How hard will this task be?

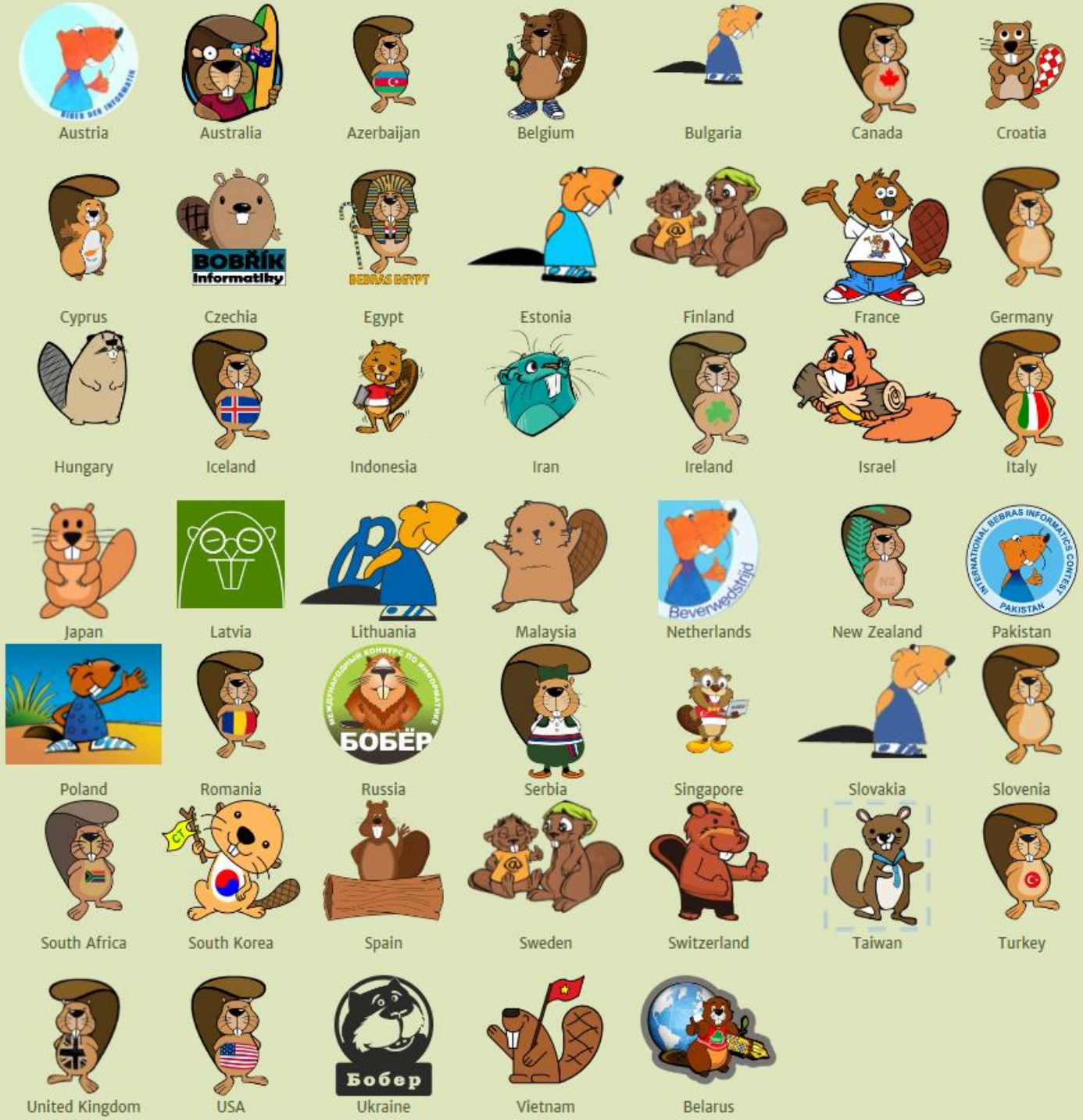
Developments in Analyzing and Predicting Question Difficulty in the Bebras Challenge

