International Journal of Engineering Education Vol. 32, No. 3(A), pp. 1141–1150, 2016
Printed in Great Britain

0949-149X/91 $3.00+0.00

Evaluation of Difficulty and Complexity of Tasks: Case Study of International On-line Competition "Beaver"*

EKATERINA YAGUNOVA
Saint-Petersburg State Electro technical Institution LETI after V.I.Ulianov (Lenin), ul. Professora Popova 5, St. Petersburg, Russia. St Petersburg Academic University, "Physical-Technical School" Lyceum, 8/3 Khlopina Str, St Petersburg, Russia.
E-mail: katrin.home@mail.ru

SERGEI POZDNIAKOV
Saint-Petersburg State Electro technical Institution LETI after V.I.Ulianov (Lenin), ul. Professora Popova 5, St. Petersburg, Russia.
E-mail: pozdnkov@gmail.com

NINA RYZHOVA
State Corporation 'Institution of Training—ARB Pro', Kaluzsky per, 3, St Petersburg, Russia. E-mail: ryzhova.nina@gmail.com

EVGENIIA RAZUMOVSKAIA
The University of Edinburgh, Old College, South Bridge, Edinburgh, United Kingdom, EH8 9YL. E-mail: evgeniar@yahoo.com

NIKOLAY KOROVKIN
St. Petersburg State Polytechnic University, 29 Polytechnicheskaya st., St. Petersburg, Russia. E-mail: nikolay.korovkin@gmail.com

Interest in the assessment of the quality of traditional educational techniques has grown, especially in relation to the subjects like IT or Computer Science because Informatics tests carried out using computers enable us to evaluate the quality of tasks using log-files. The study focuses of the assessment of difficulty and complexity of tasks for schoolchildren. Based on the analysis of results of 6588 participants in the international informatics competition 'Beaver–2012', it is shown that often usual a priori evaluation made by the organizers of the competition does not correspond to the task's real difficulty for the participant. A cluster of tasks the difficulty of which was underestimated was distinguished. The correlation between the length of the statement and difficulty for primary school children was shown. In order to make the results of the test more valid, a way of dividing the tasks according to their difficulty and complexity was found. Based on the method, recommendations for the organizers of the tests for general public were formulated in order to make the measures of educational outcomes of computer engineering knowledge more valid and accurate.

Keywords: complexity of tasks; difficulty of tasks; informatics; on-line competitions; educational tests; typology of tasks; competition "Beaver"; computer engineering education

1. Introduction

1.1 Competition as a measuring procedure

Any competition in a school subject is in fact a ***test***, i.e., it represents a standard evaluation procedure. During "The Beaver" on-line competition the competence and skills of school children in Informatics have been estimated. Mandatory test attributes are: standard set of tasks, standard representation of tasks, formal description of answers and processing procedure, adequate test key. A test key is an algorithm of mapping the protocol with answers at some point on a scoring scale. The main requirement on the key is its **matching with a measurable feature** [1]. Only upon availability of given attributes one can make ***an objective evaluation of an individual using a quantitative scale indicating the evidence of measurable feature*** [2].

The simplest test key suggests the summation of all "values" of performed tasks. The "values" of test tasks may have the same or different weights.

There are two essentially different approaches to determine a task value in the event of test with differential values: a priori determination of task "weights" by organizers (experts) or an a posteriori "weighting" while taking into account the results of test performance by participants. In case of an a posteriori "weighting" a number of participants who found the answer shall determine the weight of task. The greater value shall be attributed to tasks which have been performed by minority of participants.[1] More complicated tasks are also supposed to have greater weights in the event of an a priori weighting.

With the most simple key administration (a participant receives 1 point for any right answer and null for a wrong answer) it was assumed, according to [3], to present the tasks to participants

[1] For example, task weights during competitions on programming are usually attributes in this way: http://codeforces.ru/blog/entry/4172, http://contest.yandex.ru/cpr_rules.html

* Accepted 12 February 2016.

in increasing complexity. It should be noted that an a priori opinion on task complexity is required, at least, to divide the tasks into several levels of complexity.

When weighting a priori, the consistency of competition results shall be defined by experts' proficiency. [2] underlines that "it is not worth being under an illusion that experts can truly assess the complexity of tasks". According to him, ***the best complexity measure is the statistics of real answers given by real participants.***

1.2 Methods for estimating task complexity and difficulty

A task which is easy for one participant may be difficult for another one. The task difficulty reflects the relationships between a task and an individual who performs it. To underline this feature many authors separate the notions of "complexity" and "difficulty" [1, 4–6]. Complexity means a certain objective feature of a task while the difficulty is understood as a subjective feature, i.e. how a participant interprets a task. While speaking about the difficulty the authors focus on the individual's activity to perform a task—to analyze and to process the information, to design and to make decisions, to forecast consequences of their own decisions and to build operation images and frameworks [7, 8].

The task complexity may be measured upon competition results by counting a share of participants who got right answers. A measurement or at least evaluation of task difficulty requires serious efforts.

The difficulty of a task for a subject contains their mental workload (cognitive, informational, emotional, attentional loads) and expenditures for their own state control [6].

