How presentation affects the difficulty of computational thinking tasks: an IRT analysis

Violetta Lonati
Università degli Studi di Milano
Milan, Italy
lonati@di.unimi.it

Dario Malchiodi
Università degli Studi di Milano
Milan, Italy
malchiodi@di.unimi.it

Mattia Monga
Università degli Studi di Milano
Milan, Italy
monga@di.unimi.it

Anna Morpurgo
Università degli Studi di Milano
Milan, Italy
morpurgo@di.unimi.it

ABSTRACT

This paper discusses how a few changes in some computational thinking tasks proposed during the Bebras challenge affected the solvers' performance. After the 2016 challenge held in November in our country (Italy), we interviewed some participants on the difficulties they had faced and we modified some of the tasks accordingly. We then proposed the whole set of tasks, with some of them modified, to pupils who had not participated to the challenge in November and compared performances in the two sessions. Using Item Response Theory, we measured the change in the distribution of difficulty and discrimination of the modified tasks. On the basis of the obtained results, we tried to better understand the many factors which influenced the difference in performances, both in the conceptual and cognitive task content and in its presentation (text, images, layout).

CCS CONCEPTS

• **Social and professional topics** → **Computational thinking**; **K-12 education**; *Informal education*;

KEYWORDS

Bebras, computational thinking, K-12 education, informal education, Item Response Theory

ACM Reference Format:
Violetta Lonati, Dario Malchiodi, Mattia Monga, and Anna Morpurgo. 2017. How presentation affects the difficulty of computational thinking tasks: an IRT analysis. In *17th Koli Calling International Conference on Computing Education Research.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3141880.3141900

Koli Calling 2017, November 16–19, 2017, Koli,Finland

ACM ISBN 978-1-4503-5301-4/17/11.
https://doi.org/10.1145/3141880.3141900

1 INTRODUCTION

The Bebras International Challenge on Informatics and Computational Thinking (http://bebras.org) is a yearly contest organized in several countries since 2004 [5, 10], with almost two million participants worldwide. The contest, open to pupils of all school levels (from primary up to upper secondary), is based on tasks rooted on core informatics concepts, yet independent of specific previous knowledge such as for instance that acquired during curricular activities. The Bebras community organizes yearly an international workshop devoted to proposing a pool of tasks to be used by national organizers in order to set up the local contests. The national organizers then translate and possibly adapt the tasks to their specific educational context[1]. Having in mind the goal of proposing an entertaining learning experience, tasks should be moderately challenging and solvable in a relatively short time (three minutes on average). Besides being used during contests, Bebras tasks are more and more used as the starting points for educational activities carried out by single teachers [7, 14]. Bebras tasks were also used to measure improvements of students' attitude to computational thinking [21].

Therefore, a correct assessment of the *difficulty* (*i.e.,* how probable it is that the Bebras participants will solve it) of a task is of great importance for it to be useful for teachers and enjoyable for pupils. This is often not a primary concern of authors, which tend to be more focused on the *disciplinary interest* of tasks (that is, the informatics concepts behind them). A task perceived as straightforward by an author might prove difficult for solvers because of unspecified common hypotheses or other logical hurdles, especially hard to assess across the long span of K-12 education. Thus evaluating the difficulty of a task is actually not easy, as shown by some analyses of the participants' performance highlighting a mismatch between the difficulty as perceived by authors and by solvers [2, 6, 23].

In this paper we analyze the influence of several elements in Bebras computational thinking (CT) tasks on their difficulty. Namely, we interviewed some classes who had participated in the 2016 Italian edition of the contest, we collected their comments, and tried to detect the difficulties they faced in solving tasks, or why they misunderstood them. On the basis of these observations we

[1] For instance, the French edition is based on interactive versions of the tasks, and pupils can repeatedly submit answers until achieving a correct solution. Participation is individual in some countries and team-based in other ones; in some school systems the participation to the Bebras is compulsory.

formulated some hypotheses on the sources of these difficulties/misunderstandings, and revised the texts of some tasks accordingly. Finally, we proposed to a new group of students the modified version of the tasks, since we wanted to study the difference in performances w.r.t. the "official" contest. The two groups, however, were not easily comparable: they had a different number of participants, they were neither randomized nor pre-selected according to any predefined profile. In particular we did not know the "ability" of the participants of the two cohorts. Thus, we needed a way to measure the difficulty of tasks w.r.t. the ability of the solvers. To this end, we resorted to Item Response Theory (IRT). Specifically, we fitted difficulty and discrimination (*i.e.,* a measure of how changes in difficulty impact the probability of having correct solutions) of tasks through a two-parameter model on the basis of the observed performances, leading to an estimate of their distribution for the two sessions of the contest. By analyzing the difference between these distributions we can confirm that the observed performances in the more recent session were consistent with our prediction in several cases.

The paper is organized as follows: in Section 2 we discuss the difficulties that a Bebras task may present and the features that contribute to such difficulties, in Section 3 we present the tasks used in 2016 Bebras challenge in Italy, in Section 4 we present the methodology we adopted to analyse the effects of tasks variation and compare the two groups, in Section 5 we discuss the finding of our analysis, and in Section 6 we draw some conclusions.

2 BEBRAS TASK DIFFICULTIES

Assessing tasks' difficulty is not easy and it is especially hard with Bebras tasks, since they cover a broad spectrum of topics and skills, and they do not refer to a prefixed set of learning goals. Accordingly, the difficulty of a Bebras task is generally assigned by its authors or translators based on their subjective judgment and experience as teachers. To the best of our knowledge there is no accepted framework suitable to diagnose such difficulties for these kinds of tasks, differently from other fields like programming [19].

However, some general guidelines can be extracted from previous research on tasks' difficulties that has been conducted both on tasks in general and specifically on Bebras tasks [15, 23]. Here we report a brief account of the literature, considering what can be reasonably applied to Bebras computational thinking tasks.

Tasks can present two main kinds of difficulties [8].

- Some difficulties are "intrinsic" with the task and related to its content; they may concern both the *concepts* involved in the task, and in particular how much they are complex and/or abstract, and the *processes*, that is the cognitive operations and the use of cognitive resources implied to solve the task [16, 17]. To account for this kind of difficulties one can for instance consider the number of objects/constraints in the problem, the number of transformations required to solve it, the dimension and density of the solution space, if the solution's representation can fit in working memory, if feedback is given [23]. Another element that contributes to the process difficulty can be envisaged by considering Bloom's taxonomy of the cognitive domain, since tasks that ask to understand, apply, analyze, evaluate or create require increasingly advanced cognitive skills [1].
- Some others are "surface" difficulties, in that they depend on the task format, which includes mark schemes (which we are not considering in this study) and linguistic, structural, and visual aspects. Among them, we mention the text wording (*e.g.,* the choice of terms, the use of synonyms or repetitions, the length of sentences) [15], the presence and use of examples, the use of diagrams and images [22], and the layout of the task.

