MODESTUM

OPEN ACCESS

EURASIA Journal of Mathematics, Science and Technology Education

ISSN: 1305-8223 (online) 1305-8215 (print)

2018 14(4):1111-1132

DOI: 10.29333/ejmste/81807

Class Tournament as an Assessment Method in Physics Courses: A Pilot Study

Daniel Dziob [1,2*], Lukasz Kwiatkowski [3], Dagmara Sokolowska [1]

[1] Smoluchowski Institute of Physics, Jagiellonian University, Krakow, POLAND

[2] Department of Biophysics, Jagiellonian University Medical College, Kracow, POLAND

[3] Department of Econometrics and Operations Research, Cracow University of Economics, Krakow, POLAND

Received 23 August 2017 ▪ Revised 9 December 2017 ▪ Accepted 11 December 2017

ABSTRACT

Testing knowledge is an integral part of a summative assessment at schools. It can be performed in many different ways. In this study we propose assessment of physics knowledge by using a class tournament approach. Prior to a statistical analysis of the results obtained over a tournament organized in one of Polish high schools, all its specifics are discussed at length, including the types of questions assigned, as well as additional self- and peer-evaluation questionnaires, constituting an integral part of the tournament. The impact of the tournament upon student improvement is examined by confronting the results of a post-test with pre-tournament students' achievements reflected in scores earned in former, tests written by the students in experimental group and their colleagues from control group. We also present some of students' and teachers' feedback on the idea of a tournament as a tool of assessment. Both the analysis of the tournament results and the students' and teachers' opinions point to at least several benefits of our approach.

Keywords: team work, cooperative learning, collaborative testing, K-12 physics, assessment methods, assessment as learning

INTRODUCTION

Testing knowledge is an integral part of educational assessment, the latter being a process of documenting content knowledge, skills, attitudes and beliefs, usually focused on an individual learner or a learning community as a whole. The most popular distinction in types of assessment is founded upon the difference between formative and summative assessment (Black & Wiliam, 1998; Garriso & Ehringhaus, 2007; Harlen & James, 1997; McTighe & O'Connor, 2005; Wiliam & Black, 1996 and references therein). In general, the formative assessment is carried out throughout a unit (course, project), whereas the summative one - at the end of a unit (course, project) (Harlen & James, 1997; McTighe & O'Connor, 2005). Some authors seem to distinguish between these types of assessment arguing that the summative assessment is "assessment *of* learning", while the formative one is "assessment *for* learning" (Black et al., 2004; Earl, 2004; Looney, 2011; Taras, 2005).

Focusing on the summative assessment (SA), we can point to three major criteria defining it: i) SA is used to determine whether students have learned what they were expected to learn (Earl, 2004; Harlen & James, 1997; Torrance & Pryor, 1998); ii) SA is carried out at the end of a specific teaching period, and therefore it is generally of an evaluative nature, rather than diagnostic one (Earl, 2004; Harlen & James, 1997; Torrance & Pryor, 1998); iii) SA results are often recorded as scores or grades that are then factored into a student permanent academic record (Biggs, 1998; Bloom et al., 1971; Earl, 2004).

Summative assessment can be performed in many ways (Black et al., 2010, 2011; McTighe & O'Connor, 2005; Scriven, 1967), though written tests are still the most prevalent (Talanquer et al., 2015; Taras, 2009; Vercellati et al., 2013). However, in different fields a few researchers have come up with an idea of carrying out assessment in some alternative manners (Dochy et al., 1999; Rebello, 2011; Schuwirth & Vleuten, 2004). These include, in particular, different forms of a written test, extensively described and compared in the literature, such as free- and multiple-response tests (Wilcox & Pollock, 2014), concept tests (such as the Test of Understanding Graphs in Kinematics

✉ daniel.dziob@uj.edu.pl **(*Correspondence)** ✉ kwiatkol@uek.krakow.pl ✉ ufdsokol@cyf-kr.edu.pl

Contribution of this paper to the literature

- The study suggests and appraises a new method for evaluation, combining summative assessment with elements of formative one in a form of a tournament game taken in groups and being an example of "assessment as learning".
- The tournament is very flexible for inclusion of theoretical and practical tasks in different formats and may also comprise self- and peer-assessment questionnaires, as well as evaluation of attitude, motivation and interest.
- The analysis of the tournament results and students' opinions about the implementation in physics classes points out academic benefits for students and equal opportunities of improvement both for low- and high-performers.

(Maries & Singh, 2013), Force Concept Inventory (Hestenes et al., 1992) or Brief electricity and magnetism assessment (Ding et al., 2006) and others (Hitt et al., 2014; Wilcox et al., 2015), constructed-response tests (Slepkov & Shiell, 2014), essay tests (Kruglak, 1955), laboratory skills tests (Doran et al., 1993) and others. Also, many modifications and extensions of these tests have already been proposed in the literature, improving upon their original form (Ding, 2014; Docktor et al., 2015; Wooten et al., 2014; Zwolak & Manogue, 2015). On the other hand, some authors propose to blend formative and summative assessment techniques. According to (Wininger, 20015), such a combination, named "formative summative assessment", entails reviewing exams with students so that they get feedback about their comprehension of concepts. Nowadays, we can find different proposals of combining these two types of assessments (Fakcharoenpohl & Stelzer, 2014; Pawl et al., 2013; Wilcox & Pollock, 2015; Yu & Li, 2014), and the boundaries between them become more and more vague. One example of such an approach is "collaborative testing" – an idea of giving students the opportunity for working in groups during an exam (Guest & Murphy, 2000), at the end of an individual exam (Lusk & Conklin, 2003) or, more often, after the first, but before the second exam taken individually (Cortright et al., 2003; Ives, 2014; Rao et al., 2002) (the last two are sometimes named "two-stage exams"). Research has shown that there are many benefits of utilizing collaborative testing as a constructivist learning method. They are described in detail in (Duane & Satre, 2014; Gilley & Clarkston, 2014; Kapitanoff, 2009) and references therein.

In our study, we use a tournament – a competitive game between groups in the classroom – as a tool for summative assessment with formative evaluation elements. On the one hand, applying the mechanics of a game to make the process more appealing can be considered a gamification (Apostol et al., 2013; Deterding et al., 2011). Although the idea of introducing games in teaching is not new (Ifenthaler et al., 2012 and references therein; Moncada & Moncada, 2014), the very term of gamification has been coined only a few years ago, and has been gaining more and more popularity since then (Dicheva et al., 2015; Sadler et al., 2013; Sung & Hwang, 2013). The benefits of gamification (or, in more broad terms, game-based learning (e.g. Ifenthaler et al., 2012)) in the educational context are widely described in the literature (Banfield & Wilkerson, 2014; Dicheva et al., 2015; Hanus & Fox, 2015; Seaborn & Fels, 2015; Sung & Hwang, 2013).

Moreover, a tournament can be also considered as a kind of "collaborative testing", but unlike the forms mentioned above, we first conduct a group exam (distinguishing individual students' marks through their involvement and contribution in the group work), and, secondly, provide a control, individual test (only for the purpose of research, not influencing students' final marks). Following (Earl, 2004), where also the idea of "assessment *as* learning" is introduced (and in which student self-assessment, and, thereby, self-motivation are brought into focus (Hickey et al., 2012)), we design an alternative form of testing knowledge, combining the assessment with learning and a game at the same time. And by learning we mean not only the subject matter itself, but also acquiring and developing other skills, as well as stimulating positive, both intra- and interpersonal dispositions, such as self-motivation, language skills and group work in the form of cooperative learning (Jolliffe, 2007; Kagan, 1990; Slavin, 2000).

RESEARCH DESIGN

In this section we provide details on the tournament itself, including its organization, questions assigned and relevant evaluation procedures.

Tournament Schema

The tournament was performed in a high school in Wolbrom (a small town of ca. 9 000 inhabitants, in the South of Poland), and it involved 30 students in their final class (K-12). At the time the class had just accomplished a 22-hour course on electricity.

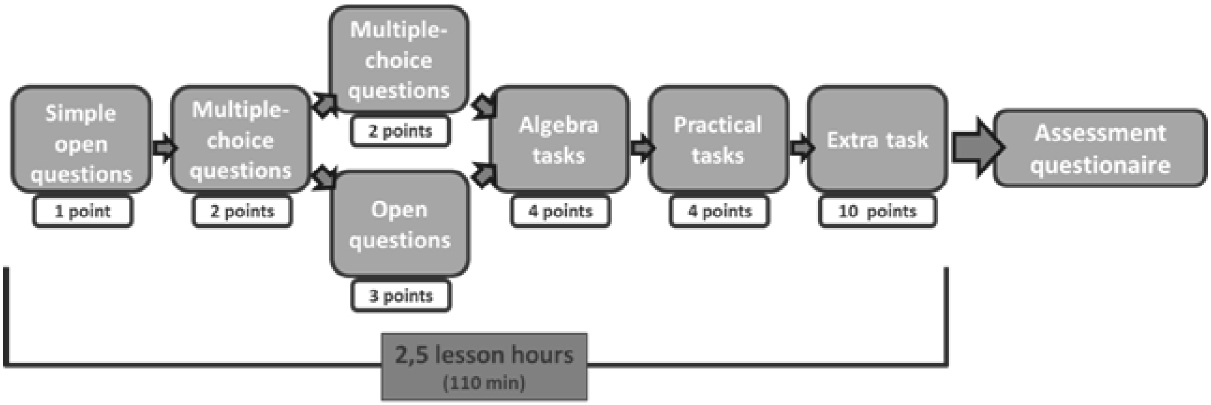


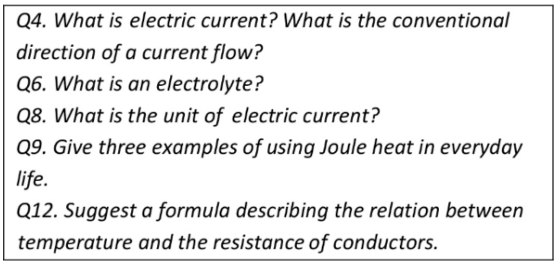
Figure 1. Tournament testing sequence

At the beginning of the actual event (lasting for 2.5 lesson hours), the students were divided randomly into 5 groups, by lottery, drawing out lots with names of fairytale heroes upon entering the classroom. After drawing a card with the name of a hero, every student held a seat at one of the five tables, each grouping heroes of one of five fairy tales. Then, actual tournament started. **Figure 1** presents the scheme of the entire process, which was designed on the basis of the former experience with utilizing different assessment formats by both the researchers and the teachers teaching in the school where the tournament was implemented. The overriding goal while formulating the scheme was to make the assessment more holistic by including tasks oriented not solely on the content-matter itself, as it often happens in typical tests, but also ones related to everyday life, and allowing the teacher to evaluate students' experimental skills as well.