The most accurate methods of work load estimation suggest measuring of various human factors. A diagnostic procedure may be accomplished only during "live" competitions with limited number of participants. The procedure itself may be an additional stress factor for participants while increasing their efforts to control their states [9]. Meanwhile, it is only possible to assess the participants' state of remote competitions with the help of self-reflection tests on their state during task performance. The results of such questionnaires shall be adequate only for senior school students because it will be difficult for primary school children to make an objective evaluation of their state and abilities [10]. According to [11], a child of primary school is at the stage of concrete operational intellectual development. Typically for this age, thinking restrictions affect not only the cognition of outside world but the manner of children to perceive themselves. It is fair to start talking about conception thinking building only by the age of 11–12 years.

Not only the appropriateness of self-assessment bus also is the level of thinking process development associated with the age. That is why primary school children may not cope with solution of tasks that require the operation with abstract notions (and this is natural!) [11]. During the period from 8 to 10 years the capacity of memory is rapidly increasing, attention can be switched much better. So, even minor disparity of years in this period may cause significant differences of results when solving the same tasks.

According to [5], the results of "human system" activity exhibit the relationship between the quality of operation information (quality and quantity of stimuli, coding, distribution etc.) and the capacity of resources available [5]. In subject competitions the attention load, the processes of short-term and operative memory may be assessed with the difficulty of the text of problem statement. Among numerous ways of assessment of the difficulty of text, according to [12], the most straightforward is the length of the statement (number of stimuli to be processed for solution).

According to [13], the workload may be a function of the level of difficulty and the number of tasks to be performed within a unit of time. Using the protocols of on-line competitions one can evaluate the workload of participants with due account made for the time spent by them to solve tasks as [14] perceptively stated it.

The rules of "The Beaver" competition assume the presentation by participants completing only a part of tasks. In this case the participant's refusal to solve the task is to be considered as their assessment of task difficulty upon binary rating scale ("difficult"—"not difficult"). Solved tasks are assessed by a participant as "not difficult", while those that are not solved—as "difficult". A share of participants who found their task as "difficult" shall determine the difficulty of a task for the whole body of competition participants.

A prior estimate of tasks by experts (weighting) shall be correct once both the task complexity and difficulty are taken into account for participants of each age group. Only analysis of the results of competition one can establish whether the difficulty has been really considered at weighting, whether a prior estimate is in agreement with an objective complexity.

It should be noted that some authors use the notions of "complexity" and "difficulty" as synonyms because not the content of mental processes is fundamental for them, but is the execution—whether a schoolchild can or cannot solve the task [15]. Further on we shall differ the terms "complex-

ity" and "difficulty" by highlighting them in bold. If the matter is a complex evaluation with account made for both parameters, we shall use the term "complexity" without highlighting it.

1.3 Research objectives

1. To evaluate the validity of measurement procedure to be performed during processing of the results of international competition in informatics "The Beaver". To assess the quality of a set of competition tasks and scoring method.
2. To estimate the adequacy of expert estimate of task complexity. To compare various estimates of task complexity and difficulty.
3. To classify competition tasks upon their complexity and difficulty.
4. To evaluate age differences of task perceiving by schoolchildren.

2. Methodology of the research

2.1 "The Beaver" competition: organization, tasks selection

"The Beaver" international on-line competition in Informatics started in 2003. Russia took part in this competition for the first time in 2012. The task pool is prepared by representatives of participating countries. Out of this pool each country prepares its own versions.

The competition in Russia is organized for six age groups. The participants are proposed to solve the tasks of three levels of complexity: for schoolchildren of 1 and 2 grades—4 tasks for each group (weighted of 3, 6 and scores respectively), for schoolchildren of 3–10 grades—5 tasks for each group (weighted of 6, 9 and 12 scores). 8 simple and 7 complicated tasks (weighted of 9 and 12 scores respectively) are proposed to senior schoolchildren 40 minutes are given for task completion. Every wrong answer is fined, the penalty rate makes one-third of task value.

Competition tasks are numbered, simple tasks are of smaller order and complicated tasks have larger numbers. The tasks of the same complexity are randomly numbered (the order is different for each participant). Participants may solve the tasks arbitrarily, they may return to solved tasks.

Participants are to choose from multiple answers, three of them are wrong and one is correct. A participant can choose the option "no answer". In this case they will receive neither score, nor penalty. Certain tasks prepared for students of 1–2 grades are dynamic—certain actions with the mouse are required to be made. These tasks will be scored in the same way as others. A sum of scores and penalties gained shall make the result of competition for each participant. Participant who gained the highest results shall become winners in each age group.

2.2 Scales and analysis

The protocol of competition where all participants' actions are recorded shall be taken as a basic data set. The time when a participant started his/her work and the time spent for selection of every answer is known. If a participant introduced successively several answers the last one is to be taken into consideration when counting the results. The total number of competition participants in 2012 was 6588 schoolchildren.

A number of scales have been used to evaluate the complexity and difficulty of competition tasks.

Scale 1—scale of expert estimation. It is performed upon three-point scale (1—simple, 3—complicated) during the meeting of international steering committee. In order to consider the estimate as correct both the objective complexity and difficulty of tasks should be taken into account.