Moreover, task difficulties may be intentional, *i.e.,* the designer wants the solvers to address them, or unintentional. For instance, in some Bebras tasks arithmetic computation is needed but it should not be the main difficulty when solving a task; whenever this happens, the designers have probably underestimated such arithmetics difficulty. In these cases the difficulty can be removed either by a content change (*i.e.,* smaller numbers to be processed) or by a surface change (like introducing the support of a calculator).

In particular, figures and examples are usually inserted into tasks to help solvers understand the problem and build a mental representation of it. However, the effect of images is particularly relevant and somewhat unpredictable depending also on the solver's past experience [4]; thus such tokens may distract the focus on specific aspects of the figure or of the example and result in misleading the solvers and adding unintentional difficulties.

3 BEBRAS CHALLENGE IN ITALY

In Italy the Bebras Challenge is proposed to five categories of pupils, from primary to secondary schools, who participate in teams of at most four pupils; the 2016 edition saw the participation of 28,407 pupils. Each category had 15 tasks to be solved within 45 minutes, using an online web-based platform.

Differently from other countries, only a few tasks are based on multiple-choice questions; most of them are *sophisticated* tasks, in that they present open answer questions, they are interactive, or they require complex, combined answers [3]. Consistently, sometimes partial scores are contemplated and there are no penalties for wrong answers.

In each category, tasks are divided by the organizers into three difficulty levels, and each level corresponds to an increasing number of points that can be achieved. Some of the tasks are repeated in more than one category with different difficulty level and scoring [2]. In particular, in the organizers' intention each category should have a couple of tasks that are accessible to everybody and a couple of tasks that require more complex reasoning and higher cognitive processes to be solved. However, apart from few exceptions, solvers are not expected to address and solve all the proposed tasks: since in Italy computer science is not a mandatory subject for all school levels and the are no standard curricula in the age groups considered, the organizers' idea is to expose pupils to a variety of different tasks and leave to them the choice of those they like the most or find most accessible. Thus, a rate of 30-40% solvers scoring some points can be acceptable for a task.

[2]However, in the analysis presented here we considered just whether the task was fully or partially resolved, ignoring the actual score assigned to the task. Since the scores were differentiated by task's difficulty level, this avoids the problems with a potential misclassification of the tasks.

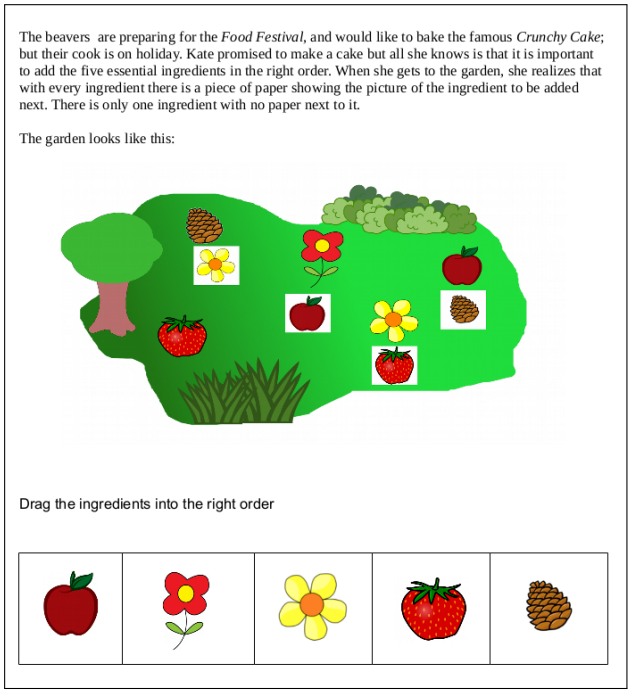

Figure 1: The original form of task "Recipe".

Difficulty levels are assigned to tasks according to designers' experience and intuition; unfortunately, such difficulty predictions often turn out to be erroneous. After the session that was held in November 2016, we analyzed pupils' performances and unexpectedly detected very low success rates for some tasks we didn't consider very challenging.

We then interviewed nine classes (around 180 pupils) who had participated in the Bebras 2016 contest with the goal of collecting comments and detecting the misunderstandings/difficulties they faced in solving the tasks. We met the classes after the contest, we discussed the tasks with each class as a whole and collected also many individual comments. During the discussion and the interviews, all tasks with their own answers and the correct solutions were available to each team. Because of organizing issues we met the classes only 4-6 weeks after the contest. However, pupils had not seen yet their results and the correct answers, which kept their attention high; indeed they were generally able to remember the tasks with some details, report the ideas and discussions they shared, and answer our questions aimed at understanding their difficulties and misunderstandings.

On the basis of the information collected, we made some hypothesis on the sources of the difficulties; we then developed new versions for seven of the tasks, trying to remove the detected unintentional difficulties and ambiguities. The original versions of the seven tasks (translated into English) are reported in Figures 1–7 whereas in Section 5 we describe how we changed them. The other tasks were kept unaltered, in order to get a reliable benchmark.

We administered the new version of the challenge to new solvers, who were sampled on a volunteer basis, with an open call to teachers in our primary and lower secondary schools network, provided

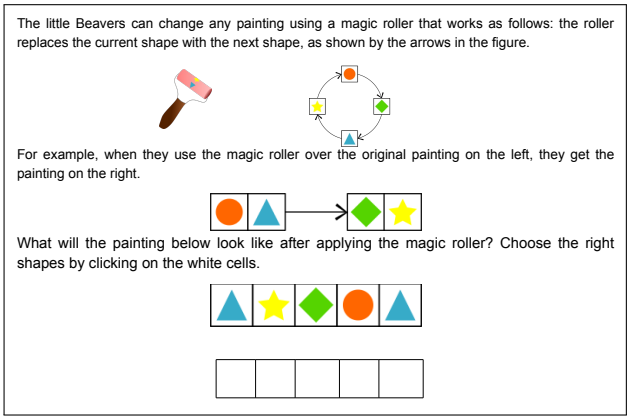

Figure 2: The original form of task "Brush".

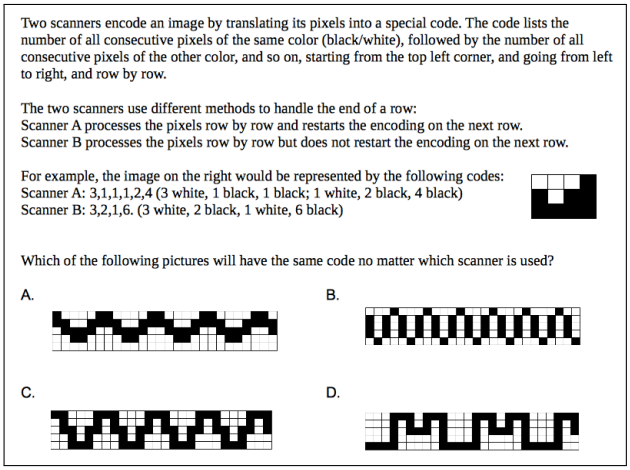

Figure 3: The original form of task "Scanner".

