Olympiads in Informatics, 2019, Vol. 13, 145–156

DOI: 10.15388/ioi.2019.09

Analyzing Task Difficulty in a Bebras Contest Using Cuttle

Willem van der VEGT[1], Eljakim SCHRIJVERS[2]
[1]*Dutch Olympiad in Informatics, Windesheim University for Applied Sciences*
PO Box 10090, 8000 GB Zwolle, The Netherlands
[2]*Dutch Olympiad in Informatics, Eljakim IT*
PO Box 85183, 3508 AD Utrecht, The Netherlands
e-mail: w.van.der.vegt@windesheim.nl , eljakim@cuttle.org

Abstract. Predicting the difficulty level of a task on the concepts of computer science or computational thinking, like in the Bebras Challenge, proves to be really hard. But the announced difficulty level is needed in the contest format used in many local challenges. The Dutch contest system Cuttle has a new module for analysis. This is applied to one specific contest in order to find parameters explaining task difficulty. Using quantitative methods we were able to confirm a relation between answer types and difficulty and a tendency that tasks on data, data structures and representation are better answered than tasks on algorithms and programming.

Keywords: Bebras contest, answer types, question difficulty, P-value, Rit-value, contest system.

1. Introduction

Founded in 2005 in Lithuania, Bebras has developed into an annual International Challenge on Informatics and Computational Thinking amongst the young (Bebras, 2019). In 2018 students from over fifty countries compete in their national contest. The questions used in these challenges are mostly chosen from a common task pool, which is composed during the annual Bebras Workshop where most of the contributing countries participate. The questions are formulated in a way that no prior knowledge is required.

The contest is about computer science and computational thinking; most of the tasks are categorized as ALP: Algorithms and Programming or DSR: Data, Data Structures, and Representations. A few tasks fit in the other three categories; CPH: Computer Processes and Hardware, COM: Communications and Networking or ISS: Interactions, Systems, and Society, based on (Dagienė and Sentance, 2016). Criteria for good Bebras tasks, using a former system for classification, have been formulated by Dagienė and Futchek (2008). Dagienė and Sturupienė (2016) give an overview of current research on Bebras.

Contestants compete in their own age division. In the Netherlands we offer the challenge in the form of a contest. In the first round contestants have 40 minutes to complete 15 tasks. Tasks can have one of three answer types: multiple choice, open ended or interactive. The contest runs during one week for five different ages groups. Some countries will also have an event for the youngest age group, 6–8 years; the Dutch contest starts with grade 3; contestants are usually aged between 8 and 18 years. The best performing contestants for the four highest age divisions are invited to a university for a second round (Beverwedstrijd, 2019).

Within the contest we present tasks to the contestants as easy, medium or hard. But in practice our own classification breaks down. Earlier (Van der Vegt, 2018) we discussed ways to predict the difficulty level of specific Bebras tasks. We applied these models to the 2017 contest for the highest age group in the Netherlands. But since using the questionnaires for predicting task difficulty is very time consuming, we want to identify a few properties of a task that can be of use in predicting task difficulty. This could be helpful for the entire Bebras community, for in a lot of national challenges the announced difficulty level of a task is part of the design.

The Cuttle contest system is developed for organizing the Bebras contest in the Netherlands. This system is used in over thirty Bebras and other scientific contests. Recently it has been extended with an analysis tool. We will use this tool to investigate aspects of the tasks in a specific contest and we try to discover a relation between properties of a task and the actual difficulty of it. In this paper we will analyze the 2017 contest for the highest age group in the Netherlands, making use of the Cuttle-tool for analysis, and develop some recommendations for possible future research. We will focus on categories of tasks and answer types.

Summarizing, we will try to answer two questions: Is it possible to use the Cuttle system to collect data that can be useful for analyzing task difficulty in Bebras? And can we formulate questions for future research, based on the findings using Cuttle?