The tournament began with open questions and multiple-choice questions with an increasing level of difficulty, and, therefore, an increasing number of available points (which were the reward for every correct answer). Further, calculus and some practical tasks were assigned. Finally, all groups faced an extra, common task, with the elements of time competition (the winning team was the first one that rang the bell and provided the correct solution to the problem faced by all teams at the same time). At the end, the students were asked to fill in a special self- and peer-assessment questionnaire. After a week, a post-test was performed. At each stage, the whole process was monitored by two independent teachers (apart from the major teacher of the class), who were responsible for the verification of verity and integrity of the student evaluation. The assistant teachers were not allowed to involve themselves in the tournament itself, with their scope of duties largely limited to overall student supervision and taking notes on the engagement and behavior of the participants. However, their role was also to aid the major teacher in assessing those students, who – at the stage of student self- and peer-evaluation (explained below) – would be found to appraise themselves or their teammates erroneously or unjustifiably.

The first two stages were organized in multiple rounds. The first phase comprised three rounds, and the second – two rounds. At the first three stages, in each round the teams attempted a task in consecutive turns. At the third phase the students faced a choice of undertaking either a 3-point open question or a 2-point multiple-choice one. It was intended to introduce some element of decision-making risk, thereby facilitating students' sense of responsibility for the choices made. The following three stages (i.e. the calculus, practical and extra tasks) were single-rounded and at each of them all teams were challenged with their tasks at the same time. The students were already familiar with the forms of all the assignments, for similar had been administered to them during previous class tests.

In the first three types of questions students from the currently "active" group were required to choose the number of a question, and then the team had the appointed time (respectively 30 seconds, 1 minute or 2 minutes) to deliver the answer. If they did not succeed or their answer was incorrect, other groups could take over the question and score extra points by ringing the bell and providing the correct answer. Allowing for such a possibility was meant to ensure attention and an active interest of each group in the question currently dealt with by any other team. During this part of the tournament, questions were projected onto the wall screen so as to make it available for all teams at the same time. In the calculus and practical tasks all groups worked simultaneously over different, randomly selected problems, received on sheets of paper. For providing the correct solution each group could earn maximally 4 points, and there was no possibility of intercepting unsolved problems by other teams. The practical task score included: 1 point for building a properly working experimental setup, 1.5 points for providing a valid explanation, and the remaining 1.5 points for answering the teacher's question on "*What would happen if…?*" The extra task was the same for all groups, and, again, it was projected on the wall screen so as to make it available to all teams at the same time. The first group which solved the problem won (according to the rule "first-come, first-



Q4. What is electric current? What is the conventional direction of a current flow?

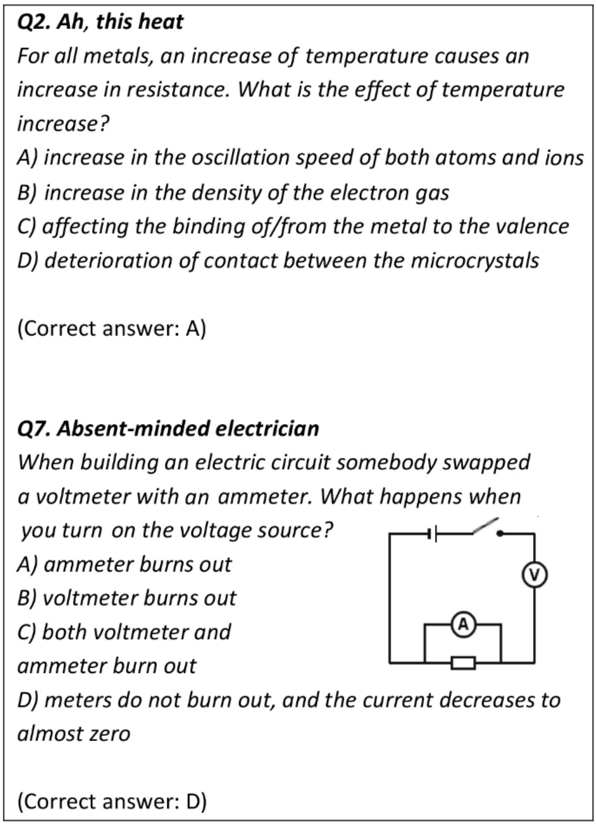
Q6. What is an electrolyte?

Q8. What is the unit of electric current?

Q9. Give three examples of using Joule heat in everyday life.

Q12. Suggest a formula describing the relation between temperature and the resistance of conductors.

Figure 2. Examples of simple open questions (1 point)



Q2. Ah, this heat

For all metals, an increase of temperature causes an increase in resistance. What is the effect of temperature increase?

A) increase in the oscillation speed of both atoms and ions

B) increase in the density of the electron gas

C) affecting the binding of/from the metal to the valence

D) deterioration of contact between the microcrystals

(Correct answer: A)

Q7. Absent-minded electrician

When building an electric circuit somebody swapped a voltmeter with an ammeter. What happens when you turn on the voltage source?

A) ammeter burns out

B) voltmeter burns out

C) both voltmeter and ammeter burn out

D) meters do not burn out, and the current decreases to almost zero

(Correct answer: D)

Figure 3. Examples of multiple-choice questions (2 points)

win"), and scored extra 10 points, which were also added to the maximum number of points possible to obtain by the team.

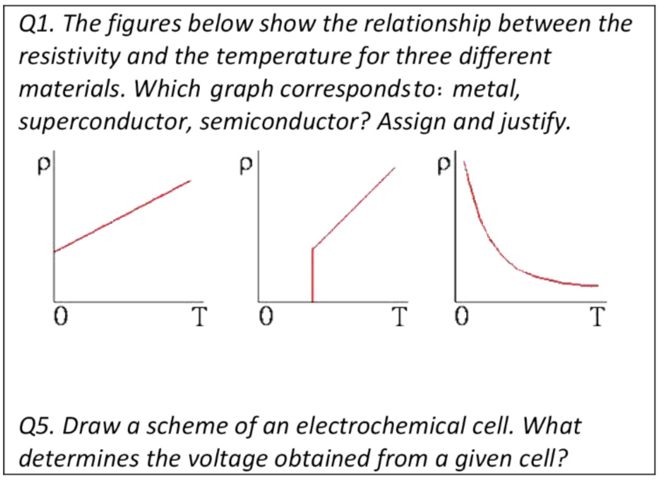
It should be clarified here that, at each stage, the correct solution along with a proper explanation to each question were delivered – either by the contestants (with or without the teacher assistance) or by the teacher himself (in those cases where the students were found incapable of delivering a valid solution on their own). Such a practice served as a means of an immediate clarification and refinement of the students' understanding of the underlying physical concepts.

Simple open questions (1 point)

Open questions, each 1-point worth, were meant "to warm up" the students. The tasks were related to some basic knowledge from the curriculum, requiring the students to provide correct simple formulae, units etc., and also examining their basic context knowledge (see **Figure 2**).

Multiple-choice questions (2 points)

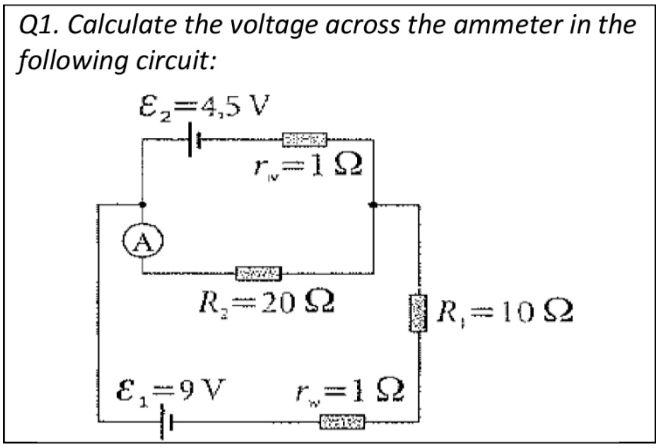
Then, two rounds ensued of multiple-choice scientific reasoning questions (each worth 2 points). The students were requested not only to point out the correct answer, e.g. "C", but also to provide a proper explanation of their choice (see **Figure 3**).



Q1. The figures below show the relationship between the resistivity and the temperature for three different materials. Which graph corresponds to: metal, superconductor, semiconductor? Assign and justify.

Q5. Draw a scheme of an electrochemical cell. What determines the voltage obtained from a given cell?

Figure 4. Examples of open questions (3 points)



Q1. Calculate the voltage across the ammeter in the following circuit:

Figure 5. Example of a calculus task

Open questions (3 points)

At the third stage of the tournament, each group faced a choice between a multiple-choice question worth 2 points and an open question for 3 points. The latter was more challenging, requiring broader knowledge and ability of connecting facts (see **Figure 4**).

Calculus tasks (4 points)

After this part, calculus tasks followed. Each group had to choose a different problem (see **Figure 5** for an example) and was given 10 minutes to provide the correct solution. As previously mentioned, this time all groups worked simultaneously. As a result each team could receive maximally 4 points, with a lower score given upon delivery of either an incomplete or partially faulty solution.

Practical tasks (4 points)

The penultimate challenge was a practical task. Each group had to pick randomly an experimental task on one of the following six themes: galvanization, electrochemical cell, electrolysis of water, Ohm’s law, building a circuit according to an assigned scheme, and voltage measurement in a designated point. Each team was requested to build a proper circuit, carry out the experiment and give valid description and explanation of the phenomenon at hand. All necessary equipment in each case, with some redundant materials mixed in, was available on a table. Then, students had to decide by themselves which objects were indispensable to accomplish the task.

Table 1. Student self-assessment questionnaire

Question	1-6 scale
Were you involved in the work group?	Communication skills
Did you communicate adequately in the group?	
Did you take part in the discussion on the problem?	
Did you take into account the opinions of others?	
Did you prepare for the test beforehand?	Subject matter contribution
Did you take part in solving problems and tasks?	
Did you have sufficient knowledge to solve the issues?	
Did you contribute to the final result of the group?	

Extra task (10 points)

The final and most demanding task was common to all groups. For all groups at the same time, a slide with the "Monstrous maze of resistors" adopted from (Halliday et al., 2001, p.728, task 8) was displayed on the wall screen. The first group which found the solution and gave the correct answer received 10 points.

Assessment questionnaires

After the tournament each student was asked to fill in individually special self- and peer-assessment questionnaires, aimed at evaluating himself/herself and other fellow players from the same group in various aspects. Each of eight questions required allotting a score between 1 and 6. Four of them were related to the "communication skills", whereas the other four were focused on assessing the "subject matter contribution". In **Table 1** we present the self-assessment questionnaire. The peer-assessment questions were designed analogously.

Evaluation Process

The final note granted to each student consisted of three components:

I. the group percentage result from the tournament questions (the first six stages) – with a weight of 0.5,

II. the questionnaire-based assessment result (in percentage terms) for the "subject matter contribution" – with a weight of 0.3,

III. the questionnaire-based assessment result (in percentage terms) for the "communication skills" – with a weight of 0.2.

The percentage score for each team was obtained through dividing the number of points accumulated by the group by the maximal number of points possible to obtain. The points scored for answering the questions taken over from other groups were not included in the maximal number of possible points.

