Olympiads in Informatics, 2013, Vol. 7, 132–139

Predicting the Difficulty Level of a *Bebras* Task

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. In the *Bebras* contest questions are marked as easy, medium or hard. Sometimes contestants perform other than expected. Since pretesting is impossible, we want to find out what kind of tasks were misplaced and develop some guidelines for predicting the difficulty level.

Key words: *Bebras* contest, task design, question difficulty.

1. Introduction

The *Bebras* contest is an international contest on informatics and computer fluency amongst the young (Bebras, 2013). Students from over twenty countries compete in their national contest. The questions used in these contests are chosen from an international task pool. No prior knowledge is required. The contest is about computer science, algorithms, structures, information processing and applications. Criteria for good *Bebras* tasks are formulated by Dagiene and Futchek (2008).

Contestants compete in their own age division. In the Netherlands we use a division for 12–14 years, the Cadets, for 14–16 years, the Juniors and for 16 years and up, the Seniors.[1] In the Netherlands the contest consists of 15 short questions, divided in three difficulty levels: easy, medium and hard. A right answer for a question is awarded with 6, 9 or 12 points (depending on the difficulty level); for a wrong answer 2, 3 or 4 points are subtracted. Skipping the question will not alter your score. Every contestant starts with 45 points, so no contestant will get a negative score; answering all questions correct gives the top score of 180. Contestants have 40 minutes to complete these 15 tasks. The contest runs for a week; the best performing contestants for every age division are invited at a university for a second round (Beverwedstrijd, 2013). Like in many countries, organizing the *Bebras* contest is done by the organizers of the Dutch Olympiad in Informatics.

One of the problems we face is that it is hard to predict the difficulty level of a task. In most cases we use the indications available in the task pool. But when we select questions that are best suitable for the Dutch contest, we sometimes end up with a lot of questions of the same difficulty level, so we need to move a few tasks to a different difficulty category. On other occasions we use questions that were designed for other age groups to complete a contest.

[1] In other countries the contest is also for age groups 10–12, the Benjamins, and sometimes even for 8–10, but our national organization is only connected with the schools of secondary education.

In this paper we will analyze the results of six different contests. A simple measure is used for comparing our predictions of the difficulty level. We will also discuss some research from other areas about predicting question difficulty. Finally we will develop a questionnaire that can be helpful for designing future contests.

2. The Difficulty Level of a Task

The easiest way to look at the difficulty of a task is to look at the result. A task that is solved by a majority of the contestants is definitely easier than a task that is solved by a small percentage. Two of the graphs of the percentages of good answers are presented in Figs. 1 and 2. With results like the one in Fig. 1 it is obvious that we made a couple of bad decisions regarding the difficulty level. In an ideal situation the five tasks in which contestants performed best should have been placed in the easy category, the five tasks in which they performed worst in the hard category. But looking this way, amongst the intended easy questions was at least one that should have been characterized as hard; four questions should have been moved to a higher difficulty level, five others to an easier one. For the Seniors we predicted rather well. Figure 2 shows that only one of the questions was designed to be medium, but turned out to be hard.

An easy measure for the quality of our predictions is the percentage of misplaced tasks. In the contest presented in Fig. 1 this was 60%, in Fig. 2 it was 13%. Table 1 gives an overview of this measure, applied to the six contests we organized in 2011 and 2012.

Are these proportional success rates in Figs. 1 and 2 reliable? Is it allowed to use these rates as a proper indication of the difficulty of a task for an age division? Well, let's state that the answer for a specific question is the result of a chance experiment. The chance for success for every experiment is an unknown value q. When we take a test sample of n contestants, the success rate p is an obvious estimator for q. The 95%-confidence interval has a maximum radius of $0.98/n$ for $p = 0.5$. With for example 1000 contestants this is less than 0.001 or 0.1%. So it is reasonable to say that the success rate p for a question can be used to describe the difficulty level q of that question.