In section 2 we will give a short summary of earlier research on predicting task difficulty. Section 3 will describe the selection process to compose the contest, characteristics of the task set and the way the task proposals were developed before, at and after the Bebras Workshop. In section 4 we give an analysis of the overall results for the contest and we will look into detail to several properties of the tasks in the contest. Finally, we give some conclusions and a few ideas for a possible research agenda in section 5.

2. Task Difficulty

Since the core of a Bebras task is answering a question, we give an brief overview of research on question difficulty, focused on Bebras and similar tasks..

Lonati, Malchiodi, Monga and Morpurgo (2017) distinguish two main kinds of difficulties: on the one side intrinsic with the task, related to its content, and on the other side surface difficulties, depending on the task format and linguistic, structural and visual aspects.

Ahmed and Pollitt (1999) distinguish three kinds of difficulties in questions. Cognitive difficulty has to do with the concepts that are used in a question. The level of abstraction

of these concepts will determine this difficulty. Question difficulty is connected with the linguistic and structural properties of a question. Process difficulty is about the difficulty of the cognitive operations and the degree in which they use cognitive resources.

Leong (2006) makes a similar distinction; he considers content difficulty, depending of the subject matter being assessed, stimulus difficulty, related to comprehending words and phrases in a test item and accompanying information, and task difficulty, referring to the work needed to formulate or discover the answer to the question.

Several questionnaires or rubrics have been proposed to predict the difficulty of a task (Van der Vegt, 2018); these instruments each try to assign proper weights to the expected difficulty on content, stimulus and task performance, in different ratios. Some items are easy to measure. Dhillon (2003) for instance states that the number of components of a question and the number of times these components have to be repeated have a high impact on the difficulty level. Estimating the number of steps to perform a task is possible for an experienced task designer. Other ways of assessing topics in these questionnaires are not yet well described.

3. Tasks

3.1. *Task Selection*

In the Netherlands we work together with some other countries in the selection process to compose the contests. We receive the results of the task selection from the German speaking countries, the UK and US task pool, as well as the Belgian team. We tend to reuse tasks in more than one age group in order to reduce the total number of tasks. This way we used 34 different tasks to organize a first round in 2017 for five different age groups with 12, 15, 15, 15 and 15 tasks. Of the 15 tasks that were selected for the highest age group, 9 have also been used in the same contest but for other age groups.

For each contest the difficulty level of a task is announced as easy, medium or hard. The score a contestant can achieve depends on the expected difficulty level. For an easy task, a contestant gets 6 points for a good answer and -2 for a wrong answer. For a medium task these numbers are 9 and -3 and for a hard task 12 and -4. The original rationale behind this was that the expected score for a task when guessing should be 0. This holds only for multiple choice question with four alternatives but we have kept this scheme also for the other types of answers, open ended and interactive. If a question stays unanswered, no points are added or subtracted. To prevent negative score in case someone has only wrong answers, we start for each contestant with an initial score of 45 points.

3.2. *Task Properties*

All tasks were taken from the international Bebras task pool 2017, developed at the Bebras Workshop in Brescia. All tasks are proposed by one of the member countries,

after which they are reviewed in the preparation weeks before the workshop. During the workshop all tasks are discussed and improved. After the workshop tasks are translated and sometimes changed in order to fit into a national contest format. It is also possible that the answer type is altered in order to make the task easier or more difficult.

In this section we investigate three aspects of a task: category, answer type and difficulty level.

3.3. *Categories*

In 2017 the Bebras community has introduced five categories for tasks, based on Dagienė and Sentence (2016):

ALP: Algorithms and Programming
DSR: Data, Data Structures, and Representations
CPH: Computer Processes and Hardware
COM: Communications and Networking
ISS: Interactions, Systems, and Society

In the Bebras task pool these categories are not mandatory. Table 1 shows the suggested category for each task, a short description of the task, without the background story. Even though there are several tasks about graphs or on the assignment problem, the differences between these proposals are large enough to justify the use of all these tasks within one contest.