Scale 2—share of participants who chose the "no answer" option for a particular task. When the answer is arbitrarily chosen out of 4 proposed options, the probability (p) to choose a correct answer makes $\frac{1}{4}$. In such case a mathematical expectation of scores gained is $p*x+ (1–p)*(–x/3) = x (4p–1)/3 = 0$ (where x – task value), that coincides with total scores received when choosing "no answer" option. If one of the answers proposed is rejected as certainly wrong, the probability of arbitrary choice of a correct answer out of remained options is greater than $\frac{1}{4}$ and a mathematical expectation of score gained for solving a task becomes positive. So, a participant has no reasonable motives to choose "no answer" button. The use of this button may only be due to psychological reasons – for example, a fear of failure (and penalty for it) or an extreme lack of self-confidence, a fear of task statement. As statistical expectation of score may be hardly calculated by schoolchildren, the choice of "no answer" option reflects their instinctive presentation of relation between winning and failing probabilities. In all cases the choice of "no answer" option is the result of an interaction between the schoolchild and the task, i.e. it features the difficulty of a task for a participant.

Scale 3—share of participants who gave a correct answer among those who decided to solve the problem. It shall be determined upon completion of a competition and shall contribute to proper evaluation of the task complexity. We must underline that this is just a task complexity because it is calculated only taking into consideration those who decided to solve it, i.e. among those who evaluated it as "not difficult" upon scale 2.

Scale 4—number of symbols in problem statement. This is an indirect assessment of task difficulty because it relates to memory and attention loads.

Because scales 1–3 are ordinal, the comparison of scale 1 with scales 2 and 3 has been made by using Spearman's correlation coefficient.

Clustering of tasks has been made by Ward's method while using Euclidean metric.

In order to assess the adequacy of scoring method chosen for this competition and to evaluate the quality of a set of competition tasks by using Lilliefors test, the control of competition results normality has been done. The following hypothetical procedures have been implemented as an alternative to used scoring method (tasks of various values set a priori with penalties for the wrong answer):

- tasks of various values established a priori without penalty scores for the wrong answer;
- tasks of the same value with penalty scores for the wrong answer;
- tasks of the same value without penalty scores for the wrong answer.

3. Results

3.1 Competition results

The normal-theory test of results of schoolchildren (composite score you got) is rejected at a level of significance $p < 0.05$ for all school grades except for 6th, 7th and 8th. The significance of Kolmogorov-Smirnov test is given on Fig. 1 (row 1). For junior (1–6) and higher (9–11) grades the distribution of results has a positive skew, i.e. relatively few schoolchildren have shown high results, while the most of them have low results.

An analysis of hypothetical versions of competition results estimation has shown that for all grades from 3rd the scoring method used is the best out of those with a priori estimation of tasks (Fig. 1). Abolishment of penalties, when tasks have different weights, and equalized tasks, when penalties are available, make a real distribution of results insignificantly worse. The most unsuccessful is the version of tasks having the same weights without penalties. For all tested hypothetical scoring methods even the distribution of results of 6–8 grades becomes far from being normal. As far as it concerns schoolchildren of junior grades (1–2) the situation is quite different: a real scoring method proved to be the least successful while penalty exception or equalizing of task weights improves the distribution. It should be noted once again that only in three of all considered cases deviations vs. normal distributions may be found incidental.

3.2 Task complexity and difficulty

Task assignments upon difficulty (scale 2) and complexity (scale 3) are far from normal (Figs. 2A-B). The skewness factor of task complexity distribution is positive while that of task difficulty distribution is negative, i.e. tasks of low difficulty prevail in competition but the most part of tasks are of high complexity. There are test versions for all grades where difficulty makes less than 10%. The task of the lowest difficulty was found in a test version for the 8th grade (1.1%) and of the highest

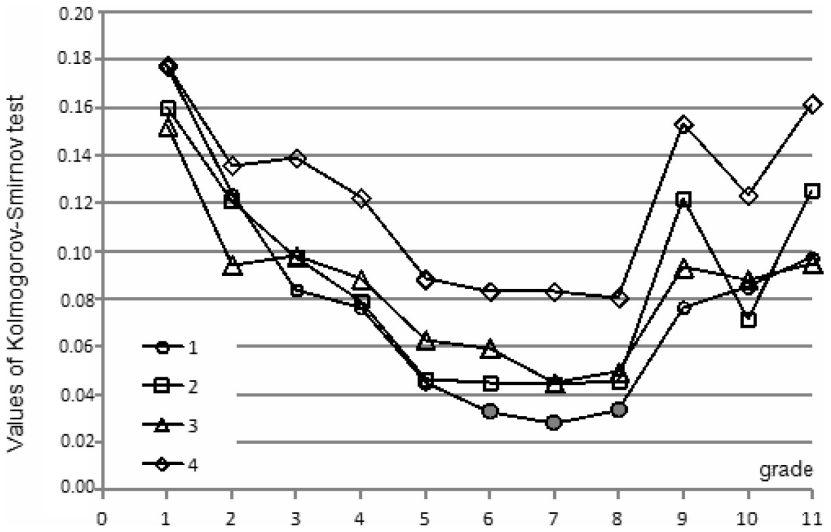

Fig. 1. Normality test of competition results at real and hypothetical scoring methods. Values of Kolmogorov-Smirnov test are given for 1—a priori assignment of different weights to tasks and penalties for the wrong answer; 2—a priori assignment of different weights to tasks without penalties for the wrong answer; 3—tasks having the same weights with penalties for the wrong answer; 4—tasks having the same weights without penalties for the wrong answer. Filled circles are distributions whose deviations vs. normal may be considered incidental at $p < 0.05$.