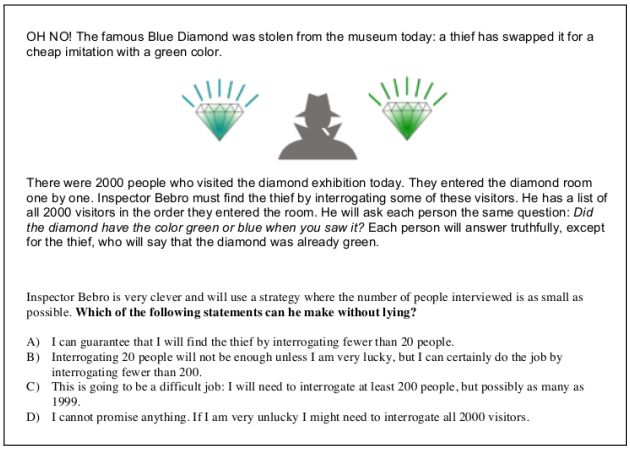

Figure 4: The original form of task "Thief".

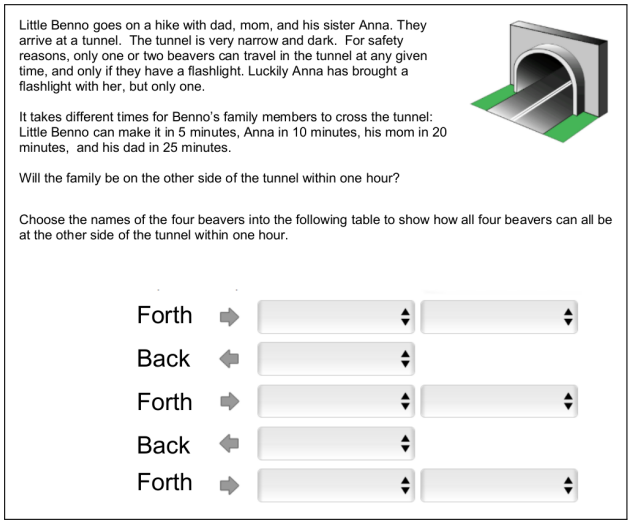

Figure 5: The original form of task "Tunnel".

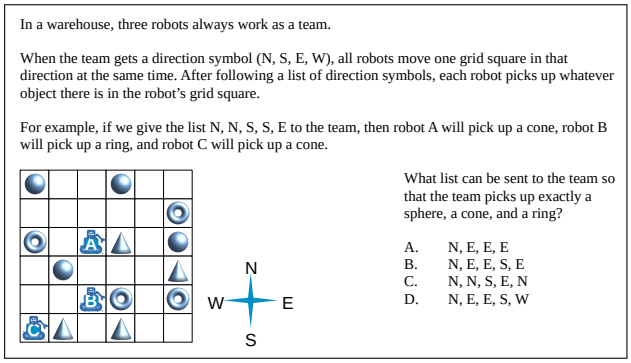

Figure 6: The original form of task "Directions".

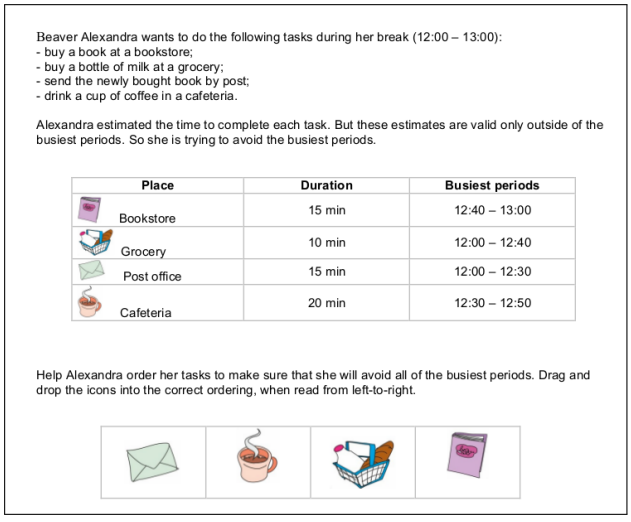

Figure 7: The original form of task "Four errands".

they had not participated in the challenge in November 2016. Hence, the experiment involved only categories I, II and III, for a total of around 650 pupils (210 teams).

Thus, we need to compare two inevitably different populations: we will call them respectively 'November 2016' (NOV16) and 'March 2017' (MAR17). In the following section we illustrate the methodology we used to compare the results achieved by the two groups.

4 METHODOLOGY

The two populations differ in cardinality: NOV16 had 5,871 solvers divided into three categories: I (primary schools, 8-10 years, 2,658 solvers), II (lower secondary schools, 10-12 years, 2,165 solvers), and III (lower secondary schools, 12-13 years, 1,048 solvers), MAR17 had 210 solvers (79, 89, and 42, respectively, in the three categories). But also the motivation and skills of the pupils can be, on average, very different: in fact when we asked the teachers to organize the March session, we explicitly asked for colleagues and pupils who had not participated before in Bebras. Table 1, 2, and 3 collect the results for all the tasks proposed to the three categories. Table 4 summarizes the performances of all the categories: the overall performances of the MAR17 population are worse than the NOV16 one, especially for categories I and III; category II gets a comparable mean but greater standard deviation from the mean.

4.1 Item Response Theory

In order to compare results achieved by groups with different skill levels we resorted to Item Response Theory (IRT) [11]. IRT is routinely used to evaluate massive educational assessment studies like OECD's PISA (Programme for International Student Assessment), and it has already been applied to Bebras and other informatics competitions [2, 12, 13]. IRT models each solver with an *ability* parameter and links it to the probability of a correct solution via a logistic function. Such a function is a characteristic of each task (*item*) and it defines its *response* to the solver ability. Response functions are described by a number of parameters: we used a model with two parameters, the *difficulty* of a task and its *discrimination*. Difficulty locates the response function: if the ability of the solver is greater than the difficulty, the probability is greater than 0.5. Discrimination defines the slope of the response curve: a high discrimination means that a small increase in the ability of the solver has a huge impact on the probability of solving it; a discrimination of 0 defines a task in which the ability of the solver does not matter at all. Figure 8 shows some examples of logistic response functions. It is worth noting that all that counts in the model are the relative values of the parameters (there is no absolute measure of ability): thus to fit it to data it is necessary to *identify* ability with conventional values. Adopting a common practice [9, 20, 24], we assumed that, overall, ability has mean = 0 with respect to an arbitrary reference point and standard deviation = 1.