How well did we predict the results of groups of questions? We may have misplaced a couple of question, but was the contest as a whole the way we intended, with a group of

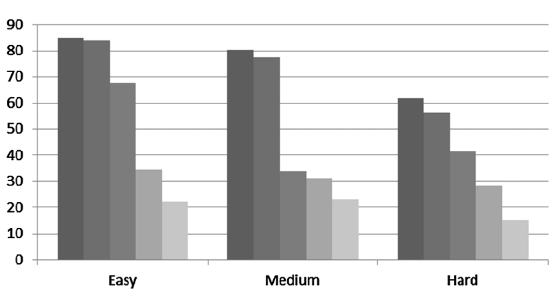

Fig. 1. Results of the 2011 contest for the Cadets.

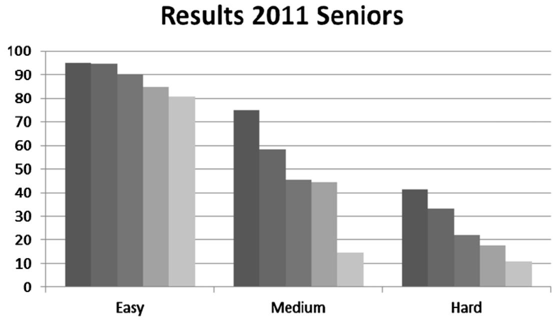

Fig. 2. Results of the 2011 contest for the Seniors.

Table 1
Percentage of misplaced tasks in 6 contests

Year	Age division	Harder than predicted	Easier than predicted	Percentage misplaced
2011	Cadets	4	5	60%
	Juniors	4	5	60%
	Seniors	1	1	13%
2012	Cadets	3	3	40%
	Juniors	1	1	13%
	Seniors	3	3	40%

Table 2

Year	Age division	Easy questions	Medium questions	Hard questions
2011	Cadets	58.64	49.18	40.74
	Juniors	57.22	67.99	30.12
	Seniors	89.13	47.61	25.11
2012	Cadets	65.14	51.48	37.08
	Juniors	72.72	58.45	29.44
	Seniors	75.09	64.93	37.17

easy questions, a medium and a hard one? The mean scores for the questions in a difficulty group are presented in Table 2. In five of the six contests the results are as expected. Only the contest for the Juniors in 2011 shows a bad result for this mean scores, where the mean score for the easy questions is much below the score for the medium intended questions. This was one of the contests where 60 % of the questions turned out to be misplaced. In this cases swapping the best answered medium question and the worst answered easy question would have made a difference of 12 % in the mean scores for the difficulty group.

3. Question Difficulty

A technique often used in test design is pretesting. A set of possible questions will be answered by a test panel, in order to compare the results, to judge about the suitability of specific questions and to get a proper indication of the difficulty level. Due to the way the international task pool for *Bebras* tasks is constructed, pretesting is not an option.

A lot of exams have a stage between doing the test and publishing he results. This gives the organizers the possibility to skip questions and to define the cut between those who passed the exam and those who did not. But the *Bebras* contest is not an exam, it is a competition. Not the score itself, but the ranking of the scores is important. 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, so it would be strange to change the difficulty level after the contest took place.

These considerations leave us no other option than to think ahead, and to try to estimate the difficulty level of a question in advance. Of course, it is possible to play with the parameters of a question to manipulate the difficulty level somewhat.

But this manipulation of question difficulty is a complex matter. Many question-setters try to achieve this merely on their intuition, of course based on experience with similar questions. Though *Bebras* contest designers are often right in describing and predicting the difficulty level of a task, this endeavour is an inexact science (Dhillon, 2003). Most of the research on question difficulty is done within specific subject domains. Dhillon however presents an overview of research results that are useful also for task settings like the *Bebras* contest.

