Olympiads in Informatics, 2018, Vol. 12, 119–132

DOI: 10.15388/ioi.2018.10

How Hard Will this Task Be? Developments in Analyzing and Predicting Question Difficulty in the Bebras Challenge

Willem van der VEGT
Dutch Olympiad in Informatics
Windesheim University for Applied Sciences
PO Box 10090, 8000 GB Zwolle, The Netherlands
e-mail: w.van.der.vegt@windesheim.nl

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. Question difficulty breaks down in content difficulty, stimulus difficulty and task difficulty. Several instruments are suggested to predict the overall difficulty level, like using a questionnaire or a rubric; these instruments are applied on the data of a recent contest and proved useful. Relative scoring could also turnout helpful. Especially on content difficulty easy applicable solutions are lacking.

Keywords: Bebras contest, question difficulty, taxonomy, cognitive load theory.

1. Introduction

The Bebras Challenge is an annual International Contest on Informatics and Computational Thinking amongst the young (Bebras, 2018). Students from over fifty countries compete in their national contest. The questions used in these contests are chosen from a common task pool, which is composed in the 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. Criteria for good Bebras tasks, using a former system for classification, are 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 contestants have 40 minutes to complete 15 tasks. These can be multiple choice questions, questions where an answer has to be given in the form of an integer or a short string, or interactive questions. The contest runs for a week in five different ages groups; some countries will also have an event for the youngest age group, 6–8 years, but the Dutch contest starts for grade 3; contestants are usually aged 8 year or (much) older. The best performing contestants for every age division are invited at a university for a second round (Beverwedstrijd, 2018).

There are several reasons why it is important to predict the difficulty level of a Bebras task in advance (Van der Vegt, 2013). In a perfect world we would always be able to pretest questions to determine their difficulty and statistical characteristics before using them in a contest (Kibble and Johnson, 2011). But pretesting for a contest you really want to engage all possible students is hard to do. However, knowing the predefined difficulty level of a question is part of the contest. It is possible that contestants take this into account when answering a question.

Lee and Heyworth (2000) state that it is general agreed that students should be able to score higher in a test if the items or exercises are arranged according to their difficulty levels. They are looking for a measure of problem difficulty that can be obtained when a problem is created. They identify four different difficulty factors for algebra problems: the perceived number of difficult steps, the number of steps required to finish the problem, the number of operations in the problem expression and students' degree of familiarity with the question.

Leong (2006) explains the need to control question difficulty in test design in general. Test that contain too many easy or too many hard questions result in skewed mark distributions. And for comparison of tests through the years the distribution of item difficulty should be comparable. 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. But Leong makes another 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. We will stick to his distinction.

In section 2 we will focus on content difficulty, including an enquiry on the possible role of taxonomies for this matter.

Section 3 handles with stimulus difficulty and possible reading problems.

Section 4 is dedicated to task difficulty, using cognitive load theory as theoretical background.

In section 5 we present a number of questionnaires, rubrics and procedures to think about question or item difficulty.

In section 6 we analyze a recent contest in the Netherlands, using some of the tools we found.

In section 7 we will discuss our findings and do some suggestions for future research.

2. Content Difficulty

Bebras is about concept in computer science and computational thinking. Barendsen et al. (2015) show that it is possible to identify concepts on programming in various questions in the Bebras task pool. Izu, Mirolo, Settle, Mannilla, and Stupuriene (2017) describe how the goals of computational thinking are reflected in Bebras tasks. Lonati, Monga, Morpurgo, Malchiodi and Calcagni (2017) categorize a lot of Bebras tasks based on computational thinking skills.

The level of complexity of an assessment task is often determined using a taxonomy. The level of mastery is determined by the use of cognitive skills. Dunham, Yapa and Yu (2015) describe a way to use Bloom's taxonomy (Bloom, Engelbart, Furst, Hill and Krathwohl, 1956) for designing assessments in statistics education with varying difficulty levels. They focus on the depth of the thought process to solve problems, and they make explicit how to align assessment tasks on the taxonomy's scale.

Newman, Kundert, Lane and Bull (1988) concluded that students obtained higher scores for harder multiple choice questions when these problems were arranged in increasing cognitive order, i.e. knowledge, comprehension, application. For medium and easy question no such effect was found. But Kindle and Johnson (2011) observe that the assignment of learning taxonomies to multiple-choice questions has no relation at all to the difficulty of questions. They conclude that the categories in a taxonomy cannot be used to control exam difficulty.