Only for 7 of the 15 tasks a domain was proposed by the original author. For the other tasks we did our own attribution and noted it in Table 1 between brackets. Most of the

Table 1
Categories, CS topics and answer types

Task-ID	Category	Computer Science Topic	Answer type
2017-CA-12	DSR	Dynamic programming	Multiple Choice Text
2017-IS-01	ALP	Sequence, binary system	Multiple Choice Text
2017-BE-05	(ALP/DSR)	A path in a graph	Multiple Choice Images
2017-RU-03	DSR	Gray code	Interactive
2017-IR-07	COM/ISS/ALP	Search in social network graph	Multiple Choice Text
2017-CA-07	ALP	Assignment problem	Interactive
2017-PL-02	(ALP/DSR)	Levenshtein distance	Open Ended Integer
2017-CH-01b	(ALP)	Programming in a maze	Interactive
2017-CZ-04c	ALP	A path in a graph	Interactive
2017-CH-07b	(ALP/DSR)	Maximum flow problem	Open Ended Integer
2017-KR-07	(DSR)	Image compression	Multiple Choice Text
2017-SK-12a	(ALP)	Turing machine	Multiple Choice Images
2017-UK-04	ALP	Assignment problem	Multiple Choice Text
2017-KR-03	(ALP)	Optimization, scheduling	Open Ended Text
2017-SI-04	(ALP/DSR)	Binary counting	Open Ended Integer

used categories are ALP (80%) and DSR (47%). Only one task was announced as a task both on ISS and on COM (both 7% of the tasks). The category CPH was never used.

3.4. *Answer Types*

Within the contest we used five different answer type:

- Multiple Choice Text means the classical form with four alternatives (33 %).
- Multiple Choice Images is somewhat similar; the alternatives are now presented as images (13%).
- Open Ended Integer asks the user to input a number (20 %).
- Open Ended Text ask the user to input a string (7 %).
- Interactive means the user has to perform some kind of action to solve the problem; a grader program checks the solution (27 %).

3.5. *Task Difficulty*

Due to the contest format we need to identify the difficulty level of each task, or to compare the tasks with each other. There are several problems in predicting difficulty level (Van der Vegt, 2013) and last year we experimented with several tools to help in this process (Van der Vegt, 2018). For the tasks in the 2017 contest we looked at the original task proposals, the tasks in the task pool and we made of course our own assessment. This is summarized in Table 2.

Table 2
Task difficulty estimations

Task-ID	Original difficulty level	Workshop difficulty level	Contest difficulty level
2017-CA-12	III-easy	V-medium	VI-easy
2017-IS-01	V-medium	V-hard	VI-easy
2017-BE-05	IV-medium	IV-medium	VI-easy
2017-RU-03	II-medium	IV-easy	VI-easy
2017-IR-07	IV-easy	V-medium	VI-easy
2017-CA-07	V-hard	V-medium	VI-medium
2017-PL-02	V-hard	V-hard	VI-medium
2017-CH-01b	IV-easy	V-medium	VI-medium
2017-CZ-04c	V-medium	V-hard	VI-medium
2017-CH-07b	VI-hard	V-hard	VI-medium
2017-KR-07	IV-medium	VI-hard	VI-hard
2017-SK-12a	VI-medium	VI-hard	VI-hard
2017-UK-04	VI-hard	VI-hard	VI-hard
2017-KR-03	VI-medium	VI-hard	VI-hard
2017-SI-04	V-medium	V-medium	VI-hard

3.6. *The Cuttle Contest System*

The Cuttle system is evolved from the system build for the first Bebras contests in the Netherlands. Since the early start in 2006 we have organized 43 contests, usually two round per year and some demonstration games. Within the system 723 Dutch tasks are available. Each task can get a difficulty level per age group, a category can be assigned and it is also possible to indicate a CS Skill: Abstraction, Algorithmic Thinking, Decomposition, Evaluation and Generalization. The system is also available in other languages.