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. Process difficulty is about the difficulty of the cognitive operations and the degree in which they use cognitive resources. Question difficulty is connected with the linguistic and structural properties of a question. Dhillon uses these three kinds of difficulties, and she makes a further distinction between 'intrinsic' question difficulty, which is content-bound, and 'surface' question difficulty, which is format-bound. In our discussion we will stick to the intrinsic question difficulty.

Sternberg suggests that intelligence can be understood in terms of the information-processing components underlying complex reasoning and problem-solving tasks such as analogies and syllogisms. Sternberg used information-processing and mathematical modeling to decompose cognitive task performance into its elementary components and strategies (MIT Encyclopedia of Cognitive Science, 2013). He analyzes analogy item types, with the form 'A is to B as C is to ?'. A series of six roughly sequential, sometimes cyclical, cognitive steps have to be taken to find a solution of an analogy problem. The different stages are encoding ('How to think about A and B'), inference ('What is the relation between A and B?'), mapping ('What is the relation between A and C?'), application ('How to transform C according to the discovered pattern?'), justification ('Is this really the intended answer?') and response ('This is what I have found.'). Sternberg proposed that the difficulty level of a question as a whole is comprised of the sum of the difficulties of each of the individual components, multiplied by the number of times each component was executed (Dhillon, 2003). Several researchers are able to show that this approach can be useful. Increasing the number of elements involved in a question increases the time needed for solving the problem. But the error rates behaved slightly different. The number of transformations per element had the greatest impact on item difficulty (Dhillon, 2003). The assumption of linearity in analogy item-difficulty breaks down at higher levels of item complexity, due to the limits of the short-term memory capacity. Miller (1956) points out that the span of immediate memory imposes severe limitations on the amount of information that we are able to receive, process and remember. If it is impossible to solve a task using only working memory, the solution process will take much more time and will be more error-prone. Sternberg's method still offers a clear operational guide to the construction of analogy problems and related question types; the decomposition of a question in elements and operations on these elements gives a first indication of the expected difficulty level.

Ahmed and Pollitt (1999) describe a model of the question answering process for students that have to respond to a reading comprehension item. Typical stages in this answering process are reading, understanding, searching the mental representation of the text, interpreting and composing the answer. Then they observe that understanding a subject is simply a macro-level version of comprehending a language. So question answering can similarly be broken down in corresponding processes. Awareness of the processing stages will not only facilitate the identification and manipulation of legitimate sources of difficulty, but also the identification and elimination of illegitimate sources of difficulty (Dhillon, 2003). The question-setter can better foresee hurdles at which contestants will fall through no fault of their own.

Katz *et al.* (2002) investigate the predictive validity of various features of generating examples test items, algebra problems that pose constraints and ask for example solutions. Students will use a generate-and-test method to find proper answers. The difficulty level of a question can be raised by increasing the amount of testing the student must do. Both the cognitive load and the potential for error would be increased, either by increasing the number of constraints or by reducing the solution density of the item (Dhillon, 2003).

4. Two Sample Problems

We took two typical *Bebras* tasks, to analyze if we can connect the results with the theory on question difficulty.

All four beavers have a hat but something went wrong (Fig. 3).

- *All four beavers have the wrong colour hat at the moment;*
- *With the hats arranged correctly, none of the beavers have the same colour hat as the colour of the shirt;*

To which beaver belongs the green (G) hat?

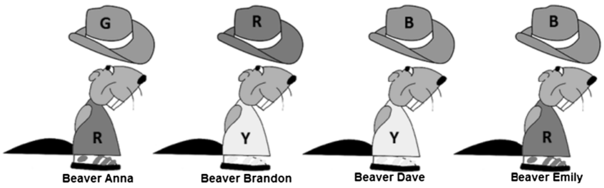

Fig. 3. Task Wrong Hat – *Bebras* Task Pool 2011.