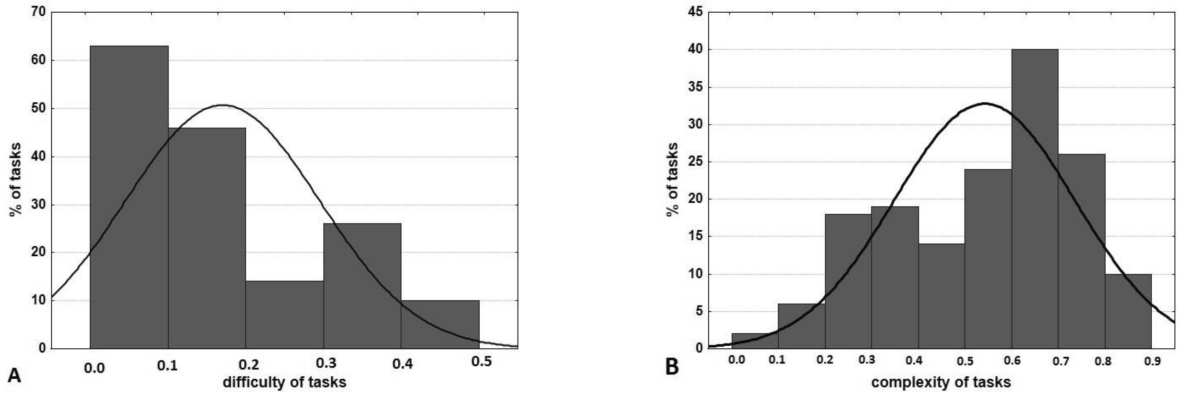

Fig. 2. A. Distribution of competition tasks upon their **difficulty.** B. Distribution of competition tasks upon their **complexity**.

difficulty—in a test version for the 10th grade (47.6%). The task complexity within test versions for each grade varies from 10–20 to 80–90%. The lowest complexity task was found in the 2nd grade version (7.7%) and the one of the highest complexity—in the 7th grade version (89.8%).

3.3 Adequacy of expert evaluation of task complexity

Spearman's rank correlation coefficients of expert tasks evaluation (scale 1) with their complexity (scale 3) and difficulty (scale 2) for participants calculated for all competition tasks are significantly positive ($p < 0.01$) and equal to 0.56 and 0.60 respectively. Table 1 exhibits rank correlation coefficients of expert evaluation between their complexity and difficulty for each grade. For junior grades (1, 2, 3) versions there was no correlation of expert evaluation of complexity with task difficulty revealed. For grades starting with the 4th there is a significantly positive correlation. The best concurrence of expert evaluation with task complexity has been found for 1-2 grades. As far as it concerns the versions for 7-8 and 11 grades the evaluation of task complexity by organizers did not correspond to real complexity of tasks for school students.

3.4 Task classification

By comparing the values of task complexity and difficulty with the use of cluster analysis 4 clusters have been marked (Fig. 3). Two clusters contain highly complicated tasks while remaining two clusters are of low and medium complexity. There was a cluster of tasks of high difficulty and the one of low difficulty tasks separated among highly complicated tasks. All tasks of low and medium-case complexity are of low difficulty.

3.5 Link between task statement length and its complexity and difficulty

Table 1 shows correlation coefficients between the number of characters in task statement (scale 4) and its complexity and difficulty (scales 2 and 3). In junior grades (from the 1st to the 4th) a significantly positive link ($p < 0.01$) has been found between task statement length and its difficulty (number of "no answer" responses). Moreover the link between the task statement length and its objective complexity was revealed only for tasks of the 3rd–4th grades.

The questionnaire among the participants of the competition was carried out in 2013. Two questions were as follows: "I didn't like some of the tasks because: (a) they have too lengthy statements; (b) we did not cover it at school; (c) the pictures were not clear; (d) I had to be very attentive"; "Sometimes I refused to solve the task straightaway if I saw that (a) the statement is too lengthy; (b) the task will take too much time; (c) the task statement did not fit on the screen; (d) we did not cover it at school; (e) I did not refuse to solve any tasks, I tried to solve all the tasks." In each of the questions, one of the listed choices had to be chosen. About 20% of more than 17000 participants responded to the questionnaire (Table 2).

Most of the participants answered that they "tried to solve all the tasks" on the question of which tasks they did not try to solve in the competition. The vast majority of those who agreed with the thesis "I refused to solve the task straightaway" indicated as the main cause that the task will take

Table 1. Spearman's rank correlation coefficients of expert evaluation of tasks complexity for each grade with their **complexity** and **difficulty.** Correlation coefficients significantly at $p < 0.05$ are highlighted in bold

Grade	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Task **difficulty**	0.41	0.46	0.51	**0.67**	**0.53**	**0.76**	**0.76**	**0.76**	**0.63**	**0.79**	**0.84**
Task **complexity**	**0.74**	**0.8**	**0.62**	**0.64**	**0.59**	**0.57**	0.28	0.25	**0.62**	**0.66**	0.49