Willem van der Vegt

Bebras

International Challenge on Informatics and Computational Thinking

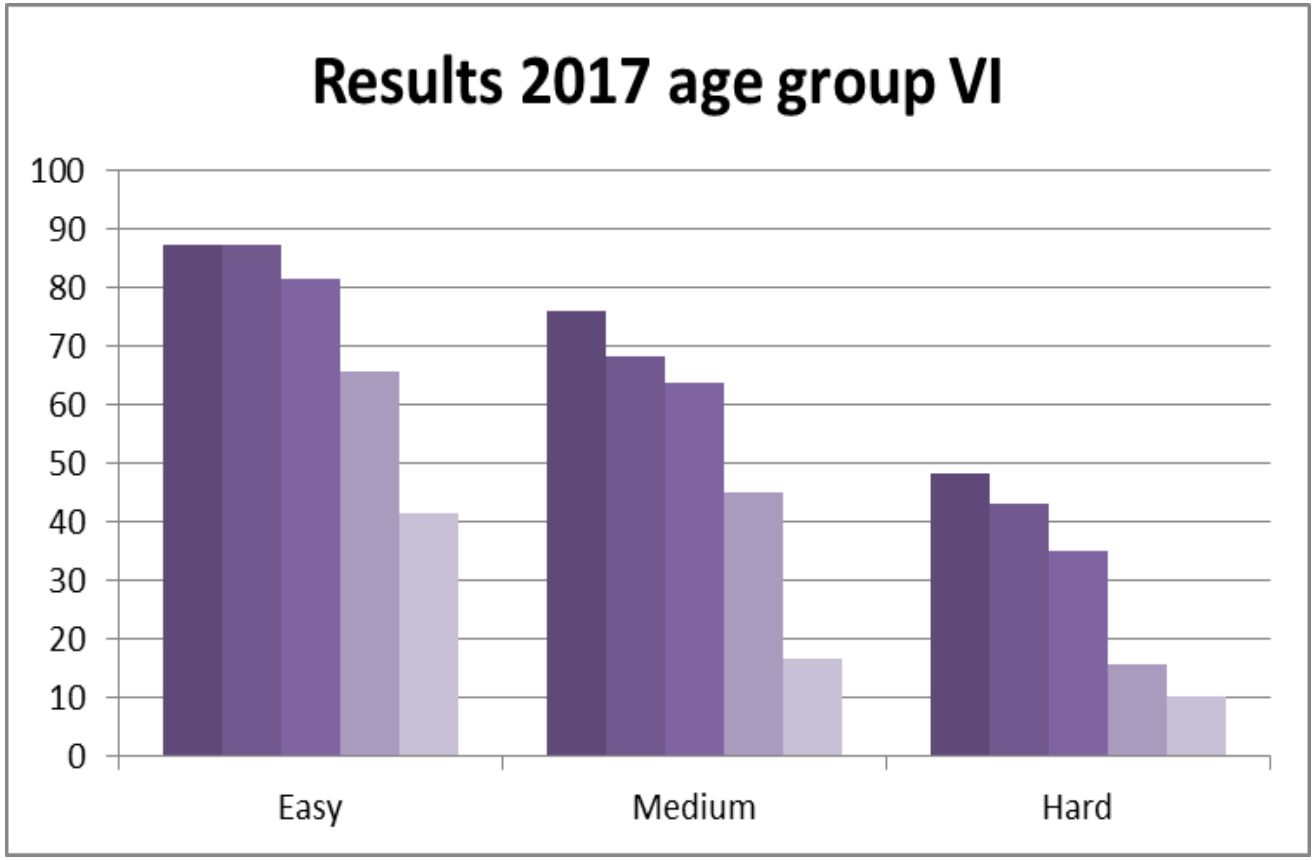
Over 50 countries
Over 2 000 000 participants
Task Pool with 150 tasklets


Austria
Australia
Azerbaijan
Belgium
Bulgaria
Canada
Croatia
Cyprus
Czechia
BOBŘÍK informatiky
Egypt
BEBRAS EGYPT
Estonia
Finland
France
Germany
Hungary
Iceland
Indonesia
Iran
Ireland
Israel
Italy
Japan
Latvia
Lithuania
Malaysia
Beverwedstrijd
Netherlands
New Zealand
INTERNATIONAL BEBRAS INFORMATICS CONTEST
PAKISTAN
Pakistan
Poland
Romania
БОБЁР
Russia
Serbia
Singapore
Slovakia
Slovenia
South Africa
South Korea
Spain
Sweden
Switzerland
Taiwan
Turkey
United Kingdom
USA
Бобер
Ukraine
Vietnam
Belarus

Contest format in the Netherlands

- 5 ages groups (8-10, 10-12, 12-14, 14-16 and 16-18 years)
- First round organized in the Bebras week
- Contestants have 40 minutes for 15 tasks
- We decide and announce which tasks are Easy, Medium or Hard
- The result for a question depends on the difficulty level

Predicting difficulty level is hard…



Question difficulty – Leong (2006)

- **content difficulty**, depending of the subject matter being assessed
  computer science, computational thinking

- **stimulus difficulty**, related to comprehending words and phrases in a test item and accompanying information
  language, information processing

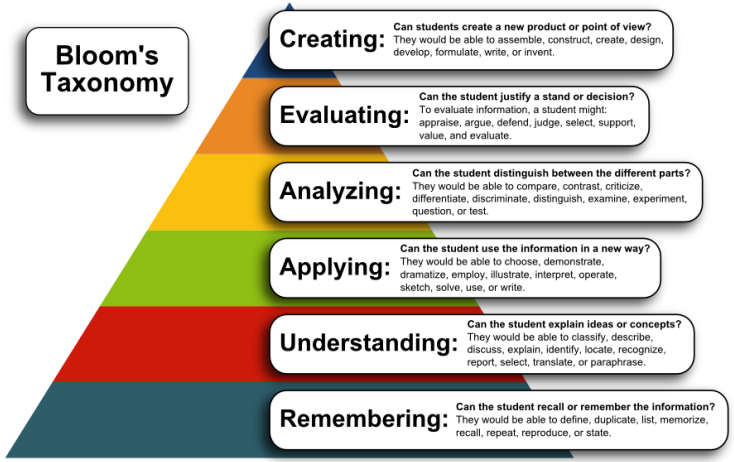
- **task difficulty**, referring to the work needed to formulate or discover the answer to the question
  cognitive process, thinking steps