The questionnaire-based assessment results were included in the final score according to the authors' own approach presented below. For each person, the algorithm proceeded as follows:

1) Firstly, the median score was calculated of "subject matter contribution" and, separately, "communication skills" points in the self-assessment results (S).
2) Secondly, the median score was calculated of "subject matter contribution" and, separately, "communication skills" points attributed to the student by all other members of the group (the peer-assessment, P).
3) Then, the "subject matter contribution" and "communication skills" scores were obtained separately according to the rule:
    - If $|S - P| \leq 1$ (a consistent evaluation): take P as the final score,
    - else (an inconsistent evaluation): take P – 0.5 as the final score.

There are three premises behind the above algorithm. Firstly, we choose to represent the "average" (benchmark) score (in both S and P) by a median rather than a mean, for the previous – as opposed to the latter – is robust to extremities. Secondly, the assumed value of "1" as a tolerable discrepancy between S and P still ensuring a consistent evaluation is our arbitrary choice that appears justifiable in view of the 6-point scale employed in the questionnaires. Note that under such a scale, a tolerable deviation span of 2 points (i.e. *plus/minus* 1 point) constitutes ca. 33% of the entire 6-point range. Finally, in the case of an inconsistent evaluation (i.e. $|S - P| \geq 1$) we penalize the P result with an arbitrary value of 0.5. Note that regardless of the precise relation between the S and P assessments, the penalization is always downward, which is intended to reduce a risk of „collusion" among the students, and to stimulate honest and reasonable both self- and peer-assessments (the students had been familiarized with the algorithm prior to the tournament). In a broader perspective, such an approach should work both ways – preventing the participants from an unduly high as well as too low self-esteem. We address this issue to a greater extent in Subsection IV.C. The final score in the tournament, calculated according to the algorithm above, is henceforth denoted as "TNT".

In addition, the students' and teachers' opinions about the tournament as an assessment method were collected just after the implementation. All students were asked to express their reflections in an open-descriptive form, whereas the teachers took part in a semi-structured interview based on three items: (1) general perception of the activity, (2) opinion on feasibility of use in other subjects, and (3) the added value of a tournament comparing to traditional assessment methods. We discuss the results in Subsection III.F.

DATA COLLECTION AND ANALYSIS

In this section we provide details about the pre-tournament test and the post-tournament test, to assess students' progress (attributable to the tournament) with respect to their former achievements. To this end, a statistical analysis of relevant results is further performed.

Post-test

The post-test was prepared in a traditional, written form, and conducted one week after the tournament, with neither a prior review of the relevant content knowledge during regular classes nor a post-tournament discussion of the tournament problems and results (let us recall, however, that all tasks administered to students during the competition were then elucidated in the process either by the students providing the solution or by the teacher).

The test was unannounced, so the students have not been induced to make any additional efforts to prepare for it. In 60% the test comprised tasks utilized during the tournament, and in the remaining 40% it was based on problems totally new to the students, though similar to the ones given in the tournament. The post-test score is expressed in percentage terms, and, henceforth, denoted as "PT".

Former Tests

Each student, during the school year and before implementation of the tournament, participated in three tests: on thermodynamics, gravitation and electrostatics. All tests were taken individually. They contained mixed problems, including content knowledge and scientific reasoning tasks, multiple choice, open-response and calculus problems. To measure each student's achievements prior to the tournament, we used the average of his/her results on the three tests. The quantity obtained (expressed in percentage terms) is further referred to as the "former tests score" and denoted as "FT".

Basic Statistical Analysis

Figure 6 presents each student's three individual scores: on the former tests (FT), the tournament (TNT), and the post-test (PT), along with horizontal bars indicating the common (for each group) result gained from the tournament. All scores are provided in percentage terms. Note that the discrepancies between the group common result and the group members' individual scores stem from the outcomes obtained in the assessment questionnaires. Notice that, incidentally, the final marks assigned to each student within the fifth team were all lower than the common result of ca. 95%. This observation can be explained by the fact that nobody in the group was perceived as a leader, and all the team members were clearly aware of the fact that their final result was the effect of their cooperation (rather than attributable to the knowledge of a single leading person).

It can be easily noticed that the TNT marks were predominantly way above the FT results. What appears far more justifiable, however, is the comparison of the students' achievements and skills prior to and after the tournament, reflected in the FT and PT results, respectively. In that regard, however, we still observe a systematic (i.e. for almost all tournament participants) increase in score, with the result hinting at a positive impact of the tournament on the students' improvement.

In what follows, to explore the results in more detail, we conduct statistical analysis. As far as the sample size is concerned, since two students (no. 8 and 21; see **Figure 6**) were absent from the post-test, we exclude them from further considerations, and carry out the necessary calculations based on the sample of n = 28 students. Note that according to such a limited sample size the statistical inferences presented below should be approached with some reserve.

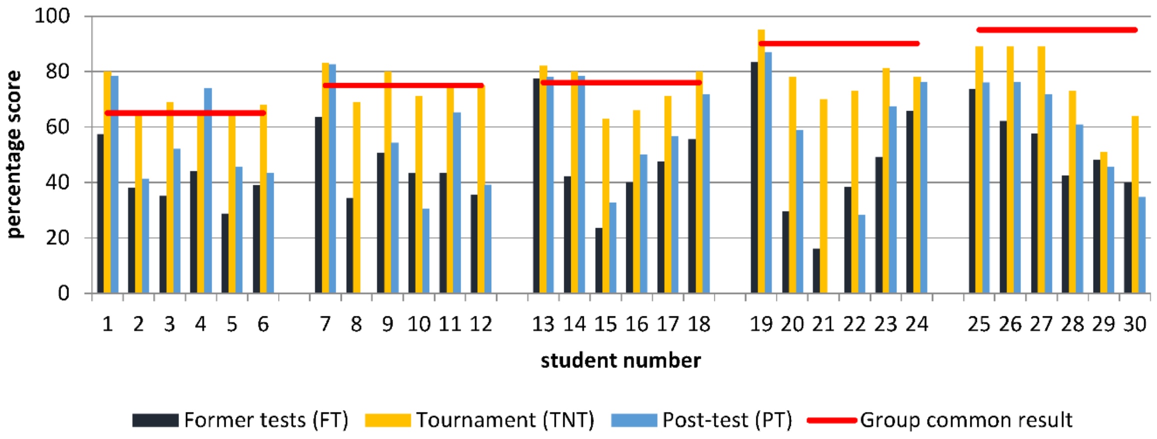


Figure 6. Student scores. For each student three vertical bars represent (starting with the leftmost one): the average score in former tests (FT), the final score obtained in the tournament (TNT), and the mark gained in the post-tournament test (PT). The horizontal lines represent the common result scored by each group in the tournament (based only on the first six stages, disregarding the qualitative component stemming from peer- and self-assessment)

Table 2. Basic statistics of the student results, including: the average score in the former tests (FT), the final score in the tournament (TNT), the result in the post-tournament test (PT), and differences between PT and FT (PT – FT). The last row contains statistics for the modified gain factor (MGF)

Variable	Characteristics						
	Mean	95%-confidence interval for mean	Median	Lower quartile	Upper quartile	Interquartile range	Standard deviation
FT [%]	48.39	(42.64; 54.14)	43.79	38.65	57.48	18.83	14.83
TNT [%]	74.96	(71.17; 78.76)	75.00	67.00	80.50	13.50	9.78
PT [%]	59.16	(52.29; 66.04)	59.78	44.57	76.09	31.52	17.73
PT – FT [pp]	10.77	(6.22; 15.32)	10.11	3.44	18.22	14.78	11.73
MGF	0.22	(0.13; 0.3)	0.23	0.07	0.37	0.30	0.22

In Table 2 and Figure 7 we present basic descriptive statistics and empirical distributions (histograms, with normality tested by the Lilliefors and the Shapiro-Wilk tests) for several variables, including the FT, TNT and PT scores, as well as the differences: TNT – FT and PT – FT (the latter measuring the "absolute" gain in student content knowledge). Moreover, we also examine a modified gain factor (MGF), which is our adaptation of the normalized gain (or the g-factor) (Hake, 1998), originally proposed in (Gery, 1972). The MGF measure is meant to relate the "absolute" gain in a student PT score to the points missed on FT, and is therefore calculated according to the formula:

$$MGF = \frac{PT-FT}{100-FT}.$$

From Table 2 it can be inferred that the students scored, on average, ca. 48.4% upon the former tests, with the standard deviation hovering around 14.8 percentage points (henceforth, pp). Half of the students recorded the FT result below ca. 43.8%, whereas the other half – above that number. (The means and medians differ on account of positive skewness of the empirical distribution; see Figure 7(a)). On the other hand, results obtained during the tournament are distinctive on two counts. Firstly, the average TNT score is much higher as compared to FT. Arguably, the difference can be attributed to the team work and cooperation among the students. Note, however, that ultimately these two scores should not be compared *per se*, since calculation of the TNT results include a strong "qualitative" component. Secondly, the TNT scores are more concentrated (as compared with FT) around the mean, with a drop in standard deviation of ca. 5 pp. Moreover, the TNT distribution is far more symmetrical than its FT counterpart (see Figure 7(b)), thereby closing the gap between the mean and median (both equal around 75%; see Table 2). In general, the TNT scores are more regularly, symmetrically distributed and strongly shifted rightwards as compared to the FT results. (Note, however, that for all but one the analyzed variables, with PT being the exception, despite more or less conspicuous irregularities such as skewness and multimodality, the null hypothesis of normality is not rejected, which, admittedly, is largely due to the low sample size. Still, as implied by the corresponding p-values, the TNT distribution is far closer to normal than actually any of the others; see Figure 7).

