) shows pupils' typical responses to this task. It is important to note that although the majority of pupils chose the correct option, b), not many of them justified it correctly (semi-peer instruction eighth grade class mentioned in Tab. 5 was taught by peer instruction in combination with classic teaching during pilot study in the 2017/2018 school year). In other words, pupils could think incorrectly even if they chose the correct answer; therefore, discussion is needed to prove their thought processes. In other words, in an elementary school environment it could be fruitful to let pupils discuss every ConcepTest even in the case of a high initial success rate during the first round of voting.

Tab. 5: Pupils' answers 603 to the example task 03 from the T1 test

	Three classic eighth grade classes (<i>N</i> = 86)	A semi-peer instruction eighth grade class (<i>N</i> = 30)	The peer instruction class (<i>N</i> = 30)
"Each pentagon consists of three triangles."	2% (2)	20% (6)	47% (14)
"Each pentagon consists of a triangle and a quadrilateral."	0% (0)	0% (0)	7% (2)
"The sums are equal because there must be some rule."	30% (26)	30% (9)	33% (10)
"Both sums are equal to 360 degrees."	10% (9)	37% (11)	10% (3)
"The second pentagon has more obtuse angles."	15% (13)	13% (4)	3% (1)
"The first pentagon has a non-convex angle."	5% (4)	0% (0)	0% (0)
"It could not be decided without measuring."	3% (3)	0% (0)	0% (0)
No answer	34% (29)	0% (0)	0% (0)

There is another reason to let pupils discuss in groups over even the simplest questions. A study by Chen et al. (2005) previously warned of pupils' insufficient social skills to maintain a fruitful group discussion over ConcepTest in physics. The easiest way to improve social skills is to practise them.

In Fig. 11 and 12, the response process (in terms of accuracy) is shown for a sequence of two consecutive ConcepTests targeting the same concept with graded difficulty. Q1 is the answer to the first question in the first vote, Q1' the answer to the first question in revised voting, and Q2 is the answer from the first vote of the follow-up ConcepTest.

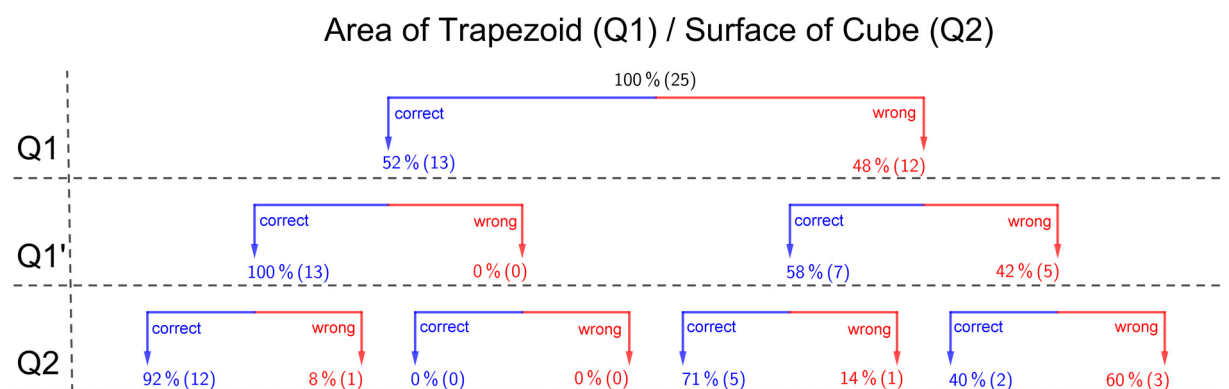


Fig. 11: Schema of response process for sequence of two ConcepTests

Circles (Q1) / Perimeter of Trapezoid (Q2)