Another issue that arises is the so called push-down effect (Merrill, 1971). A learner will attempt to perform a given response at the lowest possible level. For a novice a task can be highly demanding, while experiences students are able to push down the actual cognitive level on which they need to perform.

Finally we need to discuss the possibility to apply any taxonomy on a set of tasks that is aimed to test for insight in concepts, like in Bebras. In a school situation for many subjects reproduction can be a large part of an summative assessment. Bijsterbosch (2018) concluded for instance that in the Dutch lower level secondary education (the vmbo) over 60 % of all school exams questions on geography test reproductive skills. But Bebras provides a challenge where reproduction should be useless; the whole idea is to provide a set of tasks that do not require any pre-knowledge. This is a condition where the relation between learning objectives and the levels of mastery in a taxonomy is altered in a serious way. Adapting any form of taxonomy to this kind of contest will be needed before it will be possible to apply a taxonomy to the content difficulty of Bebras..

3. Stimulus Difficulty

Lumley, Routinksy, Mendelovits and Ramalingam (2012) created a scheme for describing the difficulty of reading items used in PISA. They compared the perceived and the empirical difficulty. They were able to conclude that their set of ten variables could be

reduced, because five variables explained about 57% of the variability in difficulty in items. and found indications that the variables in Table 1 contribute to item difficulty. Since reading and understanding a question is of course an important part of answering a task, these variables might prove useful, also for Bebras.

Remarkable is that their variable 7, Concreteness of information, which on its own correlated modestly but positively with item difficulty, was also found to be significant in the multiple regression analysis, but with a negative relation to item difficulty. Removing this variable lowered the explanatory power of the data. The authors suggest that if an item becomes more difficult, the degree of abstractness of the information readers need relates negatively to the items difficulty.

Lonati, Malchiodi, Monga and Morpurgo (2017) changed the formulation or presentation of some questions in the 2016-Bebras contest and presented these tasks to a new group of contestants. They report remarkable changes in the results for the altered tasks. On the task Recipe, which is on linked lists, the success rate was very low; in interviewing contestants they discovered that the text was not understood and generally read with no care. So they structured the problem in another way, added two ingredients and pre-filled three out of seven fields in the answer, while creating a new figure. They obtained a higher success rate in their control group and a significant decrease in discrimination.

Leong (2006) describes demands to increase stimulus difficulty. Use relevant technical terms, without elaboration or clarification, in the item; present information in such a way that requires candidates to do some re-organization. Supports that will decrease stimulus difficulty are for instance: Highlight or emphasize terms that require careful comprehension; tailor the resources to the task that candidates have to do.

Table 1

Revised PISA reading item difficulty scheme. Five most explaining variables (Lumley, Routinksy, Mendelovits and Ramalingam, 2012)

3	Competing information	This refers to information in the stimulus and/or in the distractors (if multiple choice) that the reader may mistakenly select, or that the reader may generate, because of its similarity in one or more respects to the target information
5	Relationship between task and required information	The relationship between the question (the whole task, including the multiple-choice options where relevant) and the required information – that is, the kind of answer required to gain credit
7	Concreteness of information	The kind of information that readers must identify to complete a question
8	Familiarity of information needed to answer the question	This variable distinguishes tasks that focus on information inside or outside the text, or the text structure, that is close to the experience and concerns of the reader, from those focusing on what is likely to be remote and unfamiliar
10	Extent to which information from outside the text is required to answer the question	This variable deals with the extent to which the reader needs to draw on world knowledge, experience or personal beliefs and ideas and opinions in order to answer the question

4. Task Difficulty

One of the main concerns in question difficulty is the role of the working memory of a contestant. The working memory load is affected by the inherent nature of the material, the intrinsic cognitive load and the manner in which the material is presented. According to the cognitive load theory the limitations of the working memory are rarely taken into account in conventional instruction and assessment (Kirschner, 2002).

Conventional instructions tend to impose an extraneous cognitive load on the working memory, whereas learning something requires shifting for extraneous to germane cognitive load. Germane cognitive load is the effort that contributes to the construction of schemas. Schemata categorize information elements according to how they will be used. They can also reduce working memory load, since a schema can be treated as a single element. So schema construction aids the storage and organization of information in long-term memory and reduces working load memory.