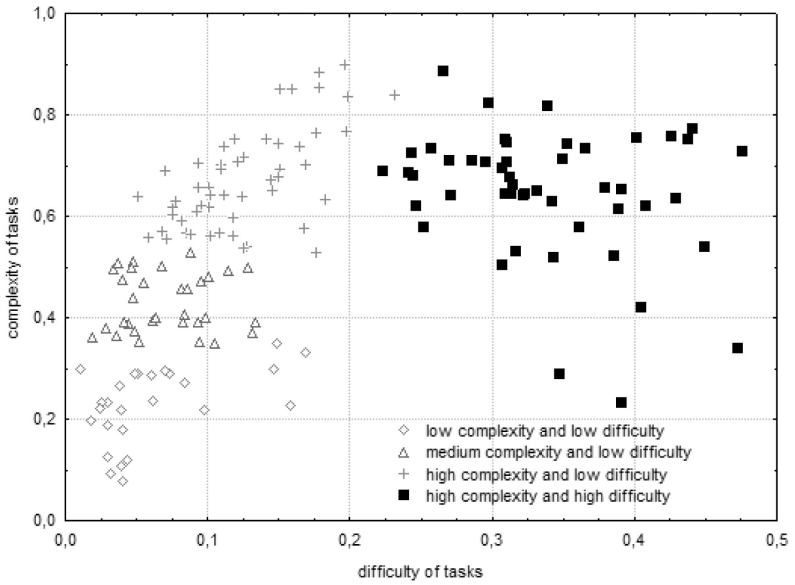

Fig. 3. Task clusters marked on the basis of comparison of their difficulty and complexity.

Table 2. The number and proportion of participants who responded to the questionnaire in each year

Year		1	2	3	4	5	6	7	8	9	10	11
Girls	number	28	115	205	297	192	222	174	165	154	112	70
	proportion	0.1	0.22	0.22	0.24	0.18	0.25	0.21	0.21	0.2	0.22	0.28
Boys	number	20	96	220	291	218	231	211	201	249	166	119
	proportion	0.08	0.19	0.24	0.24	0.2	0.24	0.22	0.22	0.22	0.21	0.29

too much time. The second most frequent cause was the length of the statement. It was mostly chosen by boys and junior schoolchildren.

At the same time, the high school children and girls marked long statement to be the main cause due to which they did not like the task. In the younger children, the main cause is the demand of high attention.

Therefore, we can see that too lengthy statements cause additional effort for all the participants of the competition. Too lengthy statement can be the cause to refuse to solve the task for junior children, while for the older ones it is not the cause to refuse to solve the task, but it is the cause to dislike the task. The task statement has bigger importance for boys than for girls. They refuse to solve the tasks because of too long statements more frequently. Also, the males responded more often than the females that they did not like the tasks with long statement.

The responses of the schoolchildren on the questionnaire confirm the result that the length of the statement of the task is one of the characteristics of its difficulty, especially for the younger pupils.

3.6 Age differences in task perception and competition results

In each of the first five competition levels schoolchildren of two grades took part. The results of comparison of junior and senior schoolchildren in each pair of grades are shown in Fig. 5. At each level (i.e. among schoolchildren who were solving the same tasks) the results of younger participants were lower (Fig. 5A). At that junior schoolchildren assessed tasks difficulty of first-second levels of competition as higher (i.e. they chose "no answer" response, Fig. 5B) and less high at fourth-fifth levels. Tasks were complicated (number of wrong answers, Fig. 5C) for junior children at all competition levels except the first one. We have to note that it is not correct to compare the results of children of different levels because the number and sets of tasks differed at different levels.

Table 3 indicates the task clustering by grades. When comparing junior and senior grades within competition levels we note that all tasks for senior schoolchildren with a single exception are of the same or of less complexity and difficulty. Only ninth task of fifth competition level for juniors (9th grade) is of low difficulty and of high difficulty for seniors (10th grade), task complexity for all participants is the same, and it is high.