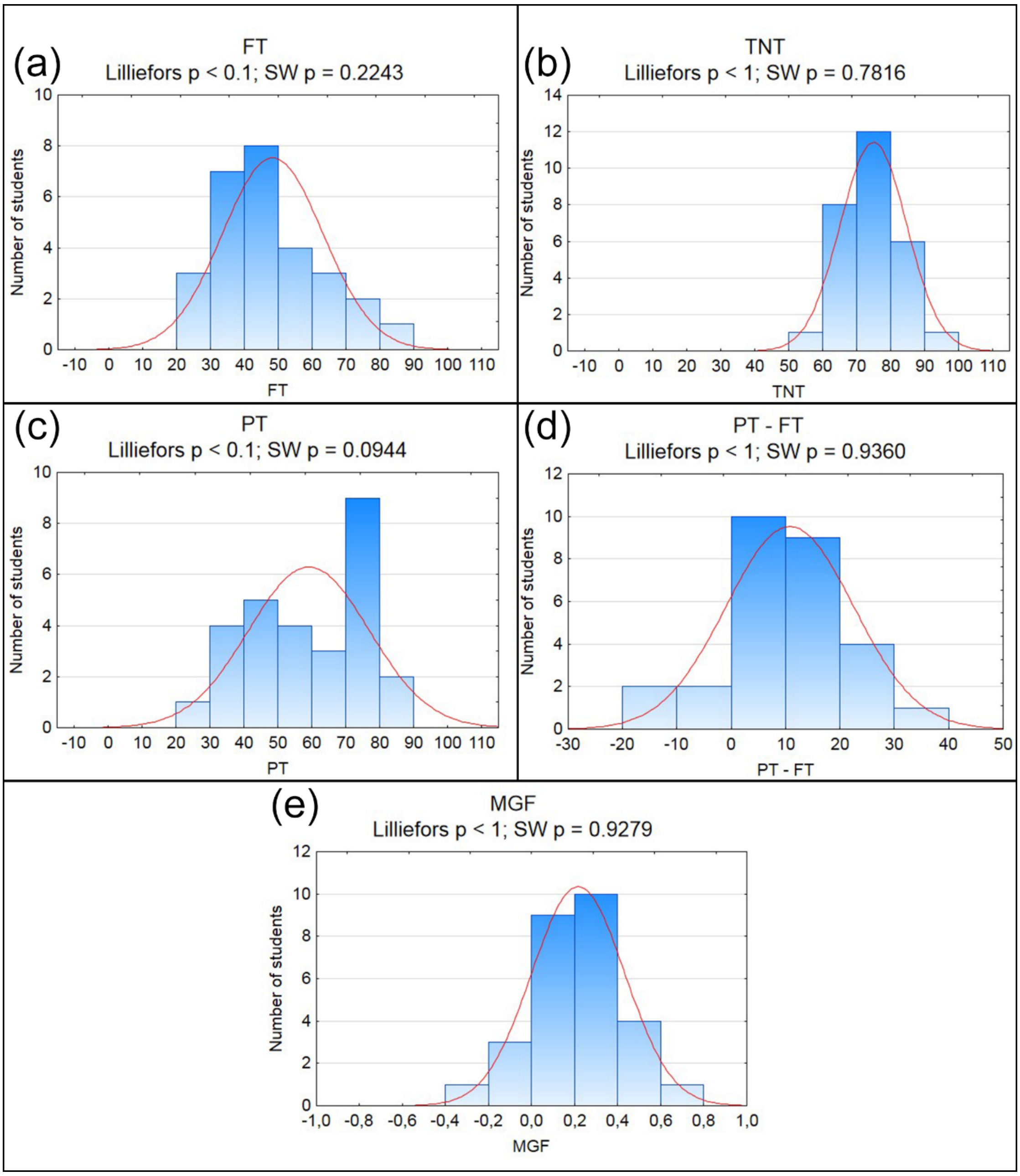


Figure 7. Histograms of the students' results: (a) the average score in the former tests (FT), (b) the final score in the tournament (TNT), (c) the result in the post-tournament test (PT), (d) differences between PT and FT (PT – FT). Panel (e) displays the histogram for the modified gain factor (MGF). In each case the normal density is fitted (solid line), accompanied by p-values for testing normality through the Lilliefors and the Shapiro-Wilk tests (denoted as "Lilliefors p" and "SW p", respectively)

Moving on to the PT results, it appears, interestingly, that these are somehow less regular than FT, on two counts. Firstly, the PT distribution has a higher dispersion, as implied by both standard deviation and, in particular, interquartile range (see **Table 2**). Secondly, as long as the FT distribution features only a single mode (somewhere between 40 and 50%), the PT histogram exhibits a pronounced bimodality. Apparently, the two PT modes correspond with the ones present in the FT and the TNT distributions, with the global PT mode (between 70 and 80%) coinciding with the TNT one, and the second, a local one (between 40 and 50%) – with the FT mode. In statistical terms, one could argue that the PT distribution is a mixture of the FT and TNT distributions. Practically speaking, it could be inferred that the PT scores are formed as a confluence of student prior physical expertise (measured by FT) and the knowledge and skills acquired during the tournament.

Finally, we proceed with the analysis of the results scored in the post-test in relation to student content knowledge and skills prior to the tournament (FT results). The average difference between the PT and FT scores

Table 3. Testing positive means for: difference between the results gained in the post-test and the former tests (PT – FT), and the modified gain factor (MGF). In the second column values of the Student-t statistics are displayed for testing a positive mean. The last column presents corresponding p-values

Characteristics	Test statistics	p-value
PT – FT	4.86	2.20×10^{-5}
MGF	5.30	6.80×10^{-6}

Table 4. Basic statistics of the student results in the control and the experimental groups, including: the average score in the former tests (FT), the final score in the tournament (TNT), the result in the post-tournament test (PT) and differences between PT and FT (PT – FT). The last row contains statistics for the modified gain factor (MGF). Results for the control group are indicated with letter "c" in the superscript

Variable	Characteristics						
	Mean	95%-confidence interval for mean	Median	Lower quartile	Upper quartile	Interquartile range	Standard deviation
FT^c [%]	48.50	(41.43; 55.56)	47.12	38.21	58.63	20.42	15.94
FT [%]	48.39	(42.64; 54.14)	43.79	38.65	57.48	18.83	14.83
PT^c [%]	47.68	(40.79; 54.57)	48.91	35.87	58.7	22.83	15.54
PT [%]	59.16	(52.29; 66.04)	59.78	44.57	76.09	31.52	17.73
$PT^c – FT^c$ [pp]	-0.82	(-5.15; 3.51)	-2.88	-6.36	3.58	9.94	9.77
PT – FT [pp]	10.77	(6.22; 15.32)	10.11	3.44	18.22	14.78	11.73
MGF^c	-0.05	(-0.16; 0.06)	-0.05	-0.17	0.11	0.28	0.25
MGF	0.22	(0.13; 0.3)	0.23	0.07	0.37	0.30	0.22

totals ca. 10.8 pp (see **Table 2**), and it is statistically significant, regardless of the α level (see **Table 3**). (Note, however, that four out of 28 students scored lower in PT than in FT, so negative increments were also reported). Improvement of the student performance is also indicated by the results obtained for the modified gain factor. A test of positive MGF mean indicated that it was significantly positive at any typical α level (see **Table 3**). Note, however, that the MGF histogram exhibits two pronounced and equivalent modes, which may question the use of the mean as a measure of central part of the distribution. Nevertheless, both modes are positive. Furthermore, almost 86% of the probability mass in the histogram is localized to the right of zero, which implies that a learner positive gain was reported for a predominant number of students (i.e. 24 out of 28; see **Figure 7(e)**).

Control Group

In order to validate a statistical approach to examining the influence of the tournament on students' achievements, we formed a control group of 22 students also attending a K-12 class. The control group students took the same former tests and the same post-test as the experimental group students (i.e. the ones analyzed in the previous subsection), but did not participate in the tournament. The former tests results, the post-test scores and the modified gain factor for the control group, which are analyzed below, are calculated in the same manner as in the case of the experimental group, and denoted analogously, i.e. FT^c, PT^c and MGF^c, respectively.

Table 4 summarizes basic statistics of the results gained by the students of each group (i.e. the control and the experimental one), whereas **Figure 8** depicts the histograms of the control group's outcomes along with the normality tests. With regard to the latter, it appears that only MGF^c features some slight departures from the normal distribution, which is attributable to the heavy left tail of the histogram. Statistics presented in **Table 4** reveal a very close similarity of the former test results in both groups not only in terms of means, but other characteristics as well, thereby indicating the validity of the control group at hand for our "experiment". A battery of statistical tests for the equality of: means, variances and the very distributions of FT and FT^c, corroborate this presumption (see **Table 5**).

A visual inspection of the mean values displayed in **Table 4** may indicate a non-negligible positive effect of participating in the tournament on student achievements. As long as there is no significant discrepancy (at α = 0.1) between the control group's former and post-tests scores (p-value ≈ 0.7; see **Table 6**), it turns out that the tournament participants scored significantly higher on the post-test than their control group counterparts, both in terms of a simple difference between PT and PT^c (p-value ≈ 0.01; see **Table 6**), and the modified gain factor (p-value ≈ 0.0001; see **Table 6**).

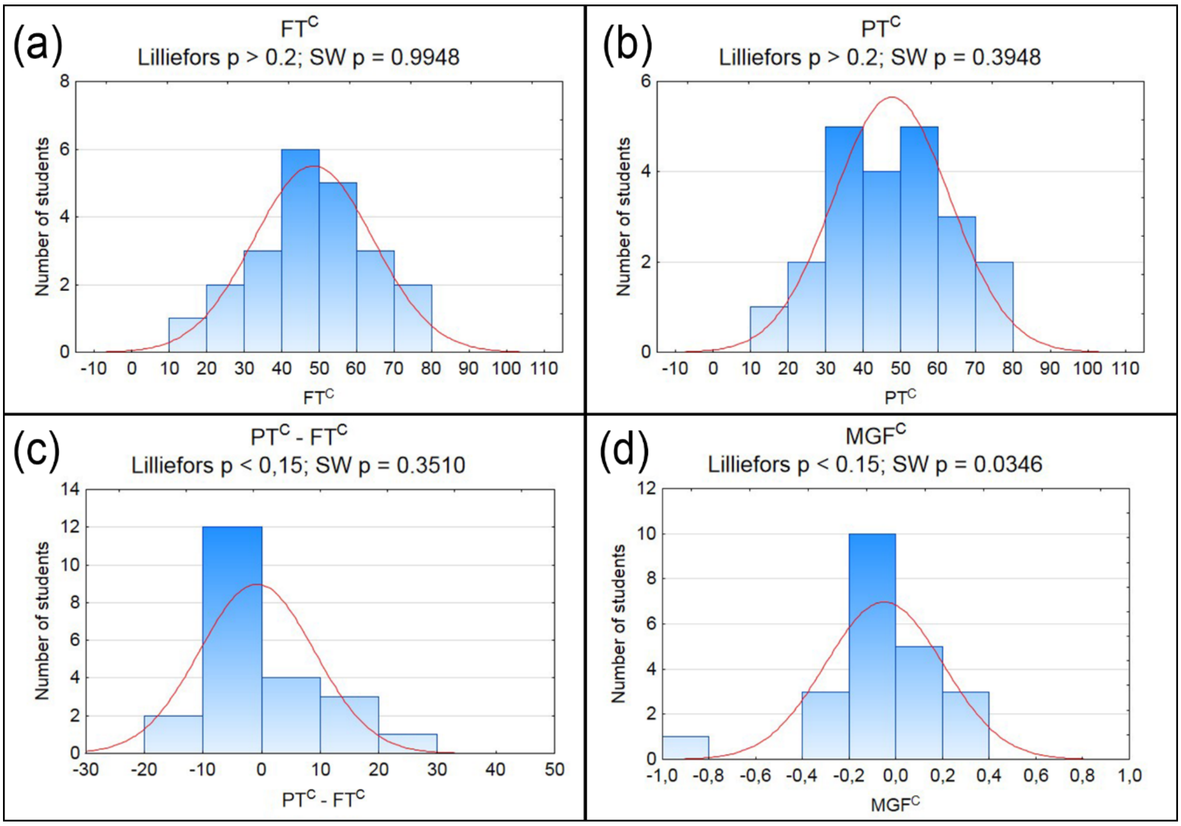


Figure 8. Histograms of the students' results in the control group: (a) the average score in the former tests (FTc), (b) the result in the post test (PT^c), (c) differences between and PT^c and FT^c (PT^c – FT^c). Panel (d) displays the histogram for the modified gain factor (MGFc). In each case the normal density is fitted (solid line), accompanied by p-values for testing normality through the Lilliefors and the Shapiro-Wilk tests (denoted as "Lilliefors p" and "SW p", respectively)

