Investigating the Psychometric Structure of Bebras Contest
Towards measuring Computational Thinking skills

Peter Hubwieser
TUM School of Education
Technische Universität München
München, Germany
Peter.hubwieser@tum.de

Andreas Mühling
Department of Informatics
Technische Universität München
München, Germany
Andreas.Muehling@tum.de

Abstract—The “Bebras” is an international large scale contest that aims to motivate children for engaging in Computer Science. In this paper, we present an investigation of the psychometric factors that determine the success in this contest, e.g., abilities in Computational Thinking or intelligence factors. For this purpose, we looked for subsets of Bebras questions that are requiring joint psychometric constructs, according to the responses of the test persons. To this end, we applied latent trait analysis on all combinations of questions of a certain test. After identifying suitable sets of questions, we tested the fit of the mono-parametric Rasch Model and evaluated the distribution of person parameters. Additionally, the identified sets of questions were analyzed by qualitative methods to find out the nature of these common psychometric constructs. As a test bed for feasibility studies, we have chosen the Bebras Contests of Germany 2009. We have found and analyzed several sets of questions that met our requirements, analyzed the questions, and identified differences in the performance of boys and girls in these sets.

Keywords—contest, psychometric construct, latent trait analysis

I. INTRODUCTION

According to the founders, the Bebras Contest aims to interest children and young adolescents in typical problems of computer science that do not require any prerequisite knowledge [1]. Given that the contest prevails, successful participation might influence the career choice towards computer science. Thus, we have to ask ourselves, which factors might determine this success in order to attract children that really have the potential to excel in our field. Such factors might be for instance prerequisite knowledge (even though this is not required officially), abilities in Computational Thinking [2], or certain intelligence factors.

The annual Bebras online contest (www.bebras.org) was founded by V. Dagiene, see [3], who named it “Bebras” according to the Lithuanian word for (busy) “Beaver”. The contest had started 2004 in Lithuania with 3470 participants and has grown up to 523,319 participants in 21 countries in 2013. This represents a very large scale similar to the PISA studies of the OECD (www.oecd.org/pisa). The German issue is called “Informatik-Biber”. It is performed in all German states and in all types of secondary schools and has the largest number of participants (206,430 in 2013) of all participating countries, followed by France (171,932 in 2013).

To investigate the psychometric structure of the contest, we applied explorative statistical means to find subsets among the Bebras questions of a certain contest that seemed to require joint psychometric constructs, according to the responses of the test persons. We have developed a specific methodology for this goal by applying latent trait analysis on all combinations of questions of a certain test. As an indicator, we investigated, if the observed responses on these combinations could be predicted satisfyingly by the unidimensional Rasch Model (RM1). After identifying suitable sets of questions, we tested the fit of the mono-parametric Rasch Model more closely by comparing it with the fit of other models with more parameters of factors, e.g., the Birnbaum Model [4]. If the fit of RM1 was acceptable, we could hypothesize that these questions might in fact measure a joint (still unknown) psychometric construct. In this case, we regarded those subsets as homogenous and evaluated the distribution of the estimated person parameters. Finally, the identified sets of questions are analyzed by qualitative methods in order to find out the nature of the common psychometric constructs. As a test bed for this feasibility study, we have chosen the German Bebras Contest of 2009. We have found and analyzed several sets of questions in one of the four age groups that met our requirements, performed qualitative text analysis on these questions and identified differences between the performance of boys and girls and between single and team participants.

© IEEE, 2015. This is the author’s version of the work. It is posted here by permission of IEEE for your personal use. Not for redistribution. The definitive version was published in 3rd International Conference on Learning and Teaching in Computing and Engineering, Taipei, April 9-12. Los Alamitos, IEEE Press (pp. 62–69)

http://ieeexplore.ieee.org/.../number=7126233
Some aspects of this methodology and some results have been described already in [5] and [6].