The task Wrong Hat in Fig. 3 was used in the 2011-contest for Cadets. We changed the proposed difficulty level from medium to easy. The success rate was 83.86%. It was a multiple choice question, and in order to find the right solution a contestant has to apply two constraints on all possible answers. One could argue that the solution space has 24 different hat distributions, but application of the first constraint reduces the actual solution space to two possible answers. The solution process fits without problems in the working memory of a contestant. The story is short and easy to understand. So it makes sense to identify this task as an easy question.

The task Three Elephants in Fig. 4 turned out to be the hardest problem that we used in the 2011-contest, with a success rate of 5.60% for Juniors and of 10.95% for Seniors. The task was interactive, but only the used programming buttons were shown, not the resulting state. This task requires close reading and solving it will take a lot of different steps. The position of every elephant has to be remembered, all possible states have to be identified and checked. It is hard to do without additional utilities, like pencil and paper, because it goes beyond the boundaries of the working memory. Experts will recognize the principle behind the task and they will use a Gray code or the analogy with the problem of traversing the vertices of a cube. But since the *Bebras* is a contest that requires no previous knowledge, we have to assume that contestants are not familiar with these problems. The difficulty level of this task could be reduced in several ways. One could of course skip one of the elephants. This would reduce the search space and the number of the consecutive steps, and this would reduce the difficulty level in a serious way. It is also possible to present the problem as a multiple choice question, reducing the number of possible answers and changing the nature of the question. In that case the question is about checking the possible solutions with the given constraints, rather than producing a solution itself. A third way to change the nature of the question is by giving some visual aid in the presentation on the screen, for instance by showing the state after each programmed step. This would make it a lot easier for a contestant to keep track and it would almost certainly reduce the unwanted errors in solving this problem.

> *A circus has an act where three elephants perform on stage. Only one elephant at a time can enter the stage or leave the stage. The director wants to show all combinations of one or more elephants on stage exactly one time.*
>
> *You get six buttons to produce a program for the elephant show. You can use a button more than once; just drag and drop a button to the program list.*

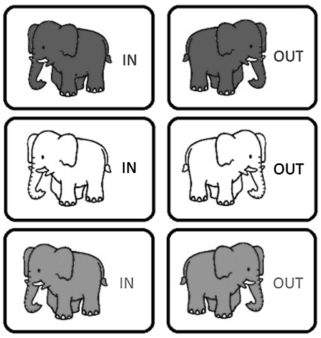

Fig. 4. Problem Three Elephants – *Bebras* Task Pool 2011.

5. Evaluation

Estimating the difficulty level of a *Bebras* task is done merely by intuition. Intuition is "the ability to acquire knowledge without interference and/or the use of reason"[2]. In some cases the intuition of tasks designers failed; tasks turned out much easier or much harder than expected. Results from research can be used to sharpen the intuition of task designers. We propose the use of a questionnaire that can be used to discover probable causes of difficulty. The answers can be used to compare questions and to discuss the estimation of the difficulty level of the question.

Table 3

Questionnaire for difficulty level estimation

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?

Next year we intend to use this questionnaire and to gather data, to investigate whether it is possible to reach a better prediction of difficulty level for *Bebras* tasks. Estimating the difficulty level will not turn into an exact science. But the use of the result of previous result should help us to improve the outcome of our intuition.

[2] Oxford English Dictionary.

References

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

Bebras website (2013).
`http://bebras.org/`.

Beverwedstrijd (2013).
`http://www.beverwedstrijd.nl/` (in Dutch only).

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

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

Katz, I.R., Lipps, A.W., Trafton, J.G. (2002). *Factors Affecting Difficulty in the Generating Examples Item Type*. GRE Board Professional Report No. 97-18P. Princeton, NJ, Educational Testing Service.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychology Review*, 63, 81–97.

MIT Encyclopedia of Cognitive Science (2013). *Intelligence*.
`http://ai.ato.ms/MITECS/Entry/sternberg.html/`.

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 was involved in the IOI-workshops on tasks in Dagstuhl (2006, 2010) and Enschede (2008). He also is one of the task designers for the *Bebras* contest.