Table 5. Testing the control group for its compatibility with the experimental group, by means of examining the equality of means, variances and distributions of the former tests results obtained in each group (FT^c and FT, respectively). The null (the alternative, correspondingly) hypothesis in each testing procedure states the equality (the inequality) of a given characteristics of the former test results in both groups. For the Mann-Whitney test two statistics are considered: U, the original one, and Z, following approximately the standard normal distribution

Equality of FT and FT^c's …	Testing procedure	Test statistics	p-value
Means	*t*-test	-0.0238	0.9811
Variances	*F*-test	1.1558	0.7141
	Levene	0.1363	0.7136
	Brown-Forsythe	0.1985	0.6579
Distributions	Kolmogorov-Smirnov	0.1266	> 0.10
	Mann-Whitney	300 (*U*) 0.1466 (*Z*)	0.8835 (*U*) 0.8845 (*Z*)

Table 6. Testing the effect of the tournament in terms of: the mean difference between the former and the post-test results in the control group (PT^c – FT^c); inequality between the mean post-test results in the experimental and the control group (PT and PT^c, respectively); inequality between the mean modified gain factors in the experimental and the control group (MGF and MGF^c, respectively)

Characteristics	Test statistics	The alternative hypothesis	p-value
PT^c – FT^c	-0.3930	Mean (PT^c – FT^c) different from zero	0.6982
PT and PT^c	2.4370	Mean PT different from mean PT^c	0.0186
		Mean PT higher than mean PT^c	0.0093
MGF and MGF^c	3.9543	Mean MGF different from mean MGF^c	0.0003
		Mean MGF higher than mean MGF^c	0.0001

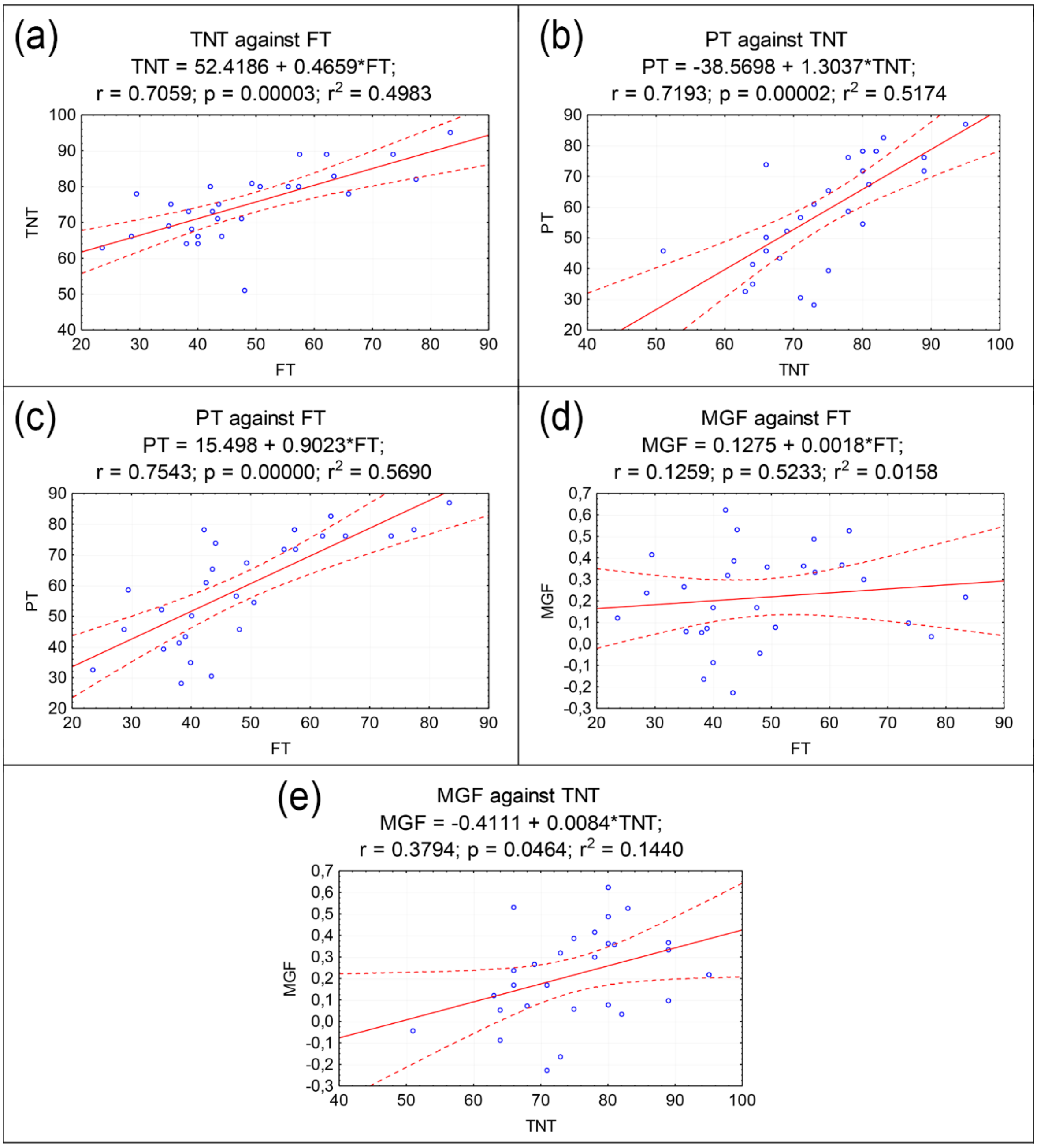


Figure 9. Scatter plots for selected pairs of the students' results. Apart from data points, in each plot a linear regression is fitted and the 95%-confidence band is marked. Below each regression equation we provide: Pearson's linear correlation coefficient (r), p-value for testing a non-zero correlation coefficient, and determination coefficient (r^2)

Correlation and Regression Analysis

Below, the analysis of correlations between selected pairs of the considered variables is performed. Figure 9 displays relevant scatter plots, along with fitted linear regressions, 95%-confidence bands, linear correlation coefficients (r), p-values for testing their significance, and, at last, the square of correlation coefficients (r^2), which coincide with determination coefficients in the fitted regressions. Based on Figure 9, the following general conclusions can be formulated:

1) A positive and statistically significant correlation between FT and TNT implies that students who performed better prior to the tournament, also scored higher in the tournament (see Figure 9(a)). It is worth underlining that the value of correlation coefficient (r = 0.7059) is negatively affected by a single outlying TNT score equal 51 (obtained by student no. 29), exclusion of which raises the coefficient value to r = 0.8021.

2) Similarly, the TNT and the PT results are positively and statistically significantly interrelated (r = 0.7193), indicating that better (likewise, worse) performance in the post-test coincides with a higher (respectively, lower) score in the tournament (see **Figure 9(b)**). Exclusion of the outlier represented by student no. 29 slightly raises the correlation coefficient to r = 0.7466.
3) As expected on the basis of the two above observations, there also occurs a positive and significant relation between the PT and FT scores, indicating that high (likewise, low) notes in the post-test were mostly obtained by those who already performed high (low, respectively) in the former tests (see Figure **9(c)**).
4) Some slight (r = 0.1259), yet statistically insignificant correlation is observed between MGF and FT, hinting at no dependence of a student gain upon his/her previous performance (see **Figure 9(d)**).
5) On the other hand, it appears that student improvement (as measured by MGF) is significantly and positively influenced by the tournament performance, though the correlation coefficient is only ca. 0.38 (see **Figure 9(e)**). The result suggests that the learner gain is generally higher in the case of those who scored higher on TNT.

The inferences formulated in items no. 1-3 point to an intuitive relation according to which the better a student has fared so far, the higher his/her performance in the tournament, and, eventually, in the post-test. Further, result no. 4 implies generally that the student gain, arguably attributable to the tournament, hardly depends on his/her former achievements. In broad terms, it would follow that the tournament provided equal opportunities of improvement to all students. Nevertheless, conclusion no. 5 would still indicate that those who performed better in the tournament (as the effect of their active involvement in cooperative work), actually improved slightly (yet significantly) more than the others.

The results presented above provided us an incentive to build two simple bivariate linear regression models in order to jointly evaluate the impact of the former tests and the tournament results on the post-test score and the modified gain factor. The two models take the following form:

$$Y = \beta_0 + \beta_1 FT + \beta_2 TNT + \varepsilon,$$

with Y representing the dependent variable (i.e. either PT or MGF), and ε denoting normally distributed random errors with zero mean and satisfying typical assumptions of a standard linear regression model. In **Table 7** the following estimation results are presented: determination coefficient (R^2), point estimates, standard errors, p-values against the alternative of a non-zero coefficient (i.e. H_1: *coefficient* ≠ 0), p-values against the alternative of either a positive or negative coefficient (i.e. H_1: *coefficient* > 0, or H_1: *coefficient* < 0), depending on the sign of the point estimate. (Though not reported in the paper, the Lillierfors and Shapiro-Wilk tests do not reject the normality of residuals in any of the regression models considered below, therefore validating testing the regression coefficients by means of a standard Student's t-test).

As regards regressing PT against FT and TNT, it appears that both regressors positively influence the PT score. More specifically, if a student scored higher in FT (likewise, TNT) by 1 pp, then he/she would score also higher in PT by ca. 0.59 pp (0.68, respectively). The results are (positively) significant at α equal 0.01 and 0.05, correspondingly. With respect to the determination coefficient, we note that about 64% of the post-test results is explained by the former tests and the tournament performance.

In view of these results, it is worth noting that also in the control group there is a positive correlation between the post-test and the former tests results. The correlation coefficient between PT^c and FT^c equals 0.81, while its counterpart in the experimental group: 0.75. A slightly higher value in the control group indicates that the PT^c and FT^c scores are more similar to each other than the corresponding results obtained by those students who participated in the tournament, which is also evidenced by the basic statistics presented in **Table 4**. Such an observation may be simply attributed to the lack of intervention in the control group (so that PT^c and FT^c are largely similar), and, at the same time, the (positive) impact of the tournament modifying the students' former achievements so that their PT scores differ more from FT than in the case of the control group. Nevertheless, one should bear in mind that comparing the two correlation coefficients at hand should be made with caution, because in the case of the experimental group the TNT score is yet another variable that is positively correlated with both: FT and PT. Hence, measuring correlation between PT and FT by means of a simple correlation coefficient, which – by construction – fails to take TNT explicitly into account, appears inadequate. Therefore, in order to disentangle the effect of the students' former achievements and the tournament upon their post-test results, we resorted to the multiple regression analysis discussed in the previous paragraphs.

With respect to the regression for MGF, we note that as long as the student gain depends positively on the tournament performance (at α = 0.05), it is not determined by FT (see **Table 7**). As already mentioned above, it would follow that the student improvement, arguably attributable to the tournament, does not depend on their former achievements, and, in broad terms, that the tournament provided equal opportunities of improvement to all students. Note, however, that only about 18% of the modified gain factor can be explained by the former achievements and the tournament performance.