In computer science education this process of schema formation has at least two effects: by building ever more complex schema by assimilating portions of lower-level schemas skills are developed, and once a particular skill is acquired, automatic processing can bypass working memory (Shaffer, Doube & Touvinen, 2003). Indications of working memory failures include: incomplete recall, failing to follow instructions, place-keeping errors and task abandonment (Shilbi & West, 2018).

Most of the available research on cognitive load is focused on education and instruction. Elliot, Kurz, Beddow and Frey (2009) apply cognitive load theory to test design. They present guidelines for testing; some for Bebras relevant suggestions are placed in Table 2.

Leong (2006) summarizes ways to decrease the task difficulty: decrease the number of steps in executing task; break up the task into a few sub-questions, order the steps such that they provide the scaffolding for subsequent steps. Increasing the task difficulty can be done by raising the number of steps in executing task, or present a task in which candidates need to devise steps to execute the task without cues and leaders.

Table 2

Recommendations for handling cognitive load in test design

5.	Use bold for vocabulary words. Use red circles, arrows and highlighting for important elements of visuals
6.	Integrate explanatory text close to related visuals on pages and screens
9.	Text economy; all included visuals are necessary
10.	Don't add words to self-explanatory visuals
13.	Train test-takers in the test-delivery system prior to the test date

5. Instruments

Several tools have been developed to help test designers to predict the difficulty level of a question. In this section we will discuss two questionnaires, a rubric and a procedure. Since these are compositions with many parts, we will also try to analyze the ratio between content difficulty, stimulus difficulty and task difficulty.

5.1. *Questionnaires*

Earlier (Van der Vegt, 2013) we proposed a questionnaire for understanding and predicting the difficulty of a specific task. We focused on the question answering process, distinguishing reading, understanding, searching a mental representation, interpreting and composing, and the size of the problem. On the latter it is possible to give numbers, but the issues on the question answering process give more qualitative data. And weighting these data is a hard task in itself.

The questions I.a and I.b (Table 3) are about stimulus difficulty, I.c, I.d and I.e mainly about content difficulty and all questions II are handling task difficulty. This gives a 30/20/50 distribution on different kinds of question difficulty.

Vora, Jain, Mehta and Sankhe (2016) developed a method to assign weightage to question difficulty. There instrument is shown in Table 4. The level of IQ is a subjective measure, the more the question makes sense, and the more it is related to the test subject, the more weightage is given. This is a measure for content difficulty. Length of question and pattern are merely on stimulus difficulty and question type is on task difficulty. So their questionnaire has a 25/50/25 distribution on different kinds of question difficulty.

The difficulty fraction is by definition the sum of the weights, minus the minimum total weight, divided by the difference of maximum and minimum weight, and if this

Table 3
Questionnaire for difficulty level estimation (Q1)

I.		The question answering process
	a.	Which problems will there be in reading the question?
	b.	Which problems will there be in understanding the question?
	c.	Which problems can arise in searching the mental representation of the text?
	d.	Which problems can arise when interpreting the answer?
	e.	Which problems can arise when composing the answer?
II.		The size of the problem
	a.	What is the number of elements in the question?
	b.	What is the number of transformations for an element in the question?
	c.	What is the number of constraints in the question?
	d.	How do you rate the solution density of the problem?
	e.	Will it be possible to solve the problem, using only your working memory?

Table 4
Weightage assignment (Vora, Jain, Mehta and Sankhe, 2016) (Q2)

Parameter		Weight range
Level of IQ (sense)		2–10
Length of question		2–10
Pattern		
a.	Repetition of keyword	2–8
b.	Image	0–2
Type of question		
a.	True/false type	2
b.	Simple MCQ	4
c.	Calculated MCQ	6
d.	Check Box (Multiple correct answers)	8
e.	Text Box	10

fraction is below 0.1667 the question is defined as easy. A question with a difficulty fraction above 0.5 is considered hard. A medium question is neither easy nor hard. So a task that scores 31 as a total weight will have a difficulty fraction of (31 – 8) / (40 – 8) = 0.72 and such a task would be considered as a hard question. A task with a score of 8–13 points will be easy, any task with between 14 and 24 will be medium.

5.2. *Rubric*

At the Bebras Workshop 2018 the Italian team presented their work on a rubric, designed to make decisions on expected difficulty level (Bellettini, Lonati, Malchiodi, Monga and Morpurgo, 2018) . For ten different aspects of the task they define three possible difficulty levels, and they distinguish between easy, medium and high difficulty. Their ten aspects are stated in Table 5.

