How Challenging are Bebras Tasks? An IRT analysis based on the performance of Italian students

Carlo Bellettini
Dept. of Computer Science
Università degli Studi di Milano
Milan, Italy
bellettini@di.unimi.it

Mattia Monga
Dept. of Computer Science
Università degli Studi di Milano
Milan, Italy
monga@di.unimi.it

Violetta Lonati
Dept. of Computer Science
Università degli Studi di Milano
Milan, Italy
lonati@di.unimi.it

Anna Morpurgo
Dept. of Computer Science
Università degli Studi di Milano
Milan, Italy
morpurgo@di.unimi.it

Dario Malchiodi
Dept. of Computer Science
Università degli Studi di Milano
Milan, Italy
malchiodi@di.unimi.it

Mauro Torelli
Dept. of Computer Science
Università degli Studi di Milano
Milan, Italy
torelli@di.unimi.it

1. INTRODUCTION

Several contests focusing on the informatics discipline have been organized worldwide in the last decades. Such events, typically arranged on a regular basis, mainly result from two attitudes of mind: one focused in selecting students particularly talented in the field (this attitude is notably reflected in the Informatics Olympiads), and another one aimed at spreading the basic concepts of the discipline to a vast audience of students, starting from the belief that such concepts should be taught even in the first stages of the educational system. Within this second vein, the Bebras contest is a competition organized on an annual basis in several countries since 2004, with an average number of participants higher than half a million in the recent editions.

The core of the Bebras contest organization is an annual international workshop gathering participants from all the involved countries, with the aim of proposing and jointly tuning an ample set of tasks. From such pool each country chooses a number of tasks to set up the local competition. Tasks are divided into six areas (such as algorithms, data structures, and so on) and their difficulty level is scored in the scale (easy, medium, hard), with the idea of proposing in each contest a suitable mix of tasks having different difficulty and belonging to different areas of informatics. Students are given a fixed amount of time to solve tasks, either choosing an answer from a set of four alternatives, or using an interactive interface based for instance on dragging and dropping items. Bebras questions should be small and moderately challenging tasks that enable an entertaining learning experience. The criteria for good tasks have been surveyed in 2, thus they well deserve a new name, tasklets: in general they should be fun and attractive, independent of specific curricular activities, be adequate for contestants’ age and the solution should take on average three minutes.

A correct assessment of the task difficulty is particularly important, as even partial failure in this job can result in a non-heterogeneous set of tasks proposed to pupils, with the effect of letting participants perceive the contest as too difficult or too easy, and ultimately not appealing. However, evaluating the difficulty of a task is actually not easy. Thus, after the conclusion of the competition it is advisable to infer the perceived task difficulty starting from the participants’ performance (see also 3, 15), in order to tune choices and strategies in the next competitions.

This paper shows the results of such an analysis on the
scores of the 2014 edition of the Italian competition, relying on the statistical techniques in the domain of Item Response Theory (IRT) \[^{[5,6]}\]. IRT is routinely used to evaluate massive educational assessment studies like OECD’s PISA (Programme for International Student Assessment), and \[^{[9]}\] used it to find psychometric constructs common to a set of tasklets of the German 2009 Bebras.

The paper is organized as follows: Sect. \[^{[2]}\] illustrates the specific features of the Italian Bebras/Kangourou and the data collected in the 2014 contest; Sect. \[^{[3]}\] describes our approach to IRT in order to model and measure the difficulty of the Bebras tasks; Sect. \[^{[4]}\] shows and discusses the results of our analysis; finally, Sect. \[^{[5]}\] draws some concluding remarks and outlines further refinements of the work.

2. DATA FROM BEBRAS/KANGOUROU

In Italy the competition is jointly organized with the Kangourou community \[^{[10]}\] since 2013. In the 2014 edition, 684 teams (2736 pupils) participated in the contest, divided into four age groups: Benjamin (grades 6–7, ages 11–12), Cadet (grades 8–9, ages 13–14), Junior (grades 10–11, ages 15–16) and Student (grades 12–13, ages 17–18). Table \[^{[2]}\] summarizes the composition and performances of the participating teams, detailed according to the corresponding age group.