Table 7. Regression results for PT and MGF. Asterisks indicate statistical significance (a non-zero coefficient): ** for $\alpha = 0.01$, * for $\alpha = 0.05$. Note that ε equals 0 in the estimated model

	Dependent variable: PT			**Dependent variable: MGF**		
Regressor	**Constant**	**FT**	**TNT**	**Constant**	**FT**	**TNT**
Parameter	**β_0**	**β_1**	**β_2**	**β_0**	**β_1**	**β_2**
Point estimate	-19.8827	0.5878**	0.6750*	-0.5421	-0.0041	0.0128*
Standard error	17.6894	0.2031	0.3077	0.3234	0.0037	0.0056
p-value against a non-zero coef.	0.2717	0.0078	0.0378	0.1062	0.2778	0.0320
p-value against a positive/negative coef.	0.1358	0.0039	0.0189	0.0531	0.1389	0.0160
Determination coefficient (R^2)		0.6386			0.1841	

Qualitative Analysis of Students' and Teachers' Opinions

Students' opinions

After the post-test, and before getting informed about their final marks, the students were asked to express anonymously their opinions about a tournament as a tool of assessment. The participants were encouraged to formulate their comments in an open, descriptive form, with no predefined questionnaire to follow. Such an approach was meant to induce student openness and spontaneity, with no intent on our part to perform any further (quantitative) analysis of the answers. Some examples of the comments are cited below:

Student A:

I think that this form of a test is good, because we can share our knowledge with others and vice versa, helping each other. We can memorize more and learn new things.

Student D:

This is a better form of consolidation and verification of our knowledge and skills.

Student E:

This is a good idea, because it was performed in the form of a game. A student can show what he or she knows without being stressed.

Student K:

Fabulous! We can integrate, everybody who had any idea but wasn't sure about it had an opportunity to consult/discuss it with other members of the group.

Student O:

I really liked explanation of each answer given afterwards. This way it was possible to understand more.

Student W:

Everybody wanted to receive a good note and knew that there is "collective responsibility" and tried to do his/her best.

Student Z:

I suggest a different way of intercepting questions. Frankly, the bell was getting on my nerves and caused me a headache.

It is worth noting that, except for the last one (regarding the bell ringing), all the students' opinions were positive and enthusiastic.

Teachers' opinions

Just after the tournament two assistant teachers and the teacher conducting the lesson were asked to take part in a semi-structured interview about their perception of the intervention. The common agreement was that the method positively influenced the engagement of the students and raised their interest in physics. All of them also admitted that the method seems to be largely universal and feasible to extend to other topics and school subjects. They also pointed out that, contrary to the traditional assessment methods, oriented mostly on the content knowledge itself, the tournament evaluated also practical and soft skills. One of the teachers said: "*... this is a good opportunity to inure students to the way they might be assessed in their future study and work where not only knowledge and individualism counts, but also cooperation skills.*" The other teacher indicated "*... the method is attractive to young students, sharpens their focus and develops positive attitudes towards science, so much emphasized in the curriculum.*"

DISCUSSION

Social Benefits

Based on the ones delivered above we proceed with a short discussion about students' social benefits arising from participation in the tournament. Note, however, that in our study we did not measure any of the effects mentioned below, including diminishing students' anxiety, improving their social skills and the ability of critical thinking. Although a relevant quantitative analysis of these psychological phenomena appears worthwhile, it is beyond the scope of the current research. Therefore, in this subsection we draw our conclusions solely on the students' and teachers' feedback, relating them to the findings commonly presented in the literature on collaborative testing and gamification.

The tournament was organized in the form of a team game, but with elements of rivalry. In this way it can be perceived as a form of activity in which group work skills, desirable in some academic areas and also by employers, are naturally activated, playing crucial role in accomplishing tasks (Dallmer, 2004; Dicheva et al., 2015; Kapitanoff, 2009; Lusk & Conklin, 2003; Sandahl, 2010; Seaborn & Fels, 2015; Shindler, 2003). Simultaneously, the tournament induced far less test anxiety (as compared with traditional, individually taken written test) by giving students a sense of being supported by the other team members tasks (Banfield & Wilkerson, 2014; Kapitanoff, 2009; Lusk & Conklin, 2003; Sandahl, 2010; Zimbardo et al., 2003). Working together may improve communication skills as well. Students learn to listen to each other, share information, and respond to ideas proposed in discussions, which stimulate knowledge assimilation (Hanus & Fox, 2015; Jolliffe, 2007). What is worth noting is that vocabulary and concepts used in group and class discussions may provide retrieval cues that help students recall relevant information. Moreover, the requirement of providing not only the answer to a question, but also the explanation for it, necessitated that the students should be able to understand and present their lines of reasoning and reconsider them, if needed. Therefore, a tournament may also yield an improvement in students' ability of critical thinking as well as facilitate their intrinsic motivation tasks (Banfield & Wilkerson, 2014; Kapitanoff, 2009; Lusk & Conklin, 2003; Shindler, 2003). Finally, an active involvement in the self- and peer-assessment process may improve student confidence and adequate self-esteem (Hendrix, 1996), thereby enhancing retention of knowledge (Sawtelle et al., 2012). Taking all the above into consideration, cooperative testing of knowledge may become a significant part of the learning process.

Academic Benefits

The main purpose of this research was to examine the impact of taking an exam in the form of a tournament on student achievements. Firstly, a statistically significant increase is observed in students' achievements in the tournament as compared to their former tests results (the average difference amounted to ca. 26 pp, in favor of the tournament scores, being positively significant at any typical α level). Secondly, we also find evidence for improvement of student content knowledge and problem solving skills, as indicated by the results of the post-test taken by the students a week after the tournament (the average difference between marks in the post-test and former tests scored ca. 11 pp; the mean of modified gain factor totaled 0.22; both results are positively significant at any typical α level). Our findings remain in accordance with much research on positive impact of collaborative testing. Studies presented in (Bloom, 2009; Haberyan & Barnett, 2010; Kapitanoff, 2009; Lusk & Conklin, 2003), focused on the effects of taking exams in a collaborative way for numerous groups with various numbers of students and of different subject/specialization, indicated higher students' achievements as compared with traditional ways of individual testing of knowledge. Moreover, in (Bloom, 2009) it was found that collaborative exam scores were also higher than the ones earned in individually taken exams during which students were allowed to use course textbooks and their notes. Further, some researchers show that students' performance also improved in a longer perspective, as indicated by post-tests taken some time after the collaborative exam (Cortright et al., 2003; Jensen et al., 2002; Simpkin, 2005). Notice that in our research we established a positive and statistically significant impact of participation in the tournament on students' achievements in the post-test.

Finally, in the context of the tournament organization, let us emphasize that the event was not preceded by any traditional, individually taken test on the subject matter (i.e. electricity), though, conceivably, it would be worth contrasting the post-test results with the ones obtained in a typical pre-test on the same content. In our approach we followed conclusions formulated in (Dahlström, 2012), who suggested that the learning gain due to taking a collaborative final exam might be higher if the students had no previous individual encounter with relevant tasks. In the cited paper it was found that in the post-test the students scored higher on new problems (i.e. the ones that had not been used in the pre-test) than on the questions they had already been given previously. A possible logic behind this observation is that the lines of reasoning followed by a student during an individually taken exam tend to persist afterwards, therefore hindering acquiring new ways of thinking and solving the problem, even after participating in a collaborative activity. It would follow then that, as claimed in (Dahlström, 2012), "it might be preferable to collaborate without first deciding on questions individually." Taking this as well as our findings into account, we infer that a class tournament is a well-justifiable and effective learning activity, in which the three approaches to assessment (i.e. *of*, *for* and *as* learning (Earl, 2004)) merge together.

Comments on TNT Grading

In our study, a tournament is proposed as a form of summative assessment with formative elements, since it served us to evaluate students' content knowledge and practical skills in a particular physics area (though, obviously, a contest-based evaluation procedures are readily adaptable to other areas of education). As implied in the previous section, the tournament assessment yielded significantly higher final scores in comparison with the results obtained in former, classical and individually written tests. On the one hand, to some, such an outcome may cast doubt on a tournament as a valid means of student evaluation, for no longer only the content knowledge is subjected to scrutiny, but also other aspects of student performance, particularly group work skills. However, as mentioned in the previous section, in view of a voluminous literature on collaborative work and group work assessment strategies, the apparent discrepancy between the TNT and FT results is perfectly justifiable. Nevertheless, let us also note, however, that addressing the issue of what a student grade should reflect actually requires settling on what exactly should be subjected to assessment. This, in turn, is often a matter of national educational regulations and curricula, differing across countries. It should be highlighted that a tournament as a tool of student evaluation leaves the teacher much space for modifications in terms of the formulation and the difficulty level of tasks, the way of calculating composite and final scores, etc.

Organizational Considerations

Another issue that may arise among teachers searching for a practically valid and feasible alternative to classical forms of student assessment is the question of the organizational effort behind it. As regards a tournament itself, we admit that it may (though need not) be a more demanding and time-consuming endeavor than preparing and conducting an individually written test. Even setting the issue of the time cost aside, the idea of a tournament may still be approached by some with reluctance due to the need of an active and ceaseless involvement of a teacher during the event itself. Nevertheless, there are manifest benefits of this additional effort, among which the most obvious one is the online feedback between students and teacher. This allows the teacher to elicit constantly, during the process, and to monitor not only the students' content knowledge, but also their ways of understanding (Stang & Roll, 2014). Once the teacher spots some deficiencies in either the content or the reasoning, he/she is enabled to straighten them out online. Obviously, a typical written test does not allow for such a possibility (Franklin & Hermsen, 2014). Therefore, during a tournament, by listening attentively to students' responses, understanding students' lines of reasoning, and addressing them relevantly, the teacher has a unique opportunity to assess the participants in a most formative manner.