For cognitive effort they specify an easy task as one that requires to understand a concept or a procedure and to perform a straightforward activity. A medium task is one

Table 5
Rubric lines (Bellettini, Lonati, Malchiodi, Monga and Morpurgo, 2018) (R)

1	Text and sentence length
2	Familiarity of terms, notations, objects, and concepts needed to understand the task
3	Consistency of terms and notations
4	Other elements beside the text (pictures, diagrams, examples, etc.)
5	Constraints, combinations, steps needed
6	Relationships among objects to take into account
7	Cognitive effort
8	Use of notes or other supported material
9	Solution space
10	Solution check

that requires to analyze a context, a procedure or some properties. And a hard task requires to evaluate a setting, a system, a procedure, possibly comparing different hypotheses or strategies. Or it requires a creative effort to invent or build something.

In this rubric items 1, 2, 3 and 4 are about stimulus difficulty, 5 and 6 on content difficulty, and the other four mainly on task difficulty. So this rubric has a 20/40/40 distribution on question difficulty.

5.3. *Procedure*

Current practice is that the author of a Bebras task suggests the appropriate age category and difficulty level. At the international workshop, in selecting for a national contest and in translating the task adjustments may be made. In some countries, like in the Netherlands, the same task will be used for several age groups, assuming that the task prove to be more easy if the contestants are older.

Holmes and Read (2018) describe that it is very hard to make absolute judgements on question difficulty. They use a technique where a number of experts independent review many pairs of items and decide each time which item is more difficult to answer. This comparative judgement can be used to capture a group consensus well, and to avoid individual biases. This approach can be useful also for Bebras tasks; we toke a set of six different tasks and asked a group of colleagues to order them from easy to hard. We scored the individual results from 1 (easy) to hard (6) and added the individual scores for each task. The total scores were a perfect match with the relative difficulty level. Kindle and Johnson (2011) report a similar practice: Each of nine faculty members was misjudging the difficulty level of some of the tasks in an exam, but the average score proved to be much better.

Bramley and Wilson (2016) asked a group of experts to estimate the mean marks for every question in a specific test, by using information on the results of previous test and looking for nearly identical questions. Statistical information can be used by the experts to guide their judgements. Developing a benchmark of used questions with known difficulty levels could be helpful in predicting the difficulty of new developed tasks.

6. A Recent Contest Analyzed

In an analysis after the contest we can define the actual difficulty level of each task as the p-value: item difficulty is simply the fraction of contestants taking the test who answered the item correctly. The larger the fraction getting an item right, the easier the question.

For the highest age groups of the first round of Bebras in the Netherlands in 2017 graphs of the percentages of good answers are presented in Fig. 1, Fig. 2 and Fig. 3. And though some sections of the questions are really well predicted, like the hard questions for age group IV and a few of the easy questions in all of these contests, a lot of questions

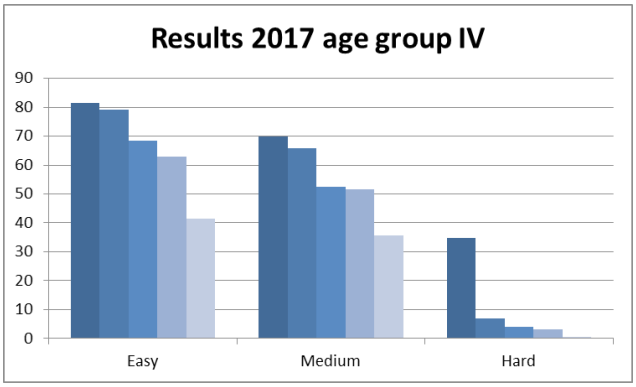

Fig. 1. Results of the 2017 contest for age group IV.

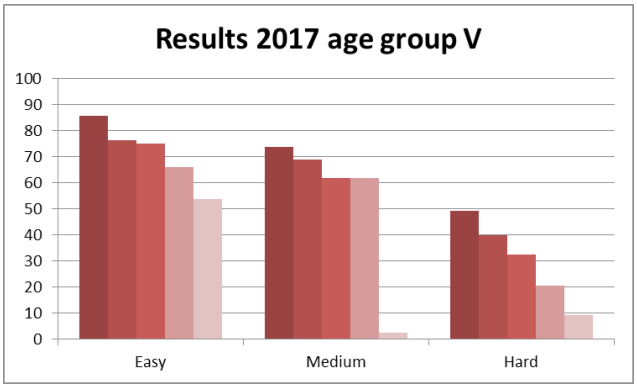

Fig. 2. Results of the 2017 contest for age group V.