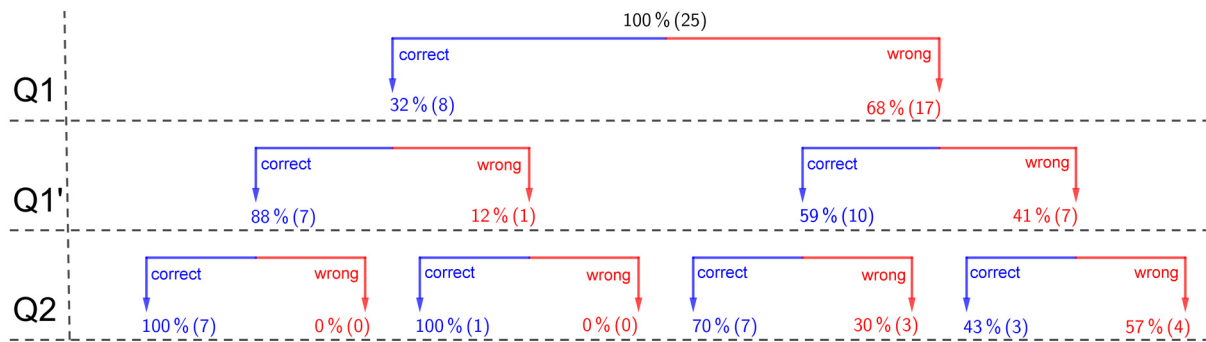


Fig. 12: Schema of response process for sequence of two ConceptTests

It is evident that in both cases approximately 70% of pupils who corrected their answer (if it were wrong the first time) were able to use the knowledge they gained in the discussion to correctly determine the answer to the follow-up question. It can also be stated that most pupils who kept the correct answer even after the group discussion answered correctly even in the case of a follow-up task. These findings are in agreement with Smith et al. (2009).

In other words, yes, group discussion can lead pupils from an initially wrong answer to a true understanding of the problem – not just a mere copy of the most common option and, yes, it is desirable to let pupils discuss every ConceptTest even in the case of a high initial success rate during the first round of voting.

5 Discussion

Studies showed (Hake, 1998; Mazur & Crouch, 2001; Michinov et al., 2015) that for peer instruction courses, a medium value (Hake, 1998) of normalized learning gain is typical. Although there was a statistically significant improvement ($p < 0.025$) between two measurements, a medium value of normalized learning gain was achieved twice by the class (see Tab. 6).

Tab. 6: Overall table

	The class	Passengers	Standard discussers	Overwhelming speakers and Advisors
Test T1	$\langle g \rangle = 0.49 \pm 0.22$	$\langle g \rangle = 0.22 \pm 0.11$	$\langle g \rangle = 0.49 \pm 0.11$	$\langle g \rangle = 0.75 \pm 0.12$
Test T2	$\langle g \rangle = 0.65 \pm 0.32$	$\langle g \rangle = 0.59 \pm 0.32$	$\langle g \rangle = 0.64 \pm 0.36$	$\langle g \rangle = 0.71 \pm 0.27$
Changeability	$c = 0.40 \pm 0.12$	$c = 0.48 \pm 0.09$	$c = 0.34 \pm 0.08$	$c = 0.35 \pm 0.07$

A relationship was found between pupils' typical roles during group discussions and normalized learning gain. Four typical roles of pupils were identified based on sixth months of observation and feedback from the reflexive group – passengers (R1), standard discussers (R2) and advisors along with overwhelming speakers (R3). Members of each of these groups were then characterized by their effort to constructively solve the ConceptTests posed and their need to dominate discussions.

Passengers were typically the most willing to change their answers. They achieved low normalized learning from the T1 test and a medium normalized learning gain from the T2 test. The improvement for passengers between these two tests was statistically significant ($p < 0.001$) and therefore implies that the addition of the pair discussion step could boost passengers' activity. The improvement between T1 and T2 test was not so significant for standard discussers nor was it significant for advisors and overwhelming speakers who achieved in that order a medium and a high normalized learning gain in the case of both tests.

A relationship was also established between group membership and pupils' test performance. The presented findings connected with normalized learning gains are all in agreement with other studies (Lucas, 2009; Michinov et al., 2015) that group discussions could be led by dominant group members at the expense of their partners. These dominant members (overwhelming speakers) profited the most from the group discussions along with advisors.

The perceived difficulty of mathematics rose significantly in the class after implementation of peer instruction; however, there was not a significant decrease detected in the perceived popularity of mathematics mentioned by Chval (2013) nor was there a decrease of pupils' motivation in mathematics. However, it was clear that passengers' attitudes towards mathematics were worse than those of any other role and that their attitudes decreased with greater abstraction in mathematics.