In this paper we report some data about the quizzes of the last edition, the names of the tasklets in Table \[^{[1]}\] are the Bebras identifiers (with three extra quizzes), a * at the end of the name indicates that the quiz was proposed as an open question, a • that the question was significantly changed with respect to the Bebras one which inspired it. The Italian version of the contest contained 16 or 17 tasks for each age group with a time limit of 45 minutes (2700 seconds).

Compared to the international Bebras, the Kangourou flavour has a significant difference: it is a competition among teams of four pupils. In 2013 we realized that proposing exactly the same (translated from English) quizzes resulted in a too easy contest because the four team members can work in parallel and exploit a coordinate effort. This “team aspect” is in our opinion something valuable that we decided to preserve even after we joined the Bebras community. In the last two editions we thus changed the tasklets slightly with respect to the ones proposed to the international Bebras contestants, mostly by transforming multiple choice questions in open or interactive ones. In the latter case, partial scores are in general admitted, since the chance to make some minor mistakes may be relevant. In this analysis, answers with partial score are considered incorrect.

Since a lot of work is needed to organize and implement the contest (although tasklet delivery is computer based), a tasklet is usually proposed to more than one category, with different difficulty: for example, (see Table \[^{[3]}\] 2014-CA-07 was used for both Benjamins and Cadets, and considered of medium difficulty for the younger contestants and easy for the older ones. In the analysis, such tasklets are repeatedly considered, since the results by a category of contestants on a task are independent from the results by another category.

3. AN IRT MODEL FOR THE CONTEST

“She merite en toutes choses est dans la dificuité,” says Aramis in the Three Musketeers, Dumas’ masterpiece. But how to survey the difficulty of a Bebras tasklet? A rough measure is the analysis of the results (a set of quiz items) in which several parameters are taken into account: the difficulty of a quiz, but also the ability of the solver. In fact, we adopted a model in which we consider the ability \( \theta \) of a team and the probability \( p \) to answer correctly a quiz as a function of \( \theta \). The function associated to each tasklet is a sigmoid characterized by three parameters:

\[
p(\theta) = \eta + \frac{(1-\eta)}{1+e^{-\alpha(\theta-\delta)}}. \tag{1}
\]

In function \(^{(1)}\) the parameter \( \eta \) gives the minimum probability to guess the answer correctly even when the ability is very low: this is indeed the case of multiple choice questions, in which there is always the possibility to correctly guess when giving a random answer. The parameter \( \delta \) models the difficulty: when \( \eta \) is zero, the probability will be \( > 0.5 \) only if the ability of a team is greater than the item difficulty. The last parameter, \( \alpha \), gives a measure of how a small change in the ability is reflected by a change in the probability: it represents the discrimination of a question.

The pictures in Figure \[^{[1]}\] show the Item-Response curves for the same tasklet proposed to different categories. The value of \( \delta \) moves the curve on the horizontal axis: the quiz resulted more difficult for the Benjamins (it needed an ability \( > 0.39 \) to have positive odds), the other categories have a decreasing \( \delta \). The discrimination \( \alpha \) changes less and resulted higher for Students. The value of \( \eta \) reflects a multiple choice question with four alternatives (it is not exactly 0.25 since it is the product of stochastic fitting, see Sect. \[^{[2]}\] ). The dotted lines use a fixed \( \eta = 0 \) to show how the curve would change if a multiple choice quiz would be changed to an open one.

We aimed at fitting this IRT model with the data collected during the Kangourou contest, that is 11483 quizzes, given by 684 teams. We designed a hierarchical statistical model following \[^{[6]}\] and our regression analysis adopted a Bayesian approach: the data are fitted with respect to a probability model of all the unknown parameters and this model is then simulated and sampled with a Markov Chain Monte Carlo (MCMC) algorithm \[^{[8]}\], in order to get the posterior probability

\[ \text{"The merit of all things lies in their difficulty."} \]

Aramis refers to his new activity of writing a poem in verses of one syllable. “Add to the merit of the difficulty that of the brevity, and you are sure that your poem will at least have two merits.” notes d’Artagnan: Bebras tasklets share some merits with Aramis’ poem!
ability distributions of latent variables conditioned on the observed data.