4. Discussion

4.1 Validity of the measuring procedure

While considering both the distribution of competi-

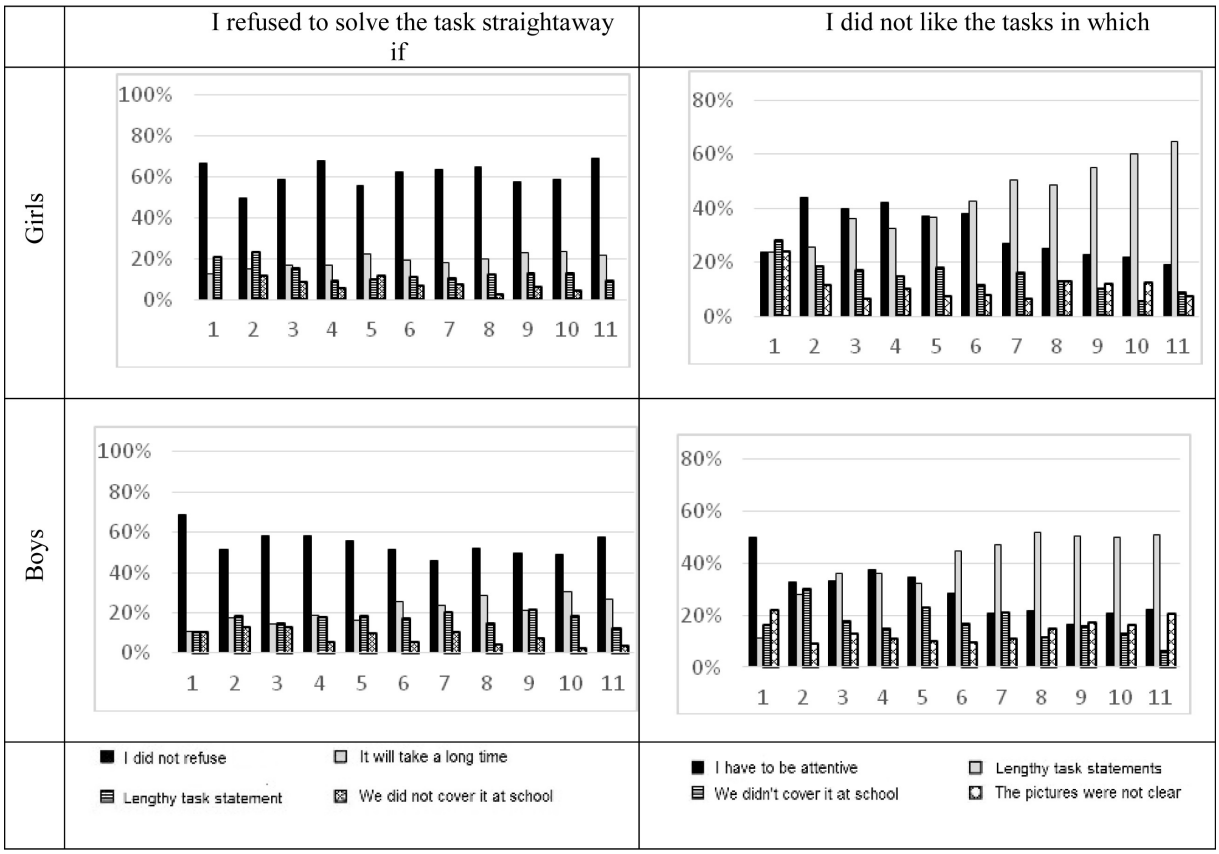

Fig. 4. Distribution of the responses of the schoolchildren on the questionnaire.

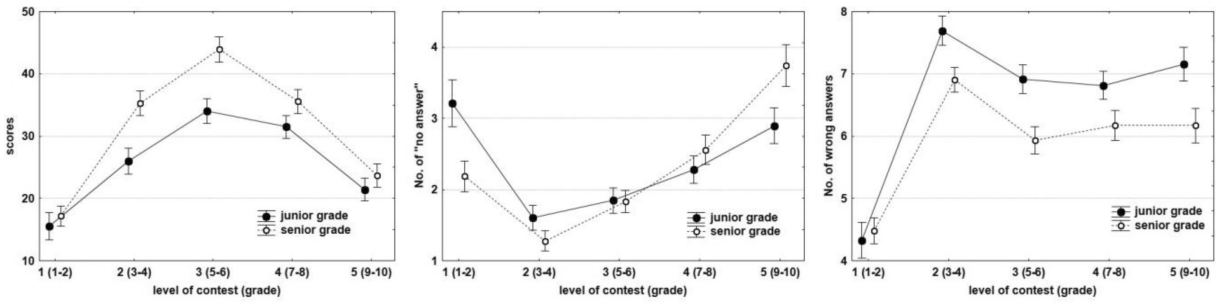

Fig. 5. Differences of results of junior and senior schoolchildren at each competition level. A. Differences of average score B. Differences in assessment of task **difficulty**. C. Differences in task **complexity**. 95%-confidence interval of averages is given.

Table 3. Spierman's rank correlation coefficients of task statement length and its **difficulty** and **complexity.** Correlation coefficients significant at $p < 0.05$ are bolded

Grade	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Task difficulty	**0.90**	**0.83**	**0.77**	**0.78**	0.17	0.13	0.04	0.06	0.10	0.04	0.14
Task complexity	0.37	0.36	**0.77**	**0.78**	0.29	0.32	0.18	0.20	0.13	0.01	0.23

tion results and the difficulty and complexity of tasks, it is possible to conclude that complicated tasks prevail in competition for school students. The most adequate to participants' level version of tasks is the one for the 7th–8th grades where the tasks of high complexity and difficulty do not make more than a half of tasks. As a consequence, just the distribution of competition results for schoolchildren of the 7th–8th grades is close to normal while in the test versions designed for the remaining groups the task complexity is beyond the abilities of participants.

Among different options of scoring supposing an a priori weighting of tasks the option chosen by

organizers (tasks of different values, penalty scores for wrong answer) is effective for a given set of tasks for all grades starting with 3rd. For juniors the introduction of penalty scores and task clustering upon complexity is an unnecessary sophistication which impairs the distribution of results which is bad enough even without that.

Due to a significant right skewness of distribution of competition composite scores the selection of winners and ranking of the strongest participants run at high point. Ranking of the main body of participants is rough. A set of tasks used for this competition could be more appropriate if its goal was to select the best students. Taking into account that the competition is aimed at the general public in order to heighten the interest in the subject and tailored for students of general education schools a set of tasks must be considered far to be successful.