I. BACKGROUND AND RELATED WORK

A. Item Response Theory

According to Classical Test Theory, the construct of interest (e.g., a certain student’s ability) is considered to be measured directly by item scores, although the results are considered to be error-prone. It is obvious that this straightforward approach is not suitable for measuring such complex constructs as competencies. In contrast, the Item Response Theory (IRT) assumes that the psychometric construct of interest, e.g., the student’s competency, is latent and can’t be measured directly [7]. Yet, the probability of correct answers is considered to depend on the manifestation of this construct in a certain way:

\[ P(X_{ik} = 1|\theta_i, \beta_k) = f(\theta_i, \beta_k) \]  (1)

where \( \theta_i \) is the parameter of person \( i \), representing the manifestation of the psychometric construct, \( \beta_k \) the parameter of Item \( k \), representing its difficulty, and \( f(\theta_i, \beta_k) \) a function that is determined by the psychometric model, for example the Rasch Model (RM), see below, that is assumed to fit the observations. In most cases these parameters have to be estimated by numeric calculations.

Usually, suitable models are preselected, driven by theory. This is followed by numerical investigations of the goodness-of-fit of the proposed models. Depending on the assumptions about the structure of the psychometric constructs that are to be measured, several different models may be considered, e.g., unidimensional models that cover only one single latent variable (competency) or alternatively multidimensional models. One of the simplest and most widely used ones is the basic unidimensional RM (RM1) [8]:

\[ P(X_{ik} = 1|\theta_i, \beta_k) = \frac{\exp(\theta_i - \beta_k)}{1 + \exp(\theta_i - \beta_k)} \]  (2)

The graph of this function is called Item Characteristic Curve (ICC) and looks as displayed in figure 3.

Provided that this model is applicable, some very convenient simplifications can be made. For example, the sum over the scores of all individual items represents a sufficient statistic, which means that the (estimated) person parameter depends only on the total number of correct answers of this person. It does not matter, which questions the person had responded to correctly. Yet, this model is applicable only if the ICCs have (at least nearly) the same slope. This slope is represented by an additional Discrimination Parameter \( \delta_k \) in the two-parametric Birnbaum Model (BM) [4]:

\[ P(X_{ik} = 1|\theta_i, \beta_k, \delta_k) = \frac{\exp(\delta_k(\theta_i - \beta_k))}{1 + \exp(\delta_k(\theta_i - \beta_k))} \]  (3)

In the case that there is more than one psychometric construct to be measured, the multidimensional RM (RMm) must be used [9]:

\[ P(X_{ik} = 1) = \frac{\sum_{f=1}^{M} \exp(q_{if}\theta_i - \beta_k)}{1 + \sum_{f=1}^{M} \exp(q_{if}\theta_i - \beta_k)} \]  (4)

In this case the probability of getting item \( i \) right is determined by the sum of all abilities \( \theta_i \) that are required for this task [10]. At the end of our research project, we aim to describe the total performance in the Bebras contest by several factors that will be integrated in one multidimensional RM.

Generally, three preconditions have to be met for the application of any RM:

- Homogeneity of items: All items are measuring the same psychometric construct (respectively a well-defined combination of constructs in the case of the multidimensional RM).
- Local stochastic independence: The underlying psychometric construct is the only coupling factor between items.
- Specific objectivity: For all samples from the population, the item parameters are independent from the specific sample; the same holds for all samples of questions and person parameters.

B. Computational Thinking

Regarding the relevant psychometric constructs that might be tested by the Bebras questions, Computational Thinking (CT) skills should be considered. This concept seems to be broadly accepted - even outside the CS community, see e.g. [11]. In [2] Wing defines CT as follows: “Computational Thinking focuses on the process of abstraction, choosing the right abstractions, operating in terms of multiple layers of abstraction simultaneously, defining the relationships between layers.” Recently a comprehensive review of the field of Computational Thinking in K–12 was published in [12].