Self- vs. Peer-assessment

Other doubts may arise with respect to the self- and peer-assessment evaluation procedure implemented in our study. There are many papers in the literature on assessing student engagement in a broadly defined group work, with many different strategies and ideas (e.g. Fernandezbreis et al., 2009; Moccozet et al., 2013). It may appear to some that the algorithm implemented in our research, primarily designed by us to encourage truthfulness in the contestants, tends to affect only those students who appraise themselves too high as compared to the evaluation by his/her teammates. However, it should be stressed that the formula hinges upon the absolute value of the difference between the self- and peer-evaluation scores, thereby equally penalizing unduly over- as well as underestimated self-assessments. Hence, we regard the scheme proposed in this paper – obviously remaining open to further enhancements and suitable adaptations – as developing a student's sense of need to provide honest evaluations, both with respect to themselves and the other members of his/her team. To this aim we deem it of utmost importance for the teacher to provide the participants – before the tournament – with an explanation of how possible discrepancies between the self- and peer-assessments are going to affect their final scores, making then clear indications that, in view of the implemented algorithm, honesty is the best policy – also for those students who tend to underestimate their achievements, skills or abilities. Then, while contrasting the self- and peer-assessments results after the tournament, the teacher is able to pinpoint those participants whose scores are overly divergent. In such cases the teacher should be prompted to take proper measures, such as discussing individually the noticed discrepancy with each of the selected students in order to trace its origins. Depending on the teacher's judgment, for some students it may emerge advisable to further seek a professional psychological advice so as to eventually develop in them a proper overall subjective emotional evaluation of his/her own worth. In addition, we also believe that performing tournaments cyclically would enable the teacher to track each student's dynamics in

Table 8. Analysis of the differences between the self- and peer-assessment scores by gender. "S" and "P" stand for the self- and peer-assessment scores, respectively. The table reports on the number of students for whom a given inequality between S and P occurred. Note that S – P > 0 (S – P < 0, respectively) indicates that a student overestimated (underestimated) his/her contribution in his/her teammates' opinion. The cases of S – P > 1 and S – P < –1 are regarded as an inconsistent evaluation, resulting in a penalization of the final score (see Sec. II B, the third point of the algorithm of the questionnaire-based part of assessment)

	Subject matter contribution		**Communication skills**	
No. of cases	**Boys**	**Girls**	**Boys**	**Girls**
S – P < 0	16	9	10	7
S – P > 0	2	2	9	4
S – P = 0	1	0	0	0
S – P < –1	3	2	1	2
S – P > 1	0	1	1	1

terms of their self-esteem (Lindsey & Nagel, 2015). Incidentally, let us note that all the tournament participants, although not used to self- and peer-assessment, embraced unequivocally the practice of mutual evaluation.

It may also be interesting to analyze the number of participants for whom the discrepancy between the self- and the peer-assessments scores (denoted as "S" and "P", respectively) was too large, resulting in a penalty for an inconsistent evaluation (which is the case when $|S - P| > 1$; see Subsection II.B). In **Table 8** we report relevant quantities by gender (there were 11 girls and 19 boys participating in the tournament; note that the total number of participants is 30, including also the two students who did not take the post-test). Overall, the students of both sexes tend to evaluate consistently both their subject matter contribution and communication skills. However, the gaps between the numbers of males and females who undervalue their contribution (i.e. S – P < 0) and the ones that overrate it (S – P > 0) are far more evident for the subject matter involvement than for the communication skills. Moreover, an overwhelming majority of students of both sexes is inclined to underestimate (rather than overestimate) their subject matter contribution. Exceptions of students overly underrating their contribution (i.e. S – P < –1) include three boys and two girls with respect to the subject matter, and only one boy and two girls in terms of the communication skills. Interestingly, only in one of these cases the tournament participant (a boy) underestimated himself on both counts. All the other students under consideration scored S – P < –1 only in one of the analyzed aspect. On the other hand, the cases of overvaluation of one's involvement were relatively rarer. In terms of the subject matter no boys and only one girl overrated their contribution, whereas with regard to communication two students (each of a different sex) evaluated their performance too enthusiastically.

Gender Differences

The final issue we would like to raise here, and the one that quite naturally spins off from the previous subsection, is an analysis of the major results (i.e. FT, TNT and PT) by gender, so as to identify and characterize possible sex-specific effects and dependencies, collectively termed as a gender gap (Kost et al., 2009; Madsen et al., 2013; Pollock et al., 2007). Basic descriptive statistics, presented in **Table 9** for the male and female students separately, imply no relevant gender discrepancies in terms of the mean and median scores, with the boys performing slightly better than the girls. However, we refrain from testing statistical significance of these differences, for under such low sample sizes no valid conclusions could be obtained. Further, we also notice that despite the similarities between the groups' means, the boys' scores are more diffused for typically-written tests (FT and PT). Although the TNT results appear more evenly dispersed in both groups (as indicated by the standard deviations and the ranges between maxima and minima), the "innermost" 50% of the scores (i.e. those between the lower and the upper quartile) obtained by the females are more scattered than for the male students.

Table 9. Basic statistics of the students' results by gender, including: the average score in the former tests (FT), the final score in the tournament (TNT), the result in the post-tournament test (PT), differences between TNT and FT (TNT – FT), as well as PT and FT (PT – FT). The last row contains statistics for the modified gain factor (MGF)

	Sex	Mean	Median	Minimum	Maximum	Lower quartile	Upper quartile	Interquartile range	Standard deviation
FT	Boys	49.08	43.79	23.57	83.33	38.33	57.58	19.25	16.99
	Girls	47.15	45.12	35.08	65.87	38.97	50.63	11.67	10.57
TNT	Boys	76.72	75.00	63.00	95.00	71.00	82.00	11.00	9.30
	Girls	71.80	73.50	51.00	83.00	64.00	80.00	16.00	10.33
PT	Boys	60.01	63.04	28.26	87.00	45.65	76.00	30.35	18.45
	Girls	57.61	53.26	34.78	82.61	43.48	76.09	32.61	17.19
TNT - FT	Boys	27.64	27.28	4.51	48.51	22.62	34.67	12.04	10.65
	Girls	24.65	27.49	2.97	37.78	19.55	31.71	12.17	10.56
PT - FT	Boys	10.93	11.99	-12.90	29.83	3.67	18.33	14.66	11.75
	Girls	10.46	7.36	-5.14	36.04	3.25	18.11	14.86	12.35
MGF	Boys	0.22	0.23	-0.23	0.53	0.09	0.37	0.28	0.21
	Girls	0.21	0.17	-0.09	0.62	0.05	0.36	0.30	0.24

CONCLUDING REMARKS

In this paper we present both, quantitative and qualitative results of a tournament as a method of assessing student performance in physics classes on electricity. Based on students' results and students' and teachers' opinions we can come up with the following conclusions:

I. As compared with the control group results, the tournament proved to significantly enhance the experimental group students performance.

II. For most learners in the experimental group their results got in an individually written post-test (taken a week after the intervention) were higher than their average performance beforehand.

III. Scores obtained during the tournament were higher than in traditionally performed tests.

IV. The alternative method of testing analyzed in our paper appears to provide equal opportunities of improvement both for low- and high-performers through the tournament approach.

V. Both students and teachers appreciated the method very much because it enabled students to help each other in solving problems in a more cooperative, less stressful way and develop soft skills.

In general, there are several advantages of such a form of examination that outweigh organizational difficulties mentioned earlier in this study. These include: supporting weaker students by collaboration with others, setting a framework of cooperative-learning among students, development of group-work skills, stress-free testing, and, in addition to these, integration of the class.

Finally, let us note that our approach can be easily transferred and adapted to testing achievements in fields other that physics, particularly the natural sciences. Nevertheless, the subject which we had chosen for testing out method was physics, which is largely due to its obvious feature of combining algebraic calculations with both the description and explanation of real-world phenomena. Implementations of the tournament as an assessment method in other areas could be the subject of further studies.

ACKNOWLEDGEMENTS

Łukasz Kwiatkowski would like to acknowledge partial financial support by the Foundation for Polish Science (within the START 2014 program), and partial support from research funds granted to Faculty of Management at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

Daniel Dziob would like to acknowledge the Polish Ministry of Science and Higher Education financial support under program 7150/E-338/M/2017.

REFERENCES

Apostol, S., Zaharescu L., & Aleze, I. (2013). Gamification of Learning and Educational Games. *eLearning & Software for Education, 2*, 67.

Banfield, J., & Wilkerson, B. (2014). Increasing Student Intrinsic Motivation and Self-Efficacy through Gamification Pedagogy. *Contemporary Issues in Education Research (CIER) CIER, 7*(4), 291. doi:10.19030/cier.v7i4.8843

Biggs, J. (1998). Assessment and Classroom Learning: A role for summative assessment? *Assessment in Education: Principles, Policy & Practice, 5*(1), 103-110. doi:10.1080/0969595980050106

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. doi:10.1080/0969595980050102

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215-232. doi:10.1080/09695941003696016

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the Black Box: Assessment for Learning in the Classroom. *Phi Delta Kappan, 86*(1), 8-21. doi:10.1177/003172170408600105

Black, P., Harrison, P., Hodgen, C., Marshall, J., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*(4), 451-469. doi:10.1080/0969594X.2011.557020

Bloom, B. S., Hastings, J. T., Madaus, G. F., & Baldwin, T. S. (1971). *Handbook on the formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bloom, D. (2009). Collaborative Test Taking. *College Teaching, 57*(4), 216-220.

Cortright, R. N., Collins, H. L., Rodenbaugh, D. W., & Dicarlo, S. E. (2003). Student Retention Of Course Content Is Improved By Collaborative-Group Testing. *Advances in Physiology Education, 27*(3), 102-108. doi:10.1152/advan.00041.2002

Dahlström, O. (2012). Learning during a collaborative final exam. *Educational Research and Evaluation, 18*(4), 321-332. doi:10.1080/13803611.2012.674689

Dallmer, D. (2004). Collaborative test taking with adult learners. *Adult Learning, 15*, 4-7. doi:10.1177/104515950401500301

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification: Using game-design elements in non-gaming contexts. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems* (p. 2425). Vancouver, US. doi:10.1145/1979742.1979575

Dicheva, D., Dichev C., Agre, G., & Angelova, G. (2015). Gamification in Education: A Systematic Mapping Study. *Educational Technology & Society, 18*(3), 75–88.

Ding, L. (2014). Seeking missing pieces in science concept assessments: Reevaluating the Brief Electricity and Magnetism Assessment through Rasch analysis. *Physical Review Special Topics - Physics Education Research, 10*(1). doi:10.1103/PhysRevSTPER.10.010105

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics - Physics Education Research, 2*(1). doi:10.1103/PhysRevSTPER.2.010105

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), 331-350. doi:10.1080/03075079912331379935

Docktor, J. L., Strand, N. E., Mestre, J. P., & Ross, B. H. (2015). Conceptual problem solving in high school physics. *Physical Review Special Topics - Physics Education Research, 11*(2). doi:10.1103/PhysRevSTPER.11.020106

Doran, R. L., Boorman, J., Chan, F., & Hejaily, N. (1993). Alternative assessment of high school laboratory skills. *Journal of Research in Science Teaching, 30*(9), 1121-1131.

Duane, B. T., & Satre, M. E. (2014). Utilizing constructivism learning theory in collaborative testing as a creative strategy to promote essential nursing skills. *Nurse Education Today, 34*(1), 31-34. doi:10.1016/j.nedt.2013.03.005

Earl, L. M. (2004). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.

Fakcharoenphol, W., & Stelzer, T. (2014). Physics exam preparation: A comparison of three methods. *Physical Review Special Topics - Physics Education Research, 10*(1). doi:10.1103/PhysRevSTPER.10.010108

Fernandezbreis, J., Castellanosnieves, D., & Valenciagarcia, R. (2009). Measuring individual learning performance in group work from a knowledge integration perspective. *Information Sciences, 179*(4), 339-354. doi:10.1016/j.ins.2008.10.014

Franklin, S. V., & Hermsen, L. M. (2014). Real-time capture of student reasoning while writing. *Physical Review Special Topics - Physics Education Research, 10*(2). doi:10.1103/PhysRevSTPER.10.020121

Garriso, C., & Ehringhaus, M. (2007). *Formative and summative assessments in the classroom.* Retrieved on March 25, 2016 from http://www.amle.org/Publications/WebExclusive/Assessment/tabid/1120/Default.aspx

Gery, F. W. (1972). *Does mathematics matter? In Research Papers in Economic Education*, A. Welsh (Ed.), 142–157. New York: Joint Council on Economic Education.