Let $T$ and $Q$ be the set of teams and quizzes, and denote by $i$ and $j$ an item in $T$ and $Q$, respectively. Each (observed) answer $y_{nj}$ is modeled with a random variable drawn from a Bernoulli distribution with parameter $p_i$, dependent on $\delta_j$, $\eta_j$, $\alpha_j$ and computed according to \[ y_{nj} \sim \text{Ber}(p_i) \quad 1 \leq n \leq 11483. \]

The model needs also the a priori distributions of the parameters, for $j \in Q$ and $i \in T$:
\[
\begin{align*}
\theta_i & \sim N(0, s_\theta), \\
\delta_j & \sim N(m_{\delta_j}, s_\delta), \\
\log(\alpha_j) & \sim N(0, s_\alpha), \\
\eta_j & \sim \text{Beta}(1, c_j),
\end{align*}
\]

where $N$ and Beta denote respectively the Gaussian and Beta distribution, and $c_j$ is the number of choices for the multiple choice question $j$ or 1000 if the question is an “open” one. We did not want to make strong assumptions about the hyper-parameters of normal distributions: we used 0 as a reference point of ability, and 1 as a reference point of discrimination and we fixed their means accordingly. The hyper-parameters of normal distributions: we used 0 as one. We did not want to make strong assumptions about the actual maximum and minimum ones, and the average and standard deviation of scores. Finally, the column “Avg. time” shows the average time (out of 45 minutes) used by teams, measured in seconds.

### Table 1: A synthesis of the students’ performance in the 2014 edition of the Italian Bebras contest.

<table>
<thead>
<tr>
<th>Age group</th>
<th>Teams</th>
<th>Students</th>
<th>Tasks</th>
<th>Tot. score</th>
<th>Max.</th>
<th>Min.</th>
<th>Avg.</th>
<th>Std.</th>
<th>Avg. time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benjamin</td>
<td>145</td>
<td>580</td>
<td>16</td>
<td>62</td>
<td>49</td>
<td>11</td>
<td>25.70</td>
<td>8.23</td>
<td>2620</td>
</tr>
<tr>
<td>Cadet</td>
<td>207</td>
<td>828</td>
<td>17</td>
<td>68</td>
<td>57</td>
<td>4</td>
<td>29.30</td>
<td>10.45</td>
<td>2661</td>
</tr>
<tr>
<td>Junior</td>
<td>181</td>
<td>724</td>
<td>17</td>
<td>68</td>
<td>62</td>
<td>8</td>
<td>25.70</td>
<td>9.04</td>
<td>2689</td>
</tr>
<tr>
<td>Student</td>
<td>151</td>
<td>604</td>
<td>17</td>
<td>68</td>
<td>56.5</td>
<td>8</td>
<td>26.39</td>
<td>11.70</td>
<td>2694</td>
</tr>
</tbody>
</table>

### Figure 1: Item-Response curves of the same tasklet proposed to different categories.

3.1 Model implementation

We implemented the model with Stan \[11\], a probabilistic programming language for Bayesian statistical inference. The main parts of the program are in Listings 1 and 2: it is virtually a transposition with Stan syntax of the statistical model described above. However, we used a couple of

\[ \text{data} \{ \text{// observed data} \]
\[
\begin{align*}
\text{int} & \text{lower}_0, \text{upper}_Q \geq y[N]; \text{// results} \\
\text{int} & \text{lower}_1, \text{upper}_T \geq y[N]; \text{// team for y[n]} \\
\text{int} & \text{lower}_2, \text{upper}_Q; \text{// quiz for y[n]} \\
\text{real} & \text{lower}_3 = -1, \text{upper}_Q; \text{// m_step}[Q]; \\
\text{real} & \text{lower}_4 = 0, \text{upper}_Q; \text{// c}[Q];
\end{align*}
\]