4.2 Hierarchical Bayesian regression

In order to estimate the difficulty and discrimination of each task, we implemented a probabilistic model with Stan [20]. Stan is a software which, given a statistical model, uses Hamiltonian Monte Carlo (HMC) sampling (a very efficient form of Markov chain Monte Carlo (MCMC) sampling) to approximate the *posterior* probability

Table 1: Number of correct answers and failure ratios for category I, obtained in the November 2016 (2658 participants) and March 2017 (79 participants) sessions; starred names denote modified quizzes; tasks are sorted as they appeared in the challenge. NS, PS, and FS columns report the number and percentages of null, partial, and full scores, respectively, while Δ shows the difference between failure rates.

		NOV16			MAR17			
	Bebras ID	NS	PS	FS	NS	PS	FS	Δ
*Brush	PK-03	1699 (64%)	35 (1%)	924 (35%)	52 (66%)	9 (11%)	18 (23%)	1.90%
*Recipe	HU-02	2198 (83%)		460 (17%)	39 (49%)		40 (51%)	-33.33%
Messages	UK-06	1063 (40%)		1595 (60%)	34 (43%)		45 (57%)	3.05%
Ladybugs	SK-10	1472 (55%)		1186 (45%)	52 (66%)		27 (34%)	10.44%
ColorFlowers	SK-04	1645 (62%)	301 (11%)	712 (27%)	59 (75%)	10 (13%)	10 (13%)	12.79%
Cones	FR-02	1678 (63%)		980 (37%)	62 (78%)		17 (22%)	15.35%
BeaverBall	JP-03	934 (35%)	1060 (40%)	664 (25%)	38 (48%)	20 (25%)	21 (27%)	12.96%
*Four errands	LT-03	2434 (92%)		224 (8%)	75 (95%)		4 (5%)	3.36%
Corks	JP-06	1464 (55%)		1194 (45%)	50 (63%)		29 (37%)	8.21%
Soccer	US-07b	1246 (47%)		1412 (53%)	51 (65%)		28 (35%)	17.68%
*Directions	IE-05	1839 (69%)		819 (31%)	61 (77%)		18 (23%)	8.03%
Robot	FR-04	1512 (57%)		1146 (43%)	53 (67%)		26 (33%)	10.20%
*Scanner	MY-02	2170 (82%)		488 (18%)	64 (81%)	5 (6%)	10 (13%)	-0.63%
BagLift	CZ-02a	726 (27%)	1702 (64%)	230 (9%)	24 (30%)	50 (63%)	5 (6%)	3.07%
Rafting	LT-02	1789 (67%)	661 (25%)	208 (8%)	48 (61%)	27 (34%)	4 (5%)	-6.55%

Table 2: Number of correct answers and failure ratios for category II, obtained in the November 2016 (2165 participants) and March 2017 (89 participants) sessions; starred names denote modified quizzes; same notations as Table 1.

		NOV16			MAR17			
	Bebras ID	NS	PS	FS	NS	PS	FS	Δ
Cones	FR-02	972 (45%)		1193 (55%)	37 (42%)		52 (58%)	-3.32%
BeaverBall	JP-03	365 (17%)	965 (45%)	835 (39%)	16 (18%)	40 (45%)	33 (37%)	1.12%
*Four errands	LT-03	1721 (79%)		444 (21%)	66 (74%)		23 (26%)	-5.33%
Corks	JP-06	813 (38%)		1352 (62%)	36 (40%)		53 (60%)	2.90%
Soccer	US-07b	918 (42%)		1247 (58%)	50 (56%)		39 (44%)	13.78%
*Directions	IE-05	1111 (51%)		1054 (49%)	45 (51%)		44 (49%)	-0.75%
Robot	FR-04	999 (46%)		1166 (54%)	33 (37%)		56 (63%)	-9.06%
*Scanner	MY-02	1780 (82%)		385 (18%)	83 (93%)	2 (2%)	4 (4%)	11.04%
BagLift	CZ-02a	919 (42%)	951 (44%)	295 (14%)	35 (39%)	39 (44%)	15 (17%)	-3.12%
Rafting	LT-02	884 (41%)	767 (35%)	514 (24%)	35 (39%)	32 (36%)	22 (25%)	-1.51%
Salad	DE-08	1201 (55%)	295 (14%)	669 (31%)	47 (53%)	12 (13%)	30 (34%)	-2.66%
Cannon	IT-06	352 (16%)	1728 (80%)	85 (4%)	20 (22%)	67 (75%)	2 (2%)	6.21%
Mug	TW-05	1815 (84%)		350 (16%)	70 (79%)		19 (21%)	-5.18%
HealthCare	CH-03	1573 (73%)		592 (27%)	61 (69%)		28 (31%)	-4.12%
*Thief	BE-02	1982 (92%)		183 (8%)	79 (89%)		10 (11%)	-2.78%

of the parameters of interest. In particular, if Y_N and Y_M denote respectively the results collected during the NOV16 and MAR17 sessions, we wanted to get the following probability densities:

$$P(\alpha_j - \alpha_{j^*} | Y_N \cup Y_M) \quad j \in \text{tasks}, \tag{1}$$

$$P(\beta_j - \beta_{j^*} | Y_N \cup Y_M) \quad j \in \text{tasks}, \tag{2}$$

where α, β are respectively the discrimination and the difficulty associated to item (task) j and its modified version j^*. The statistical model sampled is a hierarchical one, with the following *prior* distributions:

$$\overline{\beta}, \sigma_\alpha, \sigma_\beta \sim Cauchy(0, 5), \qquad \theta \sim Normal(0, 1),$$
$$\beta \sim Normal(0, \sigma_\beta), \qquad \alpha \sim LogNormal(0, \sigma_\alpha),$$
$$y \sim BernoulliLogit(\alpha \cdot (\theta - (\beta + \overline{\beta}))).$$

In this model we assumed a Cauchy weakly informative prior distribution on hyper-parameters $\overline{\beta}$ —the mean difficulty used as a reference point in the logistic—, σ_β, and σ_α —the standard deviation respectively of difficulty and discrimination—. The ability is then supposed to be normally distributed with mean = 0 and standard

Table 3: Number of correct answers and failure ratios for category III, obtained in the November 2016 (1048 participants) and March 2017 (42 participants) sessions; starred names denote modified quizzes; same notations as Table 1.