In 2018 20 countries organized their national Bebras challenge with the Cuttle system: Australia, Austria, Canada, Germany, Greece, Hong Kong, Iceland, India, Ireland, Japan, Malaysia, New Zealand, Netherlands, Norway, Romania, South Africa, Switzerland, Thailand, United States and the UK, with in total almost one million contestants. The system is also used for several other Bebras-like contests.

4. Results

In this section we will apply the new Analytics part of the Cuttle-system on the first round of the 2017 contest in the Netherlands for the highest age group (16–18 years). This contest had 1621 participants. The data we study are the results of this contest for these participants. We look at the correct answers, at the fraction of the participants not answering a specific task. And we try to relate some of the numbers to the theory on question difficulty.

In Fig. 1 the number of well-answered tasks is shown (max. 15) as well as the distribution of the scores (max. 180). The distribution patterns of both correct answers and scores appear to resemble the normal distribution.

The contest system also provides general data, like the ones shown in Fig. 2. With a maximum score of 180 the whole range from 0 to 180 turned out to be possible, with an average of 92.9 and a standard deviation of 32.4.

The system gives also the detailed scores for all tasks. Table 3 shows a part of the output, focusing on a few major measures. P_{all} is the percentage of correct answers across all participants; this P-value is often used as an indication of the difficulty level (Van der Vegt, 2013). Goldebeld (1992) states that in an ideal exam all P-values should be between 30 and 70%; but since Bebras is not an exam but a challenge, we value this not as an important restriction for our tasks. The R_{it} gives the correlation between the score of a task and the overall score as a percentage. An R_{it}-score of 40% or above is seen as an indication that the task was very good fitting in a contest (Goldeberg, 1992). An R_{it}-score below 20% indicates an atypical result for a task; if such a score occurs it means you will have to investigate if there is a problem with the task. In this perspective our outcomes were very satisfying.

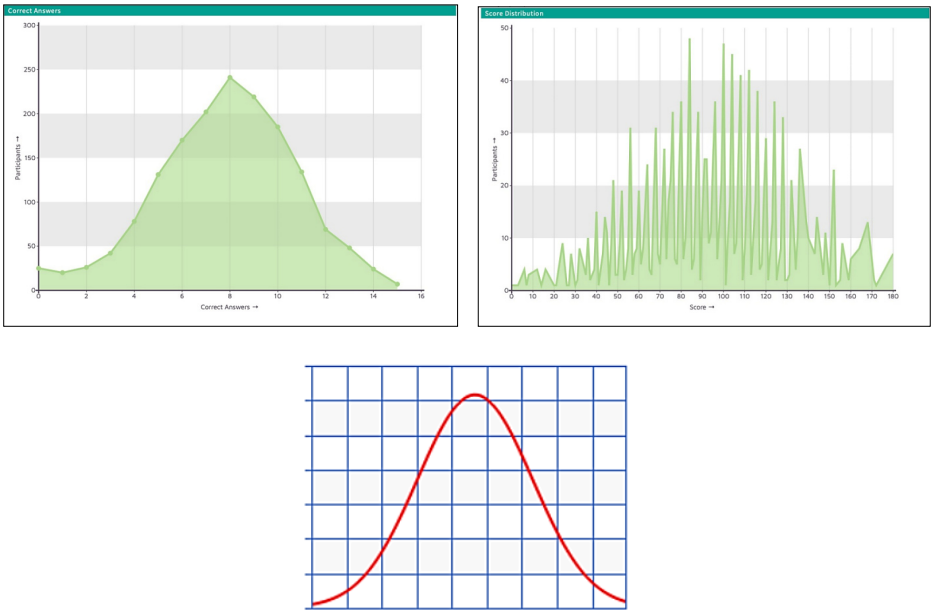

Fig. 1. Number of correct answers and score distribution, compared with a normal distribution.

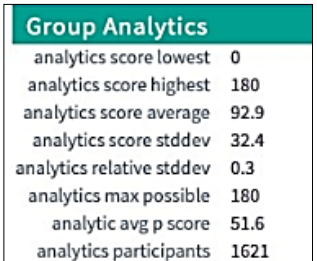

Fig. 2. General analytics.