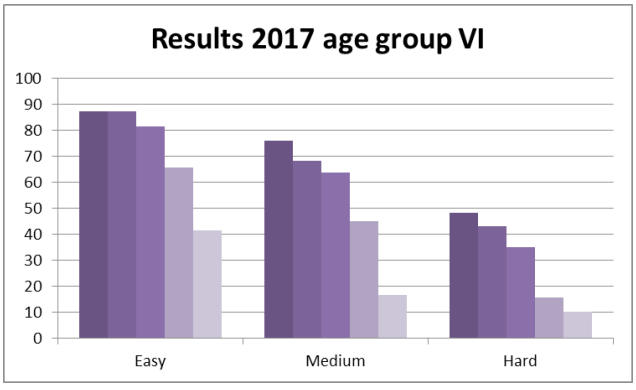

Fig. 3. Results of the 2017 contest for age group VI.

were under- or overestimated. An easy measure for the quality of our predictions is the percentage of misplaced tasks. In the contest presented in Fig. 1 this was 27%, in Fig. 2 it was 40% and in Fig. 3 it was 47%. Table 6 gives an overview of this measure.

The mean scores for the questions in a difficulty group are presented in Table 7. In all three contests the results are as expected, though the hard questions for age group IV turned out to be extremely difficult. The hardest question for age group V turned out to be a task we qualified as medium, and in age group VI the most difficult tasks of the

Table 6
Percentage of misplaced tasks in the first round of 2017

Year	Age division	Harder than predicted	Easier than predicted	Percentage misplaced
2017	IV (12–14)	2	2	27%
	V (14–16)	3	3	40%
	VI (16–18)	3	4	47%

Table 7
Results of the difficulty groups in the first round of 2017

Year	Age division	Easy questions	Medium questions	Hard questions
2017	IV (12–14)	66.72	55.13	9.93
	V (14–16)	71.39	53.86	30.29
	VI (16–18)	72.69	53.92	30.48

Table 8
Task analysis of age group VI in Dutch Bebras 2017

Task-ID	Assigned difficulty level	Success	Q1	Q2	R
2017-CA-12	Easy	87.42	0.40	0.22	0.30
2017-IS-01	Easy	86.37	0.40	0.28	0.35
2017-BE-05	Easy	81.62	0.50	0.31	0.40
2017-RU-03	Easy	65.70	0.55	0.38	0.55
2017-IR-07	Easy	41.39	0.70	0.47	0.60
2017-CA-07	Medium	75.88	0.60	0.53	0.55
2017-PL-02	Medium	68.17	0.65	0.59	0.60
2017-CH-01b	Medium	63.73	0.75	0.59	0.60
2017-CZ-04c	Medium	45.22	0.70	0.66	0.70
2017-CH-07b	Medium	16.59	0.85	0.63	0.80
2017-KR-07	Hard	48.37	0.75	0.66	0.70
2017-SK-12a	Hard	43.06	0.85	0.66	0.70
2017-UK-04	Hard	35.16	0.90	0.81	0.80
2017-KR-03	Hard	15.67	0.85	0.78	0.75
2017-SI-04	Hard	10.12	0.90	0.63	0.70

section easy and the section medium were harder than predicted. In section 6.1 we will analyze some of these tasks for age group VI more in detail.

Izu, Mirolo, Settle, Mannilla, and Stupurienė (2017) make a similar analysis for the 2014 and 2015 contest in five countries. They use the rank match to see how well the difficulty level was predicted; the outcomes are between 40% and 72%. This corresponds to 100 minus the percentage misplaced, so our Dutch values for rank match are 53%, 60% and 73%.

We did an analysis for all 15 questions of the highest age group in the Dutch Bebras 2017, using the three instruments in section 5. Q1 refers to the questionnaire of Table 3; all questions are scored as 0 for easy, 1 for medium or 2 for hard, so the total is on a scale from 0 to 20. Reported in the table is the fraction of the maximum score. Q2 stands for the questionnaire of Table 4; the result reported is the difficulty fraction. And R is the Italian rubric, presented in Table 5, also scored using 0, 1 or 2 per item and presented as a fraction.