\[
\text{parameters} \{ \text{// latent parameters} \\
\text{vector}[T] \text{theta_raw}; \text{// ability} \\
\text{vector}[Q] \text{delta_raw}; \text{// difficulty} \\
\text{vector}[Q] \text{alpha_raw}; \text{// discrimination} \\
\text{real} \text{lower}_5 = 0, \text{upper}_Q; \text{// eta}[Q]; \\
\text{real} \text{lower}_6 = 0, \text{upper}_Q; \text{// s_theta_unif}; \\
\text{real} \text{lower}_7 = 0, \text{upper}_Q; \text{// s_delta_unif}; \\
\text{real} \text{lower}_8 = 0, \text{upper}_Q; \text{// s_alpha_unif};
\]

\[
\text{transformed parameters} \{ \text{// computed from params and data} \\
\text{vector}[T] \text{theta}; \text{// ability} \\
\text{vector}[Q] \text{delta}; \text{// difficulty} \\
\text{vector}[Q] \text{alpha}; \text{// discrimination} \\
\text{real} \text{lower}_9 = 0, \text{upper}_Q; \text{// s_theta}; \\
\text{real} \text{lower}_10 = 0, \text{upper}_Q; \text{// s_delta}; \\
\text{real} \text{lower}_11 = 0, \text{upper}_Q; \text{// s_alpha};
\]

\[
\text{vector}[Q] \text{m_delta}; \\
\text{// reparameterization (see Stan manual, chapter 19)} \\
\text{// faster than s_delta \sim \text{cauchy}(0, 5);} \\
\text{// faster than theta = normal(0, s_theta);} \\
\text{// faster than delta \sim \text{normal}(m_{\delta}, s_{\delta});} \\
\text{// faster than theta \sim \text{normal}(s_{\theta}, s_{\theta});}
\]

\[
\text{for (j in 1:Q)} \{ \\
\text{// faster than m_{\delta} \sim \text{cauchy}(m_{\step}, 5);} \\
\text{m_{\delta}[j] \leftarrow m_{\step}[j] + 0.5 \times \tan(s_{\delta})) \}
\]

Listing 1: Stan program (minor parts omitted) implementing our model of task difficulty: data and parameters.
Listing 2: Stan program (minor parts omitted) implementing our model of task difficulty: model and generated quantities.

reparameterizations of Cauchy and Gaussian distributions, as suggested in Ch. 9 of [12]: this reduced the computation time from about 25 hours to 3, for a session with 10000 iterations. 8

4. RESULTS

Running a Stan program produces a sequence of samples for all the modeled parameters and the other generated quantities. The theory behind MCMC algorithms guarantees that, as the number of iterations approaches infinity, the samples are derived from the true posterior distributions of interest. No universal threshold to convergence exists across all problems: for convergence diagnostics Stan provides the Gelman-Rubin statistic \( R \). The basic idea is to use multiple independent chains to check for lack of convergence, assuming that if they have converged, by definition they should appear very similar to one another; at convergence \( R = 1 \).

We obtained stable results (that is, \( R = 1 \) for all parameters in our model) with four chains with 10000 iterations each. The first 5000 samples were used for warming up the algorithm, and the remaining samples were thinned discarding every second one. In total we got 10000 samples that we used to draw the posterior distribution of the parameters. Stan also provides a measure of the “effective sample size” (i.e., the number \( N_{\text{eff}} \) of independent samples with the same estimation power as the \( N \) autocorrelated sample) for each parameter: this can be used to estimate the standard error as standard deviation/\( \sqrt{N_{\text{eff}}} \).

The parameter we wanted to estimate is the difference between the average ability \( \bar{\theta} \) of a category and the difficulty \( \delta \) of a tasklet. The samples resulting from the Stan model give the posterior distribution of all the \( \theta \) and \( \delta \) of a tasklet. The values of the differences are collected in Table 4. We fitted two slightly different models: a “big” one in which we considered a \( \theta \) for each team (thus \( \tilde{\theta} \) is the average across all the teams of a category), and a “small” one in which all the teams of a category were aggregated (thus the observed answer of a tasklet has the multiplicity of the number of teams in a given category). We classified a tasklet as hard or easy when the absolute value of the difference was greater than 0.5. As Table 3 shows, the results of the classification are highly consistent in the two models; in 14 cases out of 67, however, they differ: the data do not support clearly the same classification in the two models. Interestingly, in 25 cases out of 67 (37%), the classification does not correspond to the authors’ one: as remarked in Sect. 1 and also shown by previous research [3], it is not easy to estimate the difficulty upfront.