		NOV16			MAR17			
	Bebras ID	NS	PS	FS	NS	PS	FS	Δ
*Directions	IE-05	452 (43%)		596 (57%)	21 (50%)		21 (50%)	6.87%
Robot	FR-04	343 (33%)		705 (67%)	16 (38%)		26 (62%)	5.37%
*Scanner	MY-02	797 (76%)		251 (24%)	38 (90%)	1 (2%)	3 (7%)	14.43%
BagLift	CZ-02a	597 (57%)	262 (25%)	189 (18%)	28 (67%)	9 (21%)	5 (12%)	9.70%
Rafting	LT-02	452 (43%)	152 (15%)	444 (42%)	18 (43%)	9 (21%)	15 (36%)	-0.27%
Salad	DE-08	513 (49%)	103 (10%)	432 (41%)	21 (50%)	6 (14%)	15 (36%)	1.05%
Cannon	IT-06	151 (14%)	846 (81%)	51 (5%)	9 (21%)	32 (76%)	1 (2%)	7.02%
Mug	TW-05	660 (63%)		388 (37%)	26 (62%)		16 (38%)	-1.07%
HealthCare	CH-03	531 (51%)		517 (49%)	25 (60%)		17 (40%)	8.86%
*Thief	BE-02	975 (93%)		73 (7%)	36 (86%)		6 (14%)	-7.32%
MedianFilter	RU-02	601 (57%)		447 (43%)	25 (60%)		17 (40%)	2.18%
Bubbles	IT-03	572 (55%)	406 (39%)	70 (7%)	17 (40%)	23 (55%)	2 (5%)	-14.10%
*Tunnel	CH-04a	571 (54%)	447 (43%)	30 (3%)	32 (76%)	6 (14%)	4 (10%)	21.71%
Islands	FR-03	736 (70%)	309 (29%)	3 (0%)	33 (79%)	9 (21%)	0 (0%)	8.34%
Colors	UK-04	1001 (96%)		47 (4%)	41 (98%)		1 (2%)	2.10%

Table 4: Distributions of full scores.

	count	mean	std	min	25%	50%	75%	max
(NOV16, I)	2658	4.61	2.52	0	3	4	6	14
(MAR17, I)	79	3.82	2.60	0	2	3	6	12
(NOV16, II)	2165	4.79	2.54	0	3	5	6	13
(MAR17, II)	89	4.83	3.03	0	2	5	7	11
(NOV16, III)	1048	4.05	2.11	0	3	4	5	12
(MAR17, III)	42	3.55	1.80	0	2	3	5	7

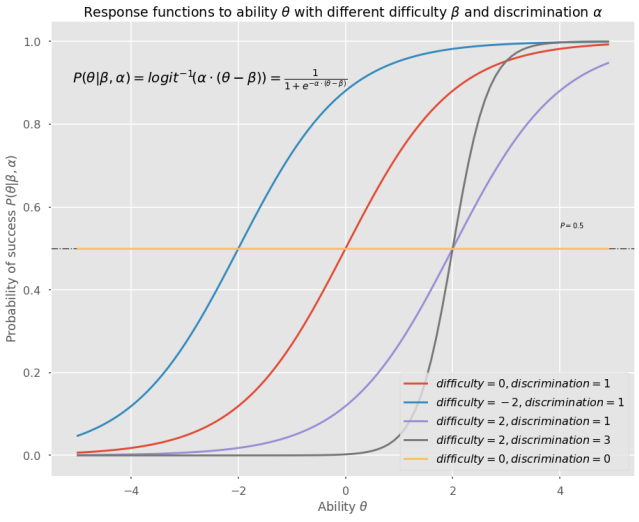

Figure 8: Logistic response functions

deviation = 1, the difficulty normally distributed with mean = 0 and standard deviation = σ_β, and the logarithm of discrimination is normally distributed with mean = 0 and standard deviation = σ_α. The correctness y of each item is finally sampled according to a Bernoulli process where the probability of success is computed with the logistic model described above. These are quite standard choices for Bayesian IRT (see [9, 20]).

We sampled the Stan Monte Carlo model for 4,000 iterations, throwing away the first 2,000 results (50% warm-up iterations). The results have all the typical properties of converging models, in particular the $\hat{R}$ statistic is close to 1 for every parameter of interest (a necessary, but unfortunately not sufficient, condition for convergence). Results are indeed sensible (*i.e.,* difficulties of *NOV*16 tasks where we have a lot of data are consistent with the observed performance), therefore we are rather confident that our model is plausible and useful to infer latent parameters.

From the model we can recover the posterior distributions of probabilities (1–2). For example, Figure 9 shows the distribution of difficulty and discrimination for a task in the NOV16 session, and Figure 10 shows the distributions of the *differences* in the difficulty and discrimination parameters for two tasks modified in the MAR17 session. The pictures show also the 95% high-density interval (HDI), *i.e.,* the interval were 95% of probability is concentrated. Thus, for example, one can conclude that, under the hypotheses given by the prior distributions, the probability that the change Δ of the difficulty of the task "Recipe" given the observed results approximately lay in the interval $(-2.4, -1.5)$, is 0.95: this interval is a negative one and it does not contain 0, therefore is highly probable that the change in the task made it easier. The key idea is that, thanks to the IRT modeling of the relationship between ability and difficulty, we are able to compare the NOV16 and MAR17 sessions: the common tasks make it possible to estimate the respective abilities of the solvers and to scale the parameters accordingly.

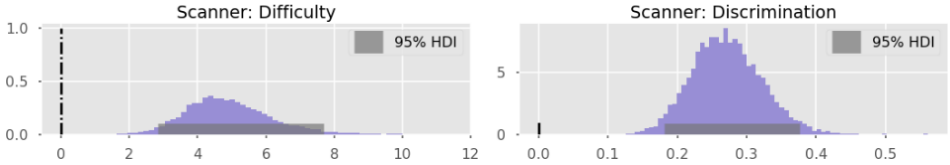

Figure 9: Distributions of difficulty and discrimination in the NOV16 session; the horizontal bar depicts the 95% high density interval.

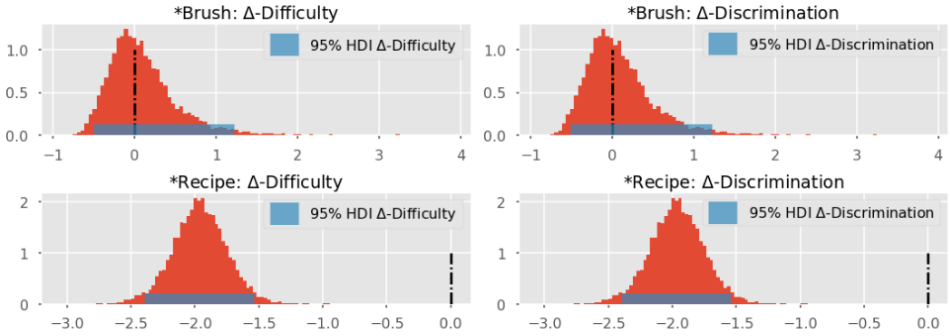

Figure 10: Distributions of the differences of difficulty and discrimination between the NOV16 and MAR17 version; the horizontal bar depicts the 95% high density interval.