It was shown that pupils could think incorrectly even if they chose the correct answer; therefore, discussion is needed to prove their thought processes. In other words, in an elementary school environment it could be fruitful to let pupils discuss every ConcepTest even in the case of a high initial success rate during the first round of voting. In agreement with Smith et al. (2009), it was also demonstrated that approximately 70% of pupils who corrected their answer (if it were wrong the first time) were able to use the knowledge they gained in the discussion to correctly determine the answer to the follow-up question.

6 Limitations

The quality of the action research depends on the quality of the story being told. Although the main mission of action research is to achieve positive change for a specific group in a specific constellation, there is also a demand to find a general model, theory, or universal modification (Mertler, 2019). The general conclusions of the research presented in this article came from the data collected on a relatively small sample of 30 pupils of a single eighth grade class. The researcher himself was personally involved as the math teacher of the class.

An attitude evaluation depends on the current situation within the class and the mood of pupils. Therefore, attitude was repeatedly measured by different questionnaires (Q1, Q2 and Q3).

Pupils were initially assigned to the roles R1–R3 by the researcher himself. To minimize errors of subjectivity, the reflection group were asked to confirm or alter proposed memberships of every pupil. Because of remaining disagreement, the pupils of the whole class were asked to evaluate each other in terms of their perceived mathematical skill and activity during group discussions. After this, there were only two pupils whose role during group discussions remained unclear.

Values of dimension of evaluation computed for taught topics did not have the same telling value as values computed in context of the whole questionnaire (Q2) and were therefore only used to track trends in pupils' attitudes towards mathematics (mathematics is evaluated in context of other fifteen concepts in Q2 – not separately as in Q3).

There was far more time between the pre/post-tests T1 than the pre/post-tests T2. In addition, the T1 test targeted more familiar mathematics topics it also contained more previously tested concepts. For this reason, the normalized learning gain of the T2 test is less objective than the normalized learning gain of the T1 test. However, the T2 test was only intended to track changes in differences between discussion roles to determine learning gains after the added pair discussion step was applied to peer instruction.

7 Conclusion

Sixteen months of action research on the implementation of peer instruction in elementary school mathematics has shown that it is possible:

- a) to satisfy assumptions of peer instruction in lower secondary mathematics education (e.g., sufficient social skills of pupils to maintain fruitful group discussions) and
- b) to achieve similar outcomes that were promised by research studies of peer instruction in physics or university level education (e.g., profound conceptual understanding, significant normalized learning gains, improved attitudes towards mathematics, etc.).

However, the concrete constellation of a class and the natural needs of its pupils raised obstacles that needed to be addressed through adequate thoughtful modification of peer instruction. Although it is not clear what exactly leads pupils to identify themselves with one of the four typical roles during group discussions, it is clear that there is a strong relationship between membership to a certain role and the positive impact of peer instruction.

Passengers are pupils who benefits the least from peer instruction (their attitudes towards mathematics decreases and their learning gains are quite low) and must therefore be identified and involved in group discussions as much as possible.

A possible solution presented in this article is the paired discussion step which can be added before the group discussion step to grant pupils more time and an additional opportunity to discuss ConcepTests (even the easiest one in order to practise social skills).

The presence of passengers is independent from the teaching method used although they are easier to detect in an active learning environment (Garcia-Souto, 2020). Our research also implies that a possible cure to “the passengers’ problem” could be provided by feedback from a reflection group. This feedback is tailored to a specific constellation and is therefore very effective in this concrete situation.

Acknowledgment

This work was financially supported by the project GA UK 680119 and by Charles University Research Centre program No. UNCE/HUM/024. In accordance with ethical requirements for pedagogical research, the data and results contained herein are published with the informed consent of the legal representatives of the minor participants.

I would like to thank my pupils, especially the members of the reflection group, without whose direct participation, opinions, and comments there would be no results suitable for publication. I would also like to thank Jarmila Robová and Isabela Pavelková for their critical notes and comments towards results published in this article.