4.2 Expert estimation of tasks

As it was mentioned above, the most part of tasks proved to be complicated for the majority of participants. There is a possibility to make some assumptions about the reasons for that basing on the results of comparison of expert estimation of tasks and their complexity and difficulty. As far as it concerns the tasks proposed to junior pupils the correlation between expert estimation and task complexity is high and that one related to difficulty is insignificant. The elder are pupils, the more precise was the task assessment by organizers in correlation with its difficulty set up in the protocol of competition. However, the correlation between expert opinion and task complexity was not always established in tasks designed for senior students. That is to say the experts do not evaluate sufficiently correctly the difficulty of tasks for juniors and their complexity for seniors. Our results confirm to some extent the opinion of [5] that one of the components making the task difficulty is its statement length. In junior classes the tasks became significantly more difficult with the increase in statement length. Perhaps just this factor was not taken into account by experts at estimating tasks for junior grades that resulted in making the task excessively complicated. We suppose that in junior classes it was just the statement length (underestimated by organizers) that became the factor that determined a large number of refusals to solve simple in fact tasks, so the results of measurement of knowledge and skills of juniors have been mispresented. Because of lower level of development of mental processes junior pupils misunderstand long texts. It was proved that complexity of text provokes loss of interest to its content [16], therefore, tasks with too intricate statements should be avoided. Another complicating factor is interface peculiarities of a competition. Tasks containing long texts may have not enough space to be presented on one screen. In this case to read it from the screen some skills related to computer-literacy will be needed (to know what ''vertical scroller'' means and how to use it), as well as fine motor skills shaped in a certain manner (to know to use a mouse).

For senior students the length of text is not an extra complicating factor. Moreover, during educative process a personal experience to assess the difficulty of a task by eye is being gained as well as stereotypes to differ ''difficult'' and ''simple'' tasks. That is why the tasks estimated by experts as difficult prove to be such for senior participants.

Having made mistake with assessment of task difficulty for junior schoolchildren, the experts evaluated best of all the complexity of tasks for them. As far as it concerns senior pupils the objective complexity of tasks did not coincide with an a priori opinion of organizers. It probably means that an objective complexity of a task is due to a large extent to students' knowledge. The knowledge of senior students participating in the competition in Informatics was overestimated by organizers. The impression is that the experts were oriented to select and estimate competition tasks for a standard student of secondary school.

4.3 Tasks categorization

Tasks of low difficulty prevail in competition. Pupils are more disposed to solve tasks than to choose ''no answer'' version. As shown above, the selection of ''no answer'' version gives no advantage in scores over a simple guessing. Even more interesting are the tasks where a large part of pupils refused to find a solution—these are tasks of second cluster. Most of them were found in versions designed for junior and senior students. We suppose that this cluster contains nonstandard tasks that frighten pupils by their presentation. The probability of giving a right answer is instinctively assessed as extremely low (this is not consistent with the reality—this probability makes not less than $\frac{1}{4}$). Three other clusters contain more intelligible and/or ordinary tasks. The most simple (tasks of low difficulty and complexity) are those called in competition slang as ''consolation tasks'', their solution is practically possible for each participant. The availability of such tasks in competition gives a positive emotion even to those who solved few tasks. The number of such tasks proposed for competition was negligible.

4.4 Age differences with regard to task perception

The above results show that one year age difference of participants makes significant differences with regard to the perception of the same tasks. Junior participants of elementary schools are prone to

Table 4. Task clustering per each grade. Numbers of tasks are given in cells. Arrows show changes of complexity or difficulty

		Low difficulty			High difficulty
Level	Grade	Low complexity	Average complexity	High complexity	High complexity
1	1	3-4-7	6	1	2-5-8-11-9-10-12
	2	3-4-7	6-1	9-10-12	2-5-8-11
2	3	2	4-3-5-10	1-6-8-9-11-12-13-14-15-7	
	4	2-3-5-10	4-7	1-6-8-9-11-12-13-14-15	
3	5	10	2-7-15-3-4-8	1-6-9-14-5	11-12-13
	6	10-3-4-8	2-7-15-5	1-6-9-14	11-12-13
4	7	10-2-12	1-3-6-4	5-7-8-9	11-13-14-15
	8	10-2-12-4	1-3-6	5-7-8-9	11-13-14-15
5	9	1	5-7	2-6-8-3-9	4-10-11-12-13-14-5
	10	1	5-7-3	2-6-8	4-10-11-12-13-14-15-9

assess tasks as more difficult and they are ready to choose "no answer" version more often than senior pupils. Elder pupils of elementary schools are prone in the contrary to assess tasks as less difficult (Table 4). The complexity of tasks for younger pupils of elementary school is in fact higher which becomes apparent in larger number of wrong answers and it is not surprising that it leads to lower results.

5. Conclusions

1. The tasks proposed for "The Beaver–2012" competition allow selecting the best students but does not allow grading the main body of participants. Different weights of tasks and penalty scores for wrong answer enhance the quality of measuring procedure for the 3rd–11th grades but do not make it accurate enough.
2. When preparing tasks for competition the organizers did not take into account the age difference of junior pupils, especially their low capacity to understand long texts. The standard of knowledge of senior pupils has been overestimated as well.
3. The proposed procedure of task clustering allows identifying nonstandard tasks.
4. One year difference makes significant differences with regard to perception of task difficulty and complexity by pupils and, therefore, their results.