Yet, for defining or identifying competencies in the field of CT, we need an operationalization of this definition. A proposal for this was recently presented by the CSTA [13]. Unfortunately its abstraction level is still too high for our purposes. Luckily, the CSTA-K12-Standards from 2011 [14] contain (among four others), a strand called “Computational Thinking”. The 48 standards of this strand can be seen as a more detailed operationalization of CT. In Table I we list several of those standards that might be relevant for the questions of the Bebras contest.
TABLE I. RELEVANT CSTA STANDARDS OF STRAND CT [14]

<table>
<thead>
<tr>
<th>Nr/ grades</th>
<th>The students shall be able to</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/3-6</td>
<td>Understand and use the basic steps in algorithmic problem-solving (e.g., problem statement and exploration, examination of sample instances, design, implementation, and testing).</td>
</tr>
<tr>
<td>2/3-6</td>
<td>Develop a simple understanding of an algorithm (e.g., search, sequence of events, or sorting) using computer-free exercises.</td>
</tr>
<tr>
<td>1/6-9</td>
<td>Use the basic steps in algorithmic problem solving to design solutions (e.g., problem statement and exploration, examination of sample instances, design, implementing a solution, testing, evaluation).</td>
</tr>
<tr>
<td>4/6-9</td>
<td>Evaluate ways that different algorithms may be used to solve the same problem.</td>
</tr>
<tr>
<td>6/6-9</td>
<td>Describe and analyze a sequence of instructions being followed (e.g., describe a character's behavior in a video game as driven by rules and algorithms).</td>
</tr>
<tr>
<td>8/6-9</td>
<td>Use visual representations of problem states, structures, and data (e.g., graphs, charts, network diagrams, flowcharts).</td>
</tr>
</tbody>
</table>

II. THE DATA

In Germany, the contest for each age group (see Table I) comprises 18 multiple choice questions in three levels of difficulty. The difficulty of the questions is assessed by the board of stakeholders that selects the questions for the national contests. The students have the choice to take the test alone or together with a partner. The test is performed online. In each age group a different set of questions - in total 18 - has to be answered, out of a pool of 39. Yet, some of the 39 different questions of the contest were presented to several age groups. One question (Nr. 172) was posed to all 4 age groups, 6 questions to 3 different groups and further 18 questions to 2 groups. The remaining 14 questions were restricted to one age group only. If the same question is posed to more than one group, a different degree of difficulty can be applied for each group.

TABLE II. AGE GROUPS OF CONTEST 33 IN 2009

<table>
<thead>
<tr>
<th>Group</th>
<th>Grades</th>
<th>Age approx.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG1</td>
<td>5-6</td>
<td>10-12</td>
</tr>
<tr>
<td>AG2</td>
<td>7-8</td>
<td>12-14</td>
</tr>
<tr>
<td>AG3</td>
<td>9-10</td>
<td>14-16</td>
</tr>
<tr>
<td>AG4</td>
<td>11-13</td>
<td>16-19</td>
</tr>
</tbody>
</table>

The relational data base was composed of 18 tables. By use of SQL statements, the raw data was read, verified and put into the format that was required for the further analysis. Basically two types of tables were produced: result-tables (one for each age group) for the responses of the participants to the questions and one participant-table for the personal attributes of the participants, e.g. gender, grade or school type. As we considered this to be relevant for the psychometrical structure, we had to distinguish between persons who worked alone and those who worked in pairs. We found 55088 singles and 13972 pairs among the participants. All computation was done in GNU R.

In a first step, we produced the pattern-tables from the result-tables, having 18 columns, each representing one question, and one row for each participant. The original score values of the questions cover a range from -3 to +12, depending on the difficulty as assessed by the authors of the questions (see table III).

Additionally, we had access to the time stamps for opening and closing each question by the participants. Missing time stamps means that the question was not opened at all by the participant.

TABLE III. POSSIBLE SCORES IN THE ORIGINAL BEBRAS DATA.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Easy</th>
<th>Medium</th>
<th>Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct answer</td>
<td>6</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>No answer</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Incorrect answer</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
</tr>
</tbody>
</table>

For our purpose, it made no sense to use the original