4.1 Model checking

We checked how our results are correlated with rough measures of failures and scores. The correlations for the “big” model, shown in Figure 3 are good, especially if (second row of the picture) four outliers are ignored. Partially surprisingly, instead, is the fact that the time spent in tasklet is almost uncorrelated with the difficulty measured by the model. The corresponding values of correlations (without outliers) for the “small” model are 0.87, 0.87, 0.02.

4.2 Discussion

The analysis points out some facts. Very few tasklets (Benjamin 2014–JP–05a, Cadet 2014–CH–02, Cadet and Junior 2014–PL–07+, Cadet and Junior 2014–FR–01) were perceived easier than expected by authors. In three cases out of four, the problem underlying the tasklet is not trivial (sorting networks, generation of a sentence by a grammar, topological sorting) but the tasklets themselves refer to rather simple instances.

However — when the perceived difficulty is different — it is in most cases greater than the expected one. This often happened for tasklets proposed in the original Bebras form but with rescaled difficulty (Benjamin 2014–BE–16b, Benjamin 2014–DE–04, Benjamin 2014–SE–04, Benjamin and Cadet 2014–SK–07), but also for modified ones (Benjamin 2014–AU–03a, Cadet and Junior 2014–CH–07, Student 2011–DE–09, Student 2014–CA–01, Student 2014–CH–06, Student 2014–RU–06). In particular, replacing a set of few alternatives with an open question often results in a problem with too many issues to be considered; for instance, this happens when the correct solution is counterintuitive (as in 2014–RU–06) or one has to find the good way to approach the problem (as in 2014–SE–04): in fact, devising the right approach is much harder than detecting the right answer when it is listed. For a few tasklets (2014–CA–07, 2014–DE–04, 2014–SE–04, 2014–AU–03a), some discontinuity in the level of perceived difficulty appears. Indeed each of them proposes a minor obstacle that appears to hinder only the youngest categories: understanding the problem in 2014–CA–07 is not difficult per se, but the execution is a bit long and error-prone; in 2014–DE–04 three answers (out of four) can be excluded, but the correct one can only be obtained assuming some hypothesis that is actually omitted in the text, thus a doubt may arise; the solution of 2014–SE–04 is immediate only after one detects a special property which is only implicit in the text of the tasklet; the correct answer for 2014–AU–03a is counterintuitive: to answer