Table 3

Details per task

Task	P_{all}	R_{it}	%NA	Task	P_{all}	R_{it}	%NA
2017-CA-12	87.4	33.5	2.04	2017-CZ-04c	45.2	55.8	12.16
2017-IS-01	86.4	44.9	3.21	2017-CH-07b	16.5	34.2	11.79
2017-BE-05	81.7	37.5	2.10	2017-KR-07	48.4	57.3	20.37
2017-RU-03	65.7	42.8	6.85	2017-SK-12a	43.1	52.5	18.64
2017-IR-07	41.4	39.1	2.53	2017-UK-04	35.2	40.1	23.09
2017-CA-07	75.9	38.6	15.93	2017-KR-03	15.7	45.5	32.10
2017-PL-02	68.1	46.0	6.05	2017-SI-04	10.1	38.8	23.77
2017-CH-01b	63.8	41.3	23.95				

4.1. *Specific Task Details*

The Cuttle contest system allows us to analyze the results of a task in more details. Fig. 3 shows the plots of the two tasks with the lowest and the highest R_{it}-score. The five values in the graph are the P-values for five different percentiles. So the lines will need to be ascending or at least not-decreasing. The low R_{it}-value in the left graph can be recognized as a bend line, where the highest line in the left graph approximates a straight line indicating a high R_{it}-value. The lower line in this graph is for a younger age group. The graph shows that for the best performing contestants in both age groups the P-values are almost similar; but the differences between age groups for less well performing contestants are much more age dependent.