References

- Balta, N., Michinov, N., Balyimez, S., & Ayaz, F. M. (2017). A meta-analysis of the effect of peer instruction on learning gain: Identification of informational and cultural moderators. *International Journal of Educational Research*, 86, 66–77. <https://doi.org/10.1016/j.ijer.2017.08.009>
- Chen, Y. F., Liu, C. C., Yu, M. H., Chang, S. B., Lu, Y. C., & Chan, T. W. (2005). Elementary science classroom learning with wireless response devices implementing active and experiential learning. In *IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'05)* (pp. 96–103). IEEE. <https://doi.org/10.1109/WMTE.2005.22>
- Chien, Y. T., Chang, Y. H., & Chang, C. Y. (2016). Do we click in the right way? A meta-analytic review of clicker-integrated instruction. *Educational Research Review*, 17, 1–18. <https://doi.org/10.1016/j.edurev.2015.10.003>
- Chvál, M. (2013). Změna postojů českých žáků k matematice během školní docházky [Changing attitudes of Czech pupils towards mathematics during school attendance]. *Orbis scholae*, 7(3), 49–71. <https://doi.org/10.14712/23363177.2015.13>
- Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American journal of physics*, 69(9), 970–977. <https://doi.org/10.1119/1.1374249>
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. <https://doi.org/10.1073/pnas.1319030111>
- Garcia-Souto, M. P. (2020). Assessing and giving feedback to students working in teams: A staff perspective of the IPAC assessment methodology, In L. G. Chova, A. L. Martínez, & I. C. Torres (Eds.), *INTED2020 Proceedings* (pp. 4416–4416). IATED. <https://doi.org/10.21125/inted.2020>
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74. <https://doi.org/10.1119/1.18809>
- Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, 53(11), 1043–1055. <https://doi.org/10.1119/1.14030>
- Lucas, A. (2009). Using peer instruction and i-clickers to enhance student participation in calculus, *PRIMUS*, 19(3), 219–231. <https://doi.org/10.1080/10511970701643970>
- Mazur, E. (1997). *Peer instruction: A user's manual*. Prentice Hall.
- Mazur, E. (2009). Farewell, lecture. *Science*, 323(5910), 50–51. <https://doi.org/10.1126/science.1168927>
- Mertler, C. A. (2019). *Action research: Improving schools and empowering educators*. SAGE Publications, Incorporated.
- Michinov, N., Morice, J., & Ferrières, V. (2015). A step further in Peer Instruction: Using the Stepladder technique to improve learning. *Computers & Education*, 91, 1–13. <https://doi.org/10.1016/j.compedu.2015.09.007>
- Olpak, Y. Z., Baltaci, S., & Arican, M. (2018). Investigating the effects of peer instruction on preservice mathematics teachers' achievements in statistics and probability. *Education and Information Technologies*, 23(6), 2323–2340. <https://doi.org/10.1007/s10639-018-9717-3>

- Pavelková, I., & Hrabal, V. (1988). *Jaký jsem učitel?: Metody získávání poznatků o vlastní vzdělávací činnosti* [What kind of teacher am I?: Methods of gaining knowledge about my own educational activities]. Státní pedagogické nakladatelství.
- Pilzer, S. (2001). Peer instruction in physics and mathematics. *PRIMUS*, 11(2), 185–192. <https://doi.org/10.1080/10511970108965987>
- Pöschl, R. (2005). *Vnímání významu matematiky a fyziky středoškolskými studenty* [Perception of the meaning of mathematics and physics by high school students]. [Diploma thesis. Charles University.] https://kdf.mff.cuni.cz/vyzkum/materialy/vnimani_vyznamu_M_a_F.pdf
- Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, 93(3), 223–231. <https://doi.org/10.1002/j.2168-9830.2004.tb00809.x>
- Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, 323(5910), 122–124. <https://doi.org/10.1126/science.1165919>
- Vickrey, T., Rosploch, K., Rahmanian, R., Pilarz, M., & Stains, M. (2015). Based implementation of peer instruction: A literature review. *CBE – Life Sciences Education*, 14(1), es3. <https://doi.org/10.1187/cbe.14-11-0198>
- Vondrová, N., Rendl, M., Havlíčková, R., Hříbková, L., Páchová, A., & Žalská, J. (2015). *Kritická místa matematiky základní školy v řešeních žáků* [Critical points of elementary school mathematics in pupils' solutions]. Karolinum.
- Zadrazil, T. (2020). Relationship between normalized learning gains of individuals and their typical role in group discussions with peers. In L. G. Chova, A. L. Martínez, & I. C. Torres (Eds.), *INTED2020 Proceedings* (pp. 3218–3227). IATED. <https://doi.org/10.21125/inted.2020>