Recommendations to competition organizers

- In general the competition needs to be simplified: to increase the number of "consolation" tasks for all grades. Nonstandard tasks for junior and middle grades are to be simplified and more carefully formulated. The tasks with long text statements should be bypassed in junior classes.
- To think over the utility of "no answer" version. Perhaps, it is worth changing the relation between added scores and penalties or the number of proposed answers, so as to add a more pragmatic sense to "no answer" version.
- To think over the possibility to determine a posteriori task weights while keeping the penalty scheme for wrong answers. Such an approach shall give a positive educative impact–the tasks "of value" shall not frighten at a first glance the participants who are not self-confident. At the same time a posteriori determination of task values shall compensate experts' mistakes at their a priori assignment.
- If school students of different grades solve the same set of tasks, the announcement of the results of competition and the selection of winners must be done separately for each grade.
- To record into protocols of competition not only the time of answer introduction but the time spent by a participant to resolve the task. By doing so it will make alternative estimation of the tasks difficulty possible.

References

1. G. A. Ball, *Theory of training tasks: Psychological and educational aspect*, Moscow, Russia: Pedagogy, 1990.
2. A. G. Shmelev, *Practical testology,* Moscow, Russia: Maska, 2013.
3. F. M. Lord, Theory of Test Scores, *Psychometric monograph*, **7**, 1952.
4. Yu. Ya. Golikov and A. N. Kostin, *Psychology of facilities management automation*, Moscow, Russia: RAS Institute of Psychology, 1996.
5. D. Navon and D. Gopher, On the economy of human

information processing systems, *Psychology Review*, **86**, 1979, pp. 214–255.
6. G. Sammer, Concepts of mental workload in psychological research, In: *Proceedings of the 13th Triennial Congress of the International Ergonomic Association*, **5**, 1997, pp. 368–370.
7. A. N. Leontiev, *Activity. Consciousness. Personality*, Moscow, Russia: Politizdat, 1975.
8. V. A. Ponomarenko, G. M. Cherniakov and V. G. Kostritsa, *Operator's mental states as the object of engineering and psychological researches*, Cybernetic issues, 2013.
9. B. Kantovits and R. Sotokin, Human factor, *Mir.,* **4**, 2013, pp. 85–113.
10. L. Towler and P. Broadfoot, Self-Assessment in the primary school, *Educational Review*, **4**(2), 1992, pp. 137–155.
11. J. Piaget, *Psychology of intelligence*, London, Great Britain: Routledge and Kegan Paul, 1951.
12. R. G. Benjamin, Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty, *Educational Psychology*, **24**, 2012, pp. 63–88.
13. V. Riley, E. Lyall and E. Wiener, Analytic workload model for flight deck design and evaluation, *Proceedings of the Human Factors and Ergonomic Society*, **38**, 1994, pp. 81–84.
14. D. Gibson and J. Clarke-Midura, *Some Psychometric and Design Implications of Game-Based Learning Analytics*, Perth, Australia: Curtin University, 2013.
15. V. M. Krotov, On complexity of problems on physics, *Physics: teaching problems*, **3**, 1999, pp. 69–74.
16. S. M. Fulmera and M. Tulis, Changes in interest and affect during a difficult reading task: Relationships with perceived difficulty and reading fluency, *Learning and Instruction.* **5**, 2013, pp. 11–20.

Ekaterina Yagunova is the research fellow at the Saint Petersburg Electrotechnical University (BSc is Mathematics and Mechanics, BSc in Biology at the Saint Petersburg State University, Ph.D). Her research interests fall into the methods of teaching, computer means of learning and data processing. She is a member of editorial board of the journal ''Computer means in education''. She is the author of more than 30 scientific papers and methodical publications. She is a mathematics teacher in schools for gifted children for over 20 years.

Sergey Pozdnyakov is the professor at the departments of higher mathematics at the Saint Petersburg Electrotechnical University and at branch of Informatics at the Mathematics and Mechanics department of the Saint Petersburg State University (Ph.D. in Education). He is the chief editor of the journals ''Computer methods in education'' and ''Computer methods at school''. He is also the chairman of Center of Information Technologies in Education CTE.

Nina Ryzhova is a methodologist, trainer-consultant and analyst of HR Consulting ''GK Institute of Training and ARB Pro'' (MA & BA in Psychology). Her main research interests fall into testing and evaluation of personnel, psychodiagnostics, diagnostics of general abilities, organizational diagnostics, education and development of adults and children. She is the author of a number of publications on evaluation of staff management systems, evaluation of the quality of testing.

Evgeniia Razumovskaia is an undergraduate student doing BSc (Hons) Cognitive Science at the University of Edinburgh. Her main research interests fall into cognitive theories of education, education and intelligence, neural networks, cognitive neuroscience, language and cognition.

Nikolay Korovkin is currently the head of Theoretical electrical Engineering Department of St.-Petersburg State Polytechnic University (MSc, Ph.D., Doctor). His main scientific areas are inverse problems, mathematical models of hard (ill-conditioned) systems and methods of their analysis, soft methods of optimization, multiextremal tasks, pedagogics, methodics of teaching and testing. He is the author of more than 250 scientific papers and methodological publications.