4.3 Model checking and cross validation

The main threat to the validity of our analysis is, of course, how much the statistical model we implemented is a useful generative abstraction of the reality which produced the observed data. Indeed the parameters estimated by our model for the benchmark tasks are consistent with the intuitive perception of their difficulty and discrimination as we get them from the performance of the participants.

As a cross-validation we fitted the data also with a prepackaged non-Bayesian IRT model (by leveraging on TAM [18]). The resulting parameters, although not identical, are strongly correlated with ours: the difficulties computed by TAM have Spearman's correlation of 0.90, 0.96, and 0.92 (for the three categories); the discriminations 0.97, 0.93, and 0.73. The effects we considered highly probable (see bold numbers in Table 5) have a TAM p-value < 0.05 in four cases out of eight: the changes in the difficulty of 'Thief, II' ($p = 0.18$) and 'Thief, III' ($p = 0.24$) would not be significant under the TAM model; the changes in the discrimination of 'Scanner, I' ($p = 0.39$) and 'Scanner, II' ($p = 0.08$) would not be confirmed to be significant under the TAM model and a $p < .05$ threshold. The TAM model is faster (seconds vs. hours of MCMC sampling) and somewhat easier to manage, at least outside computer science circles, since using Stan needs higher programming skills. Nevertheless, we found the Stan model more useful: the hypotheses on parameters and hyperparameters are explicit and they can be of arbitrary complexity (at least if one has enough time to dedicate to sampling…). For example, for abilities we tried also a Student-t prior distribution with seven degrees of freedom to allow for more outliers in skills; the results, not reported here for brevity, are consistent (almost identical, in fact) with the ones given. All in all, having an explicit generative model, as in Stan, makes the analyst well aware of what kind of machinery is crunching the data and it is easier to get a critical attitude towards the results. Moreover, the posterior probability distributions of the parameters give a quantified view of the effects, simpler to interpret than p-values.

The anonymized data, the source code of the models, and the tables of results are available at: https://gitlab.com/aladdin-unimi/bebras-stan-stats.

5 RESULTS AND DISCUSSION

For some tasks the IRT analysis of the results provides evidence to attest the validity of our hypotheses. In some other cases, data appear to be too fragile to draw firm conclusions; in general, though, our hypotheses are not disconfirmed. Table 5 summarizes the effects on difficulty and discrimination revealed by the analysis.

In what follows we analyse one by one all the tasks that we modified, discussing the difficulties we envisaged, the revision we proposed, the observed outcomes, and providing some possible interpretations.

5.1 Recipe

"Recipe" is a task on linked lists, and is depicted in Figure 1. The analysis of this task, whose success rate was incredibly low with respect to the assigned difficulty, shows some recurrences: indeed, several answers proposed ingredients in an order essentially derived from the spatial disposition of the latter in the figure (following a left-to-right path or proceeding approximately in a circular way). The interviews highlighted that the text was not understood and generally read with no care. Thus, the low score is more likely due to the chosen presentation, rather than to an intrinsic difficulty.

We tried to improve the presentation, illustrating the rule described in the text, yet avoiding the use of a separate example

Table 5: Effects on difficulty (β) and discrimination (α); bold numbers denote a probability > 0.95 or < 0.05.

	Δ_β 2.5%	Δ_β 97.5%	$P(\Delta_\beta > 0)$	$\hat{R}_\beta$	Δ_α 2.5%	Δ_α 97.5%	$P(\Delta_\alpha > 0)$	$\hat{R}_\alpha$
Brush, I	-0.51	1.22	0.53	1.0	-0.27	1.99	0.88	1.0
BrushPartial, I	-3.21	5.76	0.73	1.0	-4.33	2.78	0.31	1.0
Four errands, I	-1.04	4.15	0.72	1.0	-0.69	0.85	0.34	1.0
Four errands, II	-0.37	1.99	0.79	1.0	-1.31	0.05	**0.03**	1.0
Directions, I	-0.98	2.46	0.51	1.0	-0.29	0.88	0.70	1.0
Directions, II	-0.59	0.70	0.56	1.0	-0.64	0.62	0.36	1.0
Directions, III	-1.17	1.85	0.66	1.0	-0.86	0.52	0.18	1.0
Recipe, I	-2.40	-1.54	**0.00**	1.0	-0.46	2.90	0.82	1.0
Scanner, I	-5.54	0.88	0.06	1.0	0.01	0.93	**0.98**	1.0
Scanner, II	-3.08	4.92	0.54	1.0	-0.02	0.86	**0.96**	1.0
Scanner, III	-1.12	5.64	0.78	1.0	-0.27	1.76	0.81	1.0
Thief, II	-13.36	-2.82	**0.00**	1.0	0.06	0.96	**0.99**	1.0
Thief, III	-10.77	-0.19	**0.02**	1.0	-0.11	0.78	0.87	1.0
Tunnel, III	-2.24	5.23	0.58	1.0	-1.44	0.27	**0.07**	1.0
TunnelPartial, III	-5.89	5.71	0.50	1.0	-3.58	3.40	0.50	1.0

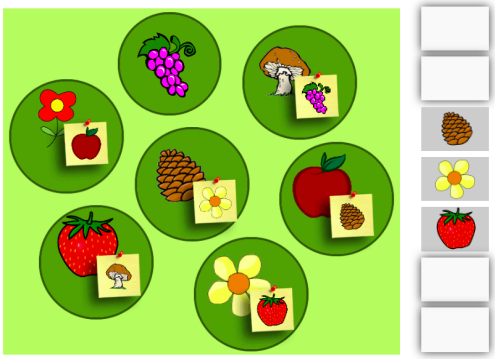

Figure 11: The new figures of task "Recipe".

(which could be misleading w.r.t. the question to be addressed). So, we added two ingredients and pre-filled three out of seven fields in the answer, and chose those ingredients avoiding regularities in their disposition, see Figure 11.

As a result, we obtained a definitely higher success rate, and a lower difficulty as confirmed by the model, as well as a significant decrease in discrimination (according to the model, the probability that discrimination was raised is essentially zero, see Figure 10 and Table 5). We may interpret this fact as follows: while in the first session only very motivated and careful solvers, paying attention to instructions and details, succeeded in answering the question, the task improvement had the effect of widening the base of pupils who correctly answered.