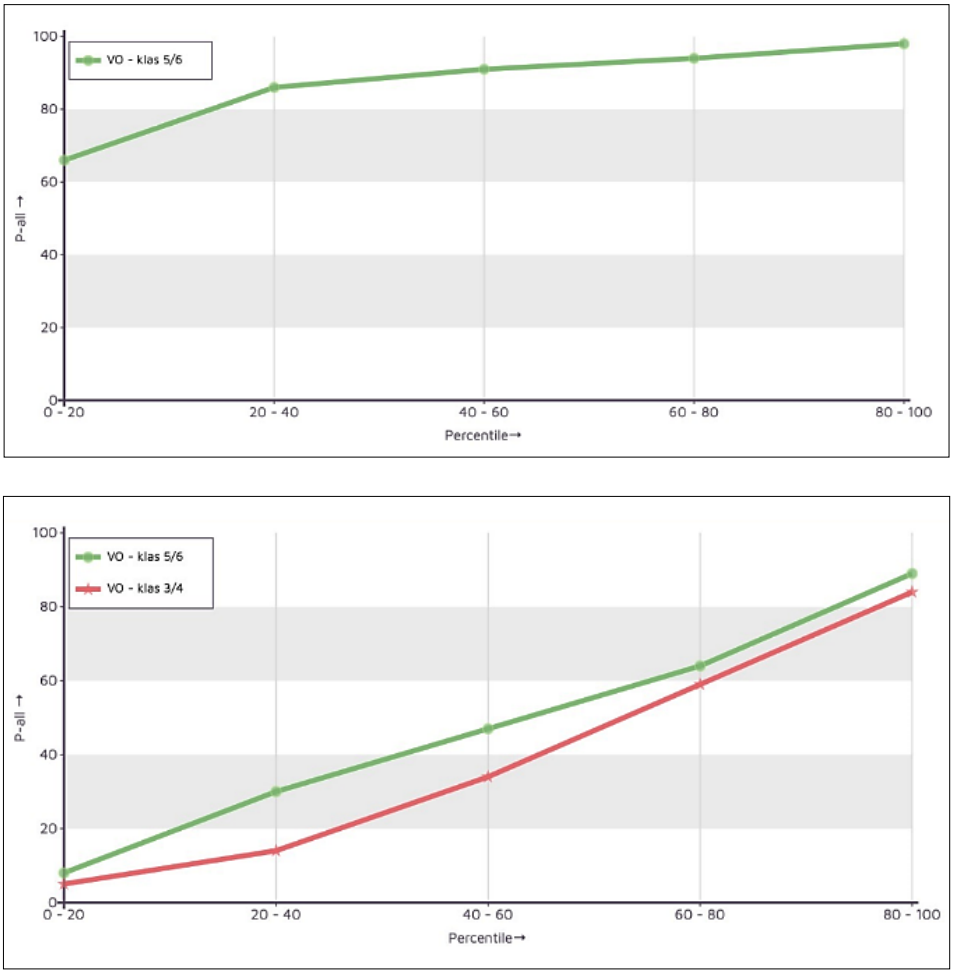

Fig. 3. The results of task 2017-CA-12 and task 2017-KR-07.

4.2. *Categories*

Using the output of the content system, we investigated the results for different categories. We looked at DSR and ALP; we will qualify the result for the one task that combined ALP, ISS and COM as a task on ALP. Though the numbers are small, there seems to be a tendency that DSR-tasks have shown to be a bit less difficult. And the combination of ALP and DSR is harder than the sole categories. This suggests that tasks on ALP require a higher cognitive load and combining both categories increases it even more. Performing an algorithm requires to make more steps in your memory or to use external memory like paper and pencil. That makes the solution process more error-prone. Another way to look at it is that DSR-tasks are more of a static nature while ALP-tasks are more dynamic. According to Leong (2006) increasing or decreasing the number of steps needed to find a solution influences task difficulty. It is interesting to investigate whether the nature of ALP-tasks makes it harder to reduce the number of steps in the solution process.

The P-value, the R_{it} and the percentages of non-answered tasks for categories are shown.

4.3. *Answer Type*

The same approach is used for analyzing the results for the different answer types. Table 5 shows the five different answer types as described in Section 3.2. As was expected, the Open Ended tasks turned out to be the hardest. The Open Ended Text task had almost one third of the participants not answering. This result can be attributed to the much larger search space in these open tasks, increasing task difficulty.

Table 4
Results per category

Categories	n	P_{all}	R_{it}	%NA
ALP/DSR	4	45.5	40.8	13.01
Only ALP	8	50.1	43.9	15.41
Only DSR	3	67.2	44.5	9.75

Table 5
Results per answer type

Answer type	n	Pall	Rit	%NA
MC Text	5	59.8	43.0	10.25
MC Images	2	62.4	45.0	10.37
Open Ended Integer	3	31.6	39.7	13.87
Open Ended Text	1	15.7	45.5	32.10
Interactive	4	62.7	44.6	14.72

4.4. *Task Difficulty*

An open question for us is whether Is it possible to make one scale for difficulty level and age group. In practice we use the assumption that the difficulty level of a task is reduced one step if you offer the task to the next higher age group. This way we can for instance offer the same task as hard for age group IV, as medium for age group V and as easy for age group VI. The results for a task for adjacent age groups can turn out to be really different, due to the computer science concepts in it or the cognitive development of the contestants of a specific age. Table 6 presents the P-values of tasks that were used in several age groups. The average difference of the P-values of age group VI and age group V is 12.8 percent, the difference between age groups VI and IV is 25.2 percent and for the two tasks that were also in the contest for age group III the difference was 42.3 percent.

An interesting research question would be to look for an explanation of the small differences in difficulty level for the one task, for instance 2017-CH-01b, and the large difference for some other tasks like 2017-CZ-04c. Understanding these differences would really help us to predict the difficulty level of a Bebras task for a certain age group.

5. Conclusions

We have tried to answer two questions. It is possible to use the Cuttle system to collect data that can be useful for analyzing task difficulty in Bebras? The new features of Cuttle offered us the chance to investigate the results of a contest in a much more detailed way. We were able to check and confirm that the contest we analyzed had a proper correlation between the individual tasks and the contest as a whole, we could reflect on the actual difficulty level and compare it to the announced difficulty.