At first sight all three instruments can be used to order the tasks from easy to hard. For each of the three scales we performed a linear regression, the results of which are seen in Fig. 4. We also calculated the correlation coefficient. In all three cases this coef-

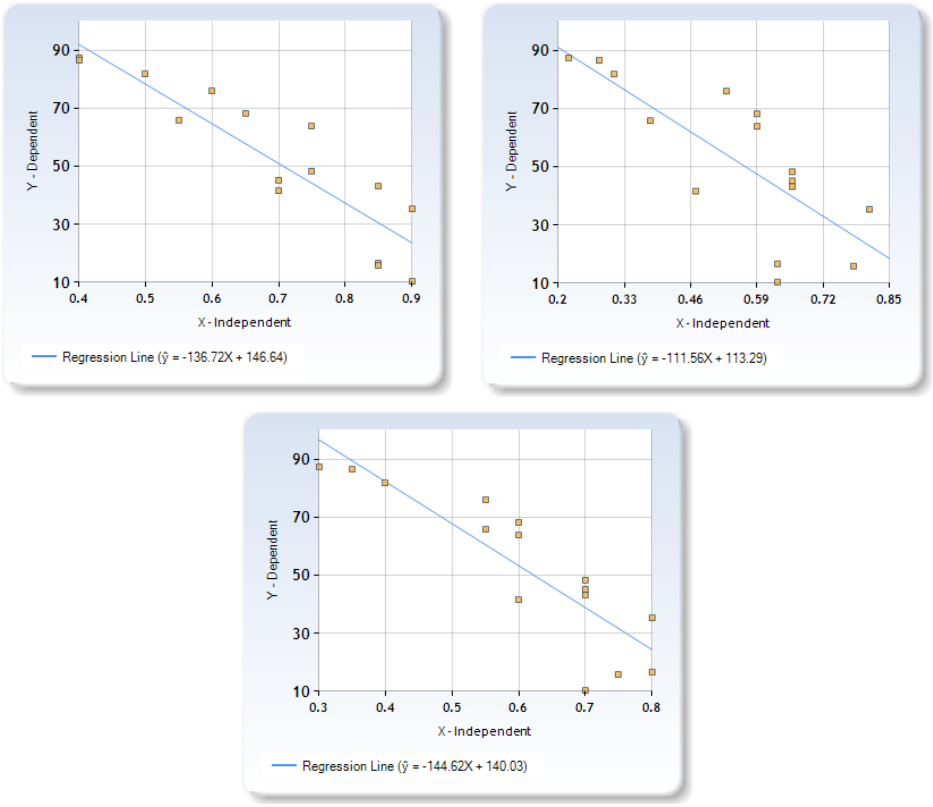

Fig.4. Lineair regression of Q1 (left), Q2 (right) and R (bottom) and success-rate for Bebras 2017.

ficient was almost -1. Q1 had a correlation coefficient of -0.90, Q2 one of -0.77 and the result for R was -0.87. So the rule of thumb can be that the higher the difficulty fraction, calculated with either of these instruments, will be, the harder the question will prove in practice.

Questionnaire 2 proved the hardest to use for a scorer. Level of IQ (sense) asks for a number on content difficulty; we used the assigned difficulty level, with a 3 for an easy task, a 6 for a medium one and a 9 for a hard task. But these judgements were of course already about more than content difficulty. So this measure has a systematic flaw. For questionnaire 1 and the rubric a lot of close calls had to be answered; the rubric has specifications when to assign a specific score for a task, but these specifications are not always decisive enough. Questionnaire 1 is lacking these specifications at all, so scoring is quite intuitive, but could easily be biased, since the actual results of the contest were already available.

One of the main factors in the actual use of this kind of instruments will be the time a scorer needs to answer all questions. The items should be well-defined and the boundaries between possible scores need to be clear; otherwise these instruments will never be used in designing an actual contest because no one has the time to fill in the forms.

7. Discussion

Application of several tools, developed to predict the difficulty level of a (Bebras) task, can help the contest designer to create a fair, balanced contest. All three investigated instruments can be used for this goal. In further research one could look to the best balance for the components and weights on content, stimulus and task difficulty.

Content difficulty is the most unclear item in predicting difficulty. More research is needed on the use of taxonomies, especially for questions that do not use any pre-knowledge, or other systematic approaches to identify content difficulty.