5.2 Brush

Task "Brush" is shown in Figure 2. We were very surprised by the low success rate, because the task seems easy to us. We analyzed the wrong answers and detected some recurrences: solvers were apparently misled by the horizontal, linear arrangement of figures to be transformed, and they simply tried to complete or repeat the

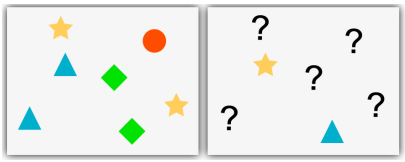

Figure 12: The new figure for the question of task "Brush".

initial pattern, like in analogy tasks (*i.e.*, "complete the sequence" tasks) [4, 8].

When we asked the pupils to explain with their own words what they were expected to do, it turned out they had looked at the figures without actually reading the text, and had been mislead by them, so most of the pupils had not caught the request correctly. Thus the low score seems to be caused by the presentation of the task (precisely, flawed in the reported example and figures), rather than by an intrinsic difficulty.

We modified the figures in the task: we changed the arrangement of the drawings to be transformed using a less regular 2D disposition, we removed the example picture, and we modified the question in order to embed the example into it, see Figure 12. We left the original scoring schema unaltered, assigning a partial score if four out of five drawings were correctly transformed, since the new 2D disposition might have introduced a distracting effect. As a result, we observed a higher number of such partial answers, thus the level of comprehension did actually rise. Even though the obtained model, trained considering as correct also a partial answer, doesn't show a substantial difference between `NOV16` and `MAR17` (see Figure 10), we see this continuity as a positive result of the introduced simplification, also taking account of the fact that the performances in `MAR17` were in average below those of `NOV16`.

5.3 Scanner

This (see Figure 3) was one of the most difficult tasks we proposed to younger solvers. The text of this task was rather long and it required

several cognitive steps: to understand two encodings, identify the critical parts within two figures in order to distinguish the corresponding encodings, and consequently analyze the figures.

The interviews highlighted that even understanding the two encodings was very difficult, so that many solvers limited themselves to guess among the multiple-choice answers, or found arbitrary relations between the figure in the example and the figures in these answers. Thus we tried to guide solvers in the various transitions toward the solution, converting the example into an intermediate open question having the aim of helping pupils to focus on understanding the encoding.

The results are puzzling, as while the difficulty of the second question turns out to be slightly lower (despite still being very high) and the discrimination is probably decreased, there is no appreciable correlation between partial answers to the two questions, and the number of correct answers is too little to be statistically measurable. We can risk the following interpretation: besides a few cases in which the first question has been correctly addressed, most of the times the answer to the second question is guessed at random, or based on misunderstandings or on inapplicable lines of reasoning. Paradoxically, partial answers to the second question have not been guessed totally at random. Indeed, in both sessions the first two answers are more frequent, and the right answer–which is in the last position–is the least frequent! This can be the result of applying the above mentioned erroneous lines of reasoning, which probably led solvers to believe that they have found a good answer before considering the last option, thus penalizing the correct one.

5.4 Thief

"Thief" (see Figure 4) is a task on binary search: a detective has the list of visitors to a museum in chronological order, among whom is a thief who has replaced a blue diamond with a faked green one. How many visitors must he interview to identify the thief, if all except the thief answer honestly when asked what color was the diamond?

Some of the teachers were surprised (and disappointed) by the low percentage of correct answers of their pupils since the topic had been proposed to the class, yet apparently they did not realize that the task could be tackled by using that strategy.

We asked pupils to tell us how they would proceed if they were the detective and they would question all the visitors in order. When we stressed the fact that it was not necessary to proceed sequentially, pupils started thinking at the possibility of "jumping" and eventually came up with binary search. We therefore thought of trying stressing this fact in the text of the task, by specifying explicitly that it was not necessary to proceed in sequential order, to see if it would impact on the success rate. The effect of this improvement is particularly evident on category II, where the discrimination definitely increased (see Table 5): in the first session, we believe answers were basically given randomly, while in the second one the recall of a known strategy was enhanced, in solvers with higher ability, by the change in the text.

5.5 Tunnel

"Tunnel" (see Figure 5) is a task on constraint scheduling: a family of four with only one flashlight have to go through a narrow and dark tunnel where only one or two people at a time can walk and the flashlight is needed. They have all different paces. The task is to schedule the back and forth walks of the characters, so that the family succeeds in passing through the tunnel in a given constrained time.

Many pupils found a way to make all characters go through the tunnel (gaining a partial score), but only a few respected the time constraint: the sequence they found took more time than allowed. When asked about this aspect, they replied they had not taken into account this request. This difficulty was an intrinsic one but of course an unintentional one: we didn't intend to test their ability to read carefully all requests or to make additions to keep track of time.

We thus decided to test if feedback on this issue would make a difference and added a time counter displaying the time their (also partial) solution was taking. This had the effect of slightly increasing the full scores rate and strongly decreasing the number of partial scores (correct solutions not satisfying the time constraint), hence the overall ratio of non-zero scores decreased. We give this fact the following interpretation: in the first session several solvers provided a wrong answer in the belief that it was correct (they simply didn't grasp that they used too much time), and in the second one they became aware of the constraint and thus did not insert such wrong answers. The discrimination is definitely decreased (see Table 5), which could be explained as for "Recipe" task (see Section 5.1): the risk of errors due to miscalculations is reduced and this broadened the basis of pupils who were able to answer correctly.

5.6 Directions

Task "Directions" is reported in Figure 6. The success rate of this task was not bad. Indeed, during the interviews pupils seemed at ease with the use of directions in order to drive the robot, although they appeared a bit confused in driving several robots simultaneously. The provided example appeared misleading, and we reckon that the order in which objects are listed (*i.e.*, not matching the order used to introduce robots), is a critical factor for a correct comprehension of the task. We modified the text, generically speaking about one object for each kind (thus avoiding to introduce an order), and we removed the example. Anyhow, in this case we didn't observe any significant improvement in the performances.

5.7 Four errands

"Four Errands" (see Figure 7) is a task on constraint scheduling: Alexandra wants to do four errands during her lunch break. She estimated the time to complete each errand, but these estimates are valid only outside of the rush hour. The task's request is find the right order in which to do the errands so as to avoid the rush hour for each one.

Also in this case the success rate was way below our expectations, and in most cases the observed recurrences suggested attempts aiming at doing the opposite of what was requested. During the interviews several pupils told they did not know the meaning of 'rush hour', thus we modified the request only by replacing such an expression with a more explicit wording. In spite of this, the collected data do not exhibit any significant improvement in the provided answers. This could be attributed to the use of negation

(also in the revised version of the task), which adds cognitive load to the solving process.

6 CONCLUSIONS

In this paper we discussed how a few changes in the presentation of computational thinking tasks may impact on the solvers' performance. After the Italian edition of the Bebras challenge held in November 2016, we interviewed some participants on the difficulties they had faced; we then modified some of the tasks and proposed them to pupils who had not participated in the challenge in November.