Content difficulty

Bebras provides a challenge where reproduction should be useless; we aim to test insight and conceptual knowledge and 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 like Bloom's 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



Stimulus difficulty

- Lumley, Routinksy, Mendelovits and Ramalingam (2012): a scheme for describing the difficulty of reading items used in PISA. They compared the perceived and the empirical difficulty
- 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

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.

Stimulus difficulty: rephrasing the question

- 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, 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 and created a new figure. They obtained a higher success rate in their control group and a significant decrease in discrimination

Task difficulty

According to the cognitive load theory the limitations of the working memory are rarely taken into account in conventional instruction and assessment (Kirschner, 2002)

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)

Task difficulty: Cognitive load in test design

Elliot, Kurz, Beddow and Frey (2009) formulate recommendations:

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.

Questionnaires and rubrics

Van der Vegt (2013)

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?

Questionnaires and rubrics

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

Questionnaires and rubrics

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

How balanced are these questionnaires?

For all three instruments each scoring item is assigned to:

Content, stimulus or task difficulty

Questionnaire	Content	Stimulus	Task
Q1	30	20	50
Q2	25	50	25
R	20	40	40

Three methods applied to a specific contest

- First round 2017 age group VI in the Netherlands
- We filled in all questionnaires for each of the 15 questions (after the contest was already closed)
- It had to be performed in limited time; we don't really can use much time in preparing an actual contest.

Results for all tasks and instruments

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

What is the correlation between scores of prediction models and actual contest?

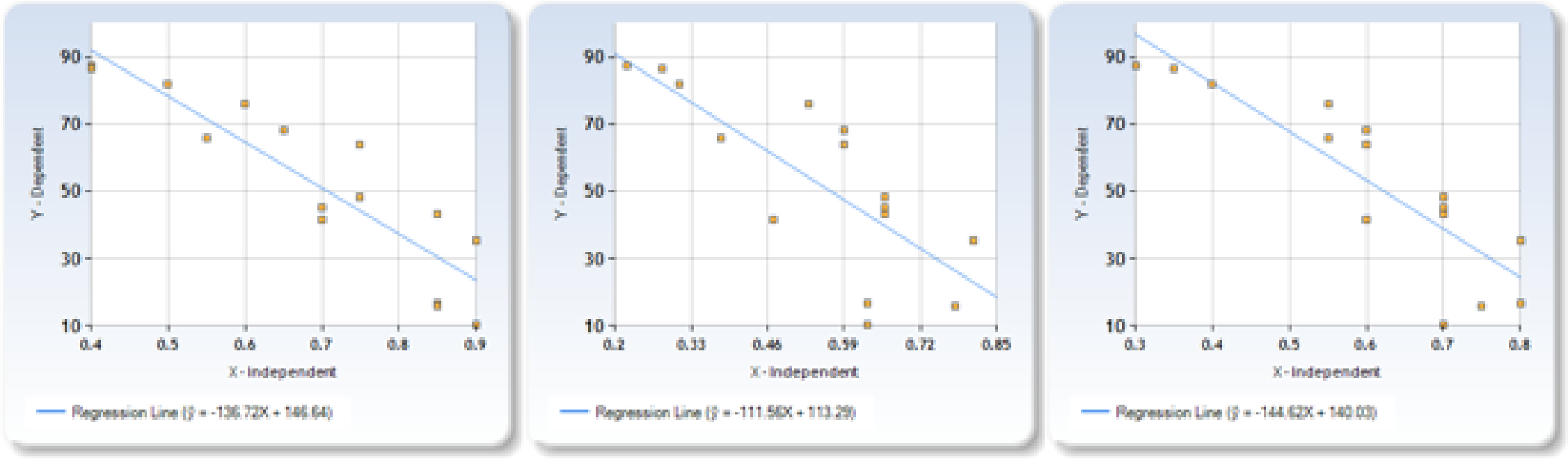


Fig.4. Linear regression for Q1, Q2 and R (from left to right) and success-rate for Bebras 2017.

Procedure using relative scoring

Holmes and Read (2018):

It is very hard to make absolute judgements on question difficulty

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

Kindle and Johnson (2011): 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.

Procedure using relative scoring

We toke a set of six different tasks

We asked a group of colleagues (researchers in computer science education) 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

And now?

- Tools developed to predict the difficulty level of a Bebras-task, can help to create a balanced contest. All three instruments looked at can be used for this goal
- Try to find the best balance for the weights on content, stimulus and task difficulty
- 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. Integrating questionnaires and relative scoring will be valuable
- Stimulus and task difficulty play an important role in the performance of contestants. Instruments used to predict question difficulty should include these insights