The use of procedures for relative scoring seems promising. Combining individual judgements on question difficulty can improve the overall decision. Integrating the use of questionnaires and relative scoring based on the output by several scorers will be a valuable condition for a more systematic preparation of this part of the contest.

Testing and the need to predict task or question difficulty go beyond the boundaries of Bebras. A lot of recent research on cognitive psychology shows that stimulus and task difficulty play an important role in the performance of contestants. Instruments used to predict question difficulty should include these insights.

References

Barendsen, E., Manilla, L., Demo, B., Grgurina, N., Izu, C., Mirolo, C., Sentence, S., Settle, A., Stupurienė, G. (2015). Concepts in K–9 computer science education. In: Dagienė, V. (Ed.), *ITICSE '15 : Proceedings of*

the 2015 ACM Conference on Innovation and Technology in Computer Science Education Conference, 2015 Vilnius, Lithuania – July 04–08. 85–116.

Bebras website (2018). `http://bebras.org/`

Belletini, C., Lonati, V., Malchiodi, D., Monga, M., Morpurgo, A. (2018). A rubric to help with Bebras tasks. Presented at the Bebras Workshop 2018, Protaras, Cyprus.

Beverwedstrijd (2018). (in Dutch) `http://www.beverwedstrijd.nl/`

Bijsterbosch, H.D. (2018). Professional Development of Geography Teachers with Regard to Summative Assessment Practices. University of Utrecht.

Bloom, B.S., Engelbart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R. (1956). Taxonomy of Educational Objects: The Classification of Educational Goals, Handbook I: Cognitive Domain. New YorkL David McKay Co Inc.

Bramley, T., Wilson, F. (2016). Maintaining test standards by expert judgement of item difficulty. In: Research Matters, Issue 21.

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, 19–30.

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.

Dunham, B., Yapa, G., Yu, E. (2015). Calibrating the difficulty of an Assessment Tool: The Blooming of a Statistics Examination. *Journal of Statistics Education*, 23(3).

Elliot, S.N., Kurz, A., Beddow, P., Frey, J. (2009). *Cognitive Load Theory: Instruction-based Research with Applications for Designing Tests*. Presented at the national Association of School Psychologists 39; annual convention, Boston, MA.

Holmes, S., Rhead, S. (2017). A level and AS mathematics: an evaluation of the expected item difficulty. Ofqual/18/6344.

Izu, C., Mirolo, C., Settle, A., Mannilla, L., Stupurienė, G. (2017). Exploring Bebras task content and performance: A multinational study. *Informatics in Education*, 16, 39–59.

Kibble, J.D., Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in Physiology Education*, 35, 396–401.

Kirschner, P.A. (2002). Cognitive load theory: implications of cognitive load theory on the design of learning. *Learning and Instruction*, 12, 1–10.

Lee, F.-H., Heyworth, R. (2000). Problem complexity: a measure of problem difficulty in algebra by using computer. *Educational Journal*, 28(1), 85-107

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.

Lonati, V., Monga, M., Morpurgo, A., Malchiodi, D., Calcagni, A. (2017). Promoting computational thinking skills: would you use this Bebras task?, In: *Proceedings of the International Conference on Informatics in Schools: Situation, Evolution and Perspectives (ISSEP2017)*. Helsinki, Finland

Lumley, T., Routitsky, A. Mendelovits, J., Ramalingam, A. (2012). A framework for predicting item difficulty in reading tests. ACEReSearch.

Merrill, M.D. (1971). Necessary psychological conditions for defining instructional outcomes. *Educational Technology*, 11(8), 34–39.

Newman, D.L., Kundert, D.K., Laner, D.S., Bull, K.S. (1988). Effect on varying item order on multiple-choice questions: Importance of statistical and cognitive difficulty. *Applied Measurement in educaion*, I(1), 89–97.

Schaffer, D., Doube, W., Tuovinen, J. (2003). Applying cognitive load theory to computer science education. In: M. Petre and D. Budgen (Eds.) *Proc. Joint Conf. EASE & PPIG*. 333–346.

Shibli, D., West, R. (2018). Cognitive load theory and its application in the classroom. *Impact, Journal of the Chartered College of Teaching*.

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

Vora, K., Jain, S., Mehta, P., Sankhe, S. (2016). Predictive analysis: Assigning weightage and difficulty level of question using data mining. *International Journal of Computer Applications*, 138(9), 31–33.

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 is a part of the international Bebras community from the start in 2005 and is nowadays a member of the Bebras board, with a specific interest in task development.