And can we formulate questions for future research, based on the findings using Cuttle? We were able to show the relation between answer type and the P-values of the

Table 6
P-values for other age groups

	III	IV	V	VI
2017-IS-01			69.0	86.4
2017-RU-03	22.8	35.7	53.8	65.7
2017-CA-07			61.8	75.9
2017-PL-02		34.9	49.3	68.1
2017-CH-01b		51.7	62.0	63.8
2017-CZ-04c	3.5	6.9	20.5	45.2
2017-CH-07b		4.1	9.2	16.5
2017-KR-07			39.9	48.4
2017-SK-12a			32.6	43.1

tasks. This is in line with earlier results on question difficulty, so answer type is useful as a parameter on task difficulty. We also showed that at least in this contest tasks of category DSR seems to be more easy than ALP tasks while combining these category increases the difficult even further. Repeating this analysis for other contests is needed to check if this reveals a general pattern and if category can be a parameter in predicting task difficulty. We found an average increase of around 13% in P-values for the same task used in the next higher age group. But there are larger differences between tasks and it will be interesting to look into these differences in order to be able to predict the difficulty level for each specific age group. The new tool for analysis can help us in this future research.

References

Ahmed, A., Pollitt, A. (1999). Curriculum Demands and Question Difficulty. Paper presented at IAEA Conference, Slovenia, May.

Bebras website (2019). http://bebras.org/

Beverwedstrijd (2019). http://www.beverwedstrijd.nl/ (in Dutch)

Dagienė, V., Futschek, G. (2008). Bebras international contest on informatics and computer literacy: Criteria for good tasks. In: R.T. Mittermeier and M.M. Syslo (Eds.), *ISSEP 2008, LNCS 5090*. Springer-Verlag Berlin Heidelberg, pp. 19–30.

Dagienė, V., Sentance, S. (2016, October). It's computational thinking! Bebras tasks in the curriculum. In: *International Conference on Informatics in Schools: Situation, Evolution, and Perspectives*. Springer International Publishing, pp. 28–39.

Dagienė, V., Sturupienė, G. (2016). Bebras – a sustainable community building model for the concept based learning of informatics and computational thinking. *Informatics in Education*, 15(1), 25–44.

Dhillon, D. (2003). *Predictive Models of Question Difficulty – A Critical Review of the Literature*. Manchester, AQA Centre for Education Research and Policy

Goldebeld. P. (1992). *Toets en – Itemanalyse Met TIA*. Cito, Arnhem. (in Dutch)

Leong, S.C. (2006). On varying the difficulty of test items. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore.

Lonati, V., Malchiodi, D., Monga, M., Morpurgo, A. (2017). How presentation affects the difficulty of computational thinking tasks: an IRT analysis. In: *Proceedings: 17th Koli Calling Conference on Computing Education Research: Koli Calling 2017: November 16–19, 2017: Koli, Finland*. ACM, 60–69.

Van der Vegt, W. (2013). Predicting the difficulty level of a Bebras task. *Olympiads in Informatics*, 7, 132–139.

Van der Vegt, W. (2018). How hard will this task be? Developments in analyzing and predicting question difficulty in the Bebras Challenge. *Olympiads in Informatics*, 12, 119–132.

W. van der Vegt is teacher's trainer in mathematics and computer science at Windesheim University for Applied Sciences in Zwolle, the Netherlands. He is one of the organizers of the Dutch Olympiad in Informatics and he joined the International Olympiad in Informatics since 1992. He has been a part of the international Bebras community from the start in 2005 and has been a member of the Bebras board, with a specific interest in task development.

E. Schrijvers is chair of the Dutch Foundation of the Informatics Olympiad. Since 1994 he is teamleader of the Netherlands at the International Olympiad in Informatics. He runs Eljakim IT, which develops and maintains Cuttle, the contest system that is used for Bebras in the Netherlands. This system is used in over thirty Bebras and other scientific contests.