Gilley, B., & Clarkston, B. (2014). Research and Teaching: Collaborative Testing: Evidence of Learning in a Controlled In-Class Study of Undergraduate Students. *Journal of College Science Teaching, 043*(03), 83.

Guest, K. E., & Murphy, D. S. (2000). In support of memory retention: A cooperative oral final exam. *Education, 121*, 350-354.

Haberyan, A. & Barnett, J. (2010). Collaborative testing and achievement: are two heads really better than one? *Journal of Instructional Psychology, 37*(1), 32-41.

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. doi:10.1119/1.18809

Halliday, D., Resnick, R. & Walker, J. (2001). *Fundamentals of Physics*, *3*. New Jersey: John Wiley & Sons, Inc.

Hanus, M., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education, 80*, 152-161.

Harlen, W., & James, M. (1997). Assessment and Learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice,* 4(3), 365-379. doi:10.1080/0969594970040304

Hendrix, J. C. (1996). Cooperative Learning: Building a Democratic Community. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 69*(6), 333-336.

Hestenes, D., Wells, M., & Swackhammer, G. (1992). Force Concept Inventory. *The Physics Teacher, 30*, 141-158. doi:10.1119/1.2343497

Hickey, D. T., Taasoobshirazi, G., & Cross, D. (2012). Assessment as learning: Enhancing discourse, understanding, and achievement in innovative science curricula. *Journal of Research in Science Teaching, 49*(10), 1240-1270.

Hitt, G. W., Isakovic, A. F., Fawwaz, O., Bawa'Aneh, M. S., El-Kork, N., Makkiyil, S., & Qattan, I. A. (2014). Secondary implementation of interactive engagement teaching techniques: Choices and challenges in a Gulf Arab context. *Physical Review Special Topics - Physics Education, 10*(2). doi:10.1103/PhysRevSTPER.10.020123

Ifenthaler, D., Eseryel, D., & Ge, X. (2012). *Assessment in game-based learning: Foundations, innovations, and perspectives*. New York: Springer.

Ives, J. (2014). Measuring the Learning from Two-Stage Collaborative Group Exams. In *2014 PERC Proceedings* (p. 123). Minneapolis, US. doi:10.1119/perc.2014.pr.027

Jensen, M., Moore, R., & Hatch, J. (2002). Cooperative Learning: Part I: Cooperative Quizzes. *The American Biology Teacher, 64*(1), 29-34. doi:10.1662/0002-7685(2002)064[0029:CLPICQ]2.0.CO;2

Jolliffe, W. (2007). *Cooperative learning in the classroom: Putting it into practice*. London: Paul Chapman.

Kagan, S. (1990). The structural approach to cooperative learning. *Educational Leadership, 47*, 12-15.

Kapitanoff, S. H. (2009). Collaborative testing: Cognitive and interpersonal processes related to enhanced test performance. *Active Learning in Higher Education, 10*(1). http://psycnet.apa.org/doi/10.1177/1469787408100195

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics. *Physical Review Special Topics - Physics Education Research, 5*(1). doi:10.1103/PhysRevSTPER.5.010101

Kruglak, H. (1965). Experimental Study of Multiple-Choice and Essay Tests. I. Am. J. Phys. *American Journal of Physics, 33*(12). doi:10.1119/1.1971143

Lindsey, B. A., & Nagel, M. L. (2015). Do students know what they know? Exploring the accuracy of students' self-assessments. *Physical Review Special Topics - Physics Education Research, 11*(2). doi:10.1103/PhysRevSTPER.11.020103

Looney, J. (2011). Integrating Formative and Summative Assessment. *OECD Education Working Papers, 58*(5).

Lusk, M. & Conklin, L. (2003) Collaborative testing to promote learning. *Journal of Nursing Education*, *42*(3), 121-124.

Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics - Physics Education Research Phys. Rev. ST Phys. Educ. Res., 9*(2). doi:10.1103/PhysRevSTPER.9.020121

Maries, A., & Singh, C. (2013). Exploring one aspect of pedagogical content knowledge of teaching assistants using the test of understanding graphs in kinematics. *Physical Review Special Topics - Physics Education Research, 9*(2). doi:10.1103/PhysRevSTPER.9.020120 · Source: arXiv

McTighe, J., & O'Connor, K. (2005). Seven practices for effective learning. *Educational Leadership, 63*(10), 10+17.

Moccozet, L., Tardy, C., Opperecht, W., & Leonard, M. (2013). Gamification-based assessment of group work. In *Proceedings of the Interactive Collaborative Learning (ICL) Conference* (p. 171). Kazan, RU.

Moncada, S., & Moncada, T. (2014). Gamification of Learning in Accounting Education. *Journal of Higher Education Theory and Practice, 14*, 9.

Pawl, A., Teodorescu, R. E., & Peterson, J. D. (2013). Assessing class-wide consistency and randomness in responses to true or false questions administered online. *Physical Review Special Topics - Physics Education Research, 9*(2). doi:10.1103/PhysRevSTPER.9.020102

Pollock, S. J., Finkelstein, N. D., & Kost, L. E. (2007). Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? *Physical Review Special Topics - Physics Education Research, 3*(1). doi:10.1103/PhysRevSTPER.3.010107

Rao, S. P., Collins, H. L., & DiCarlo, S. E. (2002). Collaborative testing enhances student learning. *Advances in Physiology Education, 26*(1). doi:10.1152/advan.00032.2001

Rebello, N. S. (2011). Comparing students' performance on research-based conceptual assessments and traditional classroom assessments. In *2011 PERC Proceedings* (pp. 66-68). Omaha, US.

Sadler, T. D., Romine, W. L., Stuart, P. E., & Merle-Johnson, D. (2013). Game-Based Curricula in Biology Classes: Differential Effects among Varying Academic Levels. *Journal of Research in Science Teaching, 50*(4), 479-499. doi:10.1002/tea.21085

Sandahl, S. S. (2010). Collaborative testing as a learning strategy in nursing education. *Nursing Education Perspectives, 31*(3), 142-147.

Sawtelle, V., Brewe, E., & Kramer, L. H. (2012). Exploring the relationship between self-efficacy and retention in introductory physics. *Journal of Research in Science Teaching, 49*(9), 1096-1121. doi:10.1002/tea.21050

Schuwirth, L. W., & Cees P M Van Der Vleuten. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Med Educ Medical Education, 38*(9), 974-979. doi:10.1111/j.1365-2929.2004.01916.x

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*, 39-83. Chicago, IL: Rand McNally.

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies, 74*, 14-31.

Shindler, J. V. (2003). "Greater Than the Sum of the Parts?" Examining the Soundness of Collaborative Exams in Teacher Education Courses. *Innovative Higher Education, 28*(4), 273-283.

Simpkin, M. G. (2005). An Experimental Study of the Effectiveness of Collaborative Testing in an Entry-Level Computer Programming Class. *Journal of Information Systems, 16*, 273-280.

Slavin, R. E. (2000). *Cooperative Learning: Theory, research, and practice*. New Jersey: Prentice-Hall.

Slepkov, A. D., & Shiell, R. C. (2014). Comparison of integrated testlet and constructed-response question formats. *Physical Review Special Topics - Physics Education Research, 10*(2). doi:10.1103/PhysRevSTPER.10.020120

Stang, J. B. & Roll, I. (2014). Interactions between teaching assistants and students boost engagement in physics labs. *Physical Review Special Topics - Physics Education Research, 10*(2). doi:10.1103/PhysRevSTPER.10.020117

Sung, H. & Hwang, G. (2013). A collaborative game-based learning approach to improving students' learning performance in science courses. *Computers & Education, 63*, 43-51.

Talanquer, V., Bolger, M., & Tomanek, D. (2015). Exploring prospective teachers' assessment practices: Noticing and interpreting student understanding in the assessment of written work. *Journal of Research in Science Teaching, 52*(5), 585-609. doi:10.1002/tea.21209

Taras, M. (2005). Assessment – Summative and Formative – Some Theoretical Reflections. *British Journal of Educational Studies, 53*(4), 466-478.

Taras, M. (2009). Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education, 33*(1), 57-69. doi:10.1080/03098770802638671

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning and assessment in the classroom*. Buckingham: Open University Press.

Vercellati, S., Michelini, M., Santi, L., Sokolowska, D., & Brzezinka, G. (2013). Investigating MST curriculum experienced by eleven-year-old Polish and Italian pupils. In *E-Book Proceedings of the ESERA 2013 Conference: Science Education Research For Evidence-based Teaching and Coherence in Learning* (Vol. 10, pp. 180-189). Cyprus.

Wilcox, B. R., & Pollock, S. J. (2014). Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics. *Physical Review Special Topics - Physics Education Research, 10*(2). doi:10.1103/PhysRevSTPER.10.020124

Wilcox, B. R., & Pollock, S. J. (2015). Upper-division student difficulties with the Dirac delta function. *Physical Review Special Topics - Physics Education Research, 11*(1). doi:10.1103/PhysRevSTPER.11.010108

Wilcox, B. R., Caballero, M. D., Baily, C., Sadaghiani, H., Chasteen, S. V., Ryan, Q. X., & Pollock, S. J. (2015). Development and uses of upper-division conceptual assessments. *Physical Review Special Topics - Physics Education Research, 11*(2). doi:10.1103/PhysRevSTPER.11.020115

Wiliam, D., & Black, P. (1996). Meanings and Consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537-548.

Wininger, S. (2005). Using Your Tests to Teach: Formative Summative Assessment. *Teaching of Psychology, 32*(3), 164-166.

Wooten, M. M., Cool, A. M., Prather, E. E., & Tanner, K. D. (2014). Comparison of performance on multiple-choice questions and open-ended questions in an introductory astronomy laboratory. *Physical Review Special Topics - Physics Education Research, 10*(2). doi:10.1103/PhysRevSTPER.10.020103

Yu, H., & Li, H. (2014). Group-based Formative Assessment: A Successful Way to Make Summative Assessment Effective. *TPLS Theory and Practice in Language Studies, 4*(4), 839-844. doi:10.4304/tpls.4.4.839-844

Zimbardo, P. G., Butler, L. D., & Wolfe, V. A. (2003). Cooperative College Examinations: More Gain, Less Pain When Students Share Information and Grades. *The Journal of Experimental Education, 71*(2), 101-125. doi:10.1080/00220970309602059

Zwolak, J. P., & Manogue, C. A. (2015). Assessing student reasoning in upper-division electricity and magnetism at Oregon State University. *Physical Review Special Topics - Physics Education Research, 11*(2). doi:10.1103/PhysRevSTPER.11.020125

http://www.ejmste.com