We compared performances in the two sessions and analyzed the effect of the changes on the difficulty and discrimination of each task by fitting a two-parameter logistic Item Response model, which allowed us to compare two populations that strongly differ both in cardinality and in ability. It is worth noting that a more traditional A/B testing approach that proposes two versions of the tasks to randomized solvers would not be fair here: while Bebras does not emphasize competition, it is in fact a contest and participants have the right to get an unbiased ranking.

For some tasks the IRT analysis of the results provides evidence to attest the validity of our hypotheses. In some other cases, data appear to be too fragile to draw firm conclusions; in general, though, our hypotheses are not disconfirmed.

A clear indication from the study warns about the use of examples and figures that must be chosen and designed with much care, since their effect can be distracting or distortionary instead of useful for understanding and addressing the question.

The study leaves us to the relevant problem of how to design brief tasks that still promote more complex reasoning and higher cognitive processes. This is generally not the case with simple multiple choice tasks, *e.g.,* "Scanner" or "Thief", where the multiple choice form of the question does not help guiding complex reasonings, resulting in excessive difficulties.

The study confirms that predicting tasks difficulty is far from being an exact science, and understanding difficulties and mistakes afterwards is not that easy too. In fact, although conceptual models of the difficulty of learning in specific computing activities (for example, programming) have been thoroughly considered, in this regard computational thinking has been the subject of far less studies. This paper represents a preliminary groundwork on which we plan to investigate proper difficulty models.

REFERENCES

[1] Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. 2000. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Edition* (2 ed.). Pearson, New York, USA.

[2] Carlo Bellettini, Violetta Lonati, Dario Malchiodi, Mattia Monga, Anna Morpurgo, and Mauro Torelli. 2015. How challenging are Bebras tasks? An IRT analysis based on the performance of Italian students. In *Proceedings of ITiCSE 2015*. ACM, Vilnius, Lithuania, 27–32.

[3] Andrew Boyle and Dougal Hutchison. 2009. Sophisticated tasks in e-assessment: what are they and what are their benefits? *Assessment & Evaluation in Higher Education* 34, 3 (2009), 305–319.

[4] Victoria Crisp and Ezekiel Sweiry. 2006. Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research* 48, 2 (2006), 139–154.

[5] Valentina Dagienė. 2010. Sustaining Informatics Education by Contests. In *Proceedings of ISSEP 2010 (Lecture Notes in Computer Science)*, Vol. 5941. Springer, Zurich, Switzerland, 1–12.

[6] Valentina Dagienė, Linda Mannila, Timo Poranen, Lennart Rolandsson, and Pär Söderhjelm. 2014. Students' Performance on Programming-related Tasks in an Informatics Contest in Finland, Sweden and Lithuania. In *Proceedings of ITiCSE 2014*. ACM, Uppsala, Sweden, 153–158.

[7] Valentina Dagienė and Sue Sentance. 2016. It's Computational Thinking! Bebras Tasks in the Curriculum. In *Proceedings of ISSEP 2016 (Lecture Notes in Computer Science)*, Vol. 9973. Springer, Cham, 28–39.

[8] Debra Dhillon. 2011. *Predictive models of question difficulty: A critical review of the literature.* Technical Report CERP-RP-DD-01022003_0. AQA Centre for Education Research and Policy, Manchester, UK.

[9] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press, Cambridge, UK.

[10] Bruria Haberman, Avi Cohen, and Valentina Dagienė. 2011. The Beaver Contest: Attracting Youngsters to Study Computing. In *Proceedings of ITiCSE 2011*. ACM, Darmstadt, Germany, 378–378.

[11] Ronald K. Hambleton and H. Swaminathan. 1985. *Item Response Theory: Principles and Applications.* Springer-Verlag, Berlin.

[12] Peter Hubwieser and Andreas Mühling. 2014. Playing PISA with Bebras. In *Proceedings of the 9th WiPSCE*. ACM, New York, NY, USA, 128–129.

[13] Graeme Kemkes, Troy Vasiga, and Gordon V. Cormack. 2006. Objective Scoring for Computing Competition Tasks. In *Proceedings of 2nd ISSEP (Lecture Notes in Computer Science)*, Vol. 4226. Springer, Berlin, Germany, 230–241.

[14] Violetta Lonati, Mattia Monga, Anna Morpurgo, Dario Malchiodi, and Annalisa Calcagni. 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) (Lecture Notes in Computer Science)*. Springer International Publishing AG, Cham, CH, 12. To appear.

[15] Wolfgang Pohl and Hans-Werner Hein. 2015. Aspects of quality in the presentation of informatics challenge tasks. In *Local proceedings of ISSEP 2015*. Ljubljiana University, Ljubljiana, Slovenia, 21–22.

[16] Alastair Pollitt. 1985. *What Makes Exam Questions Difficult?* Scottish Academic Press, Edinburgh.

[17] Alastair Pollitt and Ayesha Ahmed. 1999. A new model of the question answering process. In *IAEA conference, Slovenia, May 1999*. Cambridge Assessment, Bled, Slovenia, 1–14.

[18] Alexander Robitzsch, Thomas Kiefer, and Margaret Wu. 2017. *TAM: Test Analysis Modules.* CRAN. https://CRAN.R-project.org/package=TAM R package version 2.5-14.

[19] Judy Sheard, Simon, Angela Carbone, Donald Chinn, Tony Clear, Malcolm Corney, Daryl D'Souza, Joel Fenwick, James Harland, Mikko-Jussi Laakso, and Donna Teague. 2013. How Difficult Are Exams?: A Framework for Assessing the Complexity of Introductory Programming Exams. In *Proceedings of the Fifteenth Australasian Computing Education Conference - Volume 136 (ACE '13)*. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, Article 16, 10 pages. http://dl.acm.org/citation.cfm?id=2667199.2667215

[20] Stan Development Team. 2016. Stan Modeling Language Users Guide and Reference Manual Version 2.14.0. (2016). Retrieved Apr 2017 from http://mc-stan.org

[21] Suzanne Straw, Susie Bamford, and Ben Styles. 2017. *Randomised Controlled Trial and Process Evaluation of Code Clubs.* Technical Report CODE01. National Foundation for Educational Research. Available at: https://www.nfer.ac.uk/publications/CODE01.

[22] Monika Tomcsányiová and Martina Kabátová. 2013. Categorization of Pictures in Tasks of the Bebras Contest. In *Proceedings of ISSEP 2013 (Lecture Notes in Computer Science)*, Vol. 7780. Springer, Berlin, Germany, 184–195.

[23] Willelm van der Vegt. 2013. Predicting the difficulty level of a Bebras Task. *Olympiads in Informatics* 7 (2013), 132–139.

[24] Margaret Wu, Hak Ping Tam, and Tsung-Hau Jen. 2016. *Educational Measurement for Applied Researchers.* Springer, Singapore.