8 All the Stan programs, data, and further graphics are available at: https://bitbucket.org/monga/bebrastan

9 This tasklet resulted also much more difficult than expected for Benjamins.
Figure 2: Examples of posterior distributions of $\delta_j$ (in red), $\tilde{\theta}_k$ (in blue), and $(\tilde{\theta}_k - \delta)$ (in green).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-CZ-03a</td>
<td>e</td>
<td>(2.2 ± 0.0; 2.2 ± 0.0)</td>
<td>m</td>
<td>(1.1 ± 0.0; 0.7 ± 0.0)</td>
<td>m</td>
<td>(1.4 ± 0.0; 2.9 ± 0.0)</td>
<td></td>
</tr>
<tr>
<td>2014-CZ-06</td>
<td>e</td>
<td>(0.9 ± 0.0; 1.2 ± 0.0)</td>
<td>e</td>
<td>(13 ± 0.0; 0.9 ± 0.0)</td>
<td>e</td>
<td>(3.2 ± 0.0; 2.0 ± 0.0)</td>
<td></td>
</tr>
<tr>
<td>2014-DE-16</td>
<td>e</td>
<td>(3.4 ± 0.0; 7.9 ± 0.0)</td>
<td>c</td>
<td>(2.1 ± 0.0; 1.0 ± 0.0)</td>
<td>c</td>
<td>(0.9 ± 0.0; 0.3 ± 0.0)</td>
<td></td>
</tr>
<tr>
<td>2014-DA-24</td>
<td>e</td>
<td>(0.7 ± 0.0; 0.1 ± 0.0)</td>
<td>e</td>
<td>(2.1 ± 0.0; 0.2 ± 0.0)</td>
<td>c</td>
<td>(1.1 ± 0.0; 0.3 ± 0.0)</td>
<td></td>
</tr>
<tr>
<td>2014-DE-30</td>
<td>e</td>
<td>(0.9 ± 0.0; 0.2 ± 0.0)</td>
<td>e</td>
<td>(13 ± 0.0; 0.9 ± 0.0)</td>
<td>c</td>
<td>(0.5 ± 0.0; 0.1 ± 0.0)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Difficulty classification of tasklets (easy, medium, or hard): for each category the first column is the difficulty given by the authors (a.d.), the second (perceived difficulty) the output of the big and small models (in bold when it differs from authors’ one). For each classification the value of $\tilde{\theta}_k - \delta_j$ is given, within Monte Carlo Standard Error. Values in italics are outliers. The mods column indicates if the tasklet was modified w.r.t. the Bebras original (see Sect. 1).

Figure 3: Correlations between $\delta$ and failures, average score, and total time spent in a tasklet. In the second row outliers are ignored.
correctly one needs to build the right algorithm (the original, multiple choice, question would probably result much easier). Some tasklets resulted generally hard, with no significant differences among the categories, but were classified by the authors with different difficulty levels (2011-CH-06, 2014-CH-07, Cruc, 2011-DE-09, 2014-CA-01, 2014-CH-06, 2014-RU-06). Most of them were significantly changed from the original form, and hence admitted partial scores. However, in the analysis presented here we considered only full-score answers as correct.

5. CONCLUSIONS
Assessing the difficulty of tasks in an informatics contest aiming at spreading the discipline has a critical importance in order to avoid letting participants perceive the contest as too difficult or too easy, and thus not appealing. As several studies have pointed out, such estimation is not easy to carry out. Thus it is advisable to reconsider a proposed set of tasks after the competition ends, using the attained scores in order to infer the difficulty actually perceived by participants, compare it with the difficulty initially estimated by those who proposed the same tasks and gain knowledge in order to tune the future contest editions. This paper presents the results of an analysis aimed at inferring the perceived difficulty of tasks proposed in the 2014 edition of the Italian Bebras/Kangourou competition, based on the scores of more than 2000 participants. Such analysis, exploiting the methodology of IRT, highlights a substantial match between planned and perceived difficulty, meanwhile also emphasizing that in roughly one third of the cases the tasks were either easier or more difficult than expected. In the future we plan to further refine the analysis, also in view of providing an automated version of the workflow to be executed shortly after the contest ends but before the process of checking answers and ranking participants. However we can already infer some suggestions to better estimate the difficulty of tasklets. First, rescaling the difficulty proposed by the Bebras community according to the greater ability of team w.r.t. an individual is in general a good idea, but that should be done carefully because this criterion is not always applicable: not for all tasklets it is possible to work in parallel, thus the time needed may not decrease; replacing a set of few alternatives with an open question may result in a problem with too many issues to be considered. Second, assigning the same tasklet to a different category with varied level of difficulty is in general fair, except in the cases of tasklets just requiring special competences usually reached at a certain age (this is related very often to a high value of the $\alpha$ parameter). The teams of a younger category will not be able to solve it despite the team effort, and for the teams of an older category they are too easy. Last, many tasks appear to be too hard for every category, but often they called for partial scores that were not considered in this analysis. A refinement of this hypothesis could give more insights also on this issue.

6. ACKNOWLEDGMENTS
We would like to thank all the Bebras community for the great effort spent in producing exciting tasklets. Their Italian version saw also the fundamental contribution of Lorenzo Repetto, Fabrizio Carlinati, Angelo Lissoni, and all the staff of the Italian Kangourou.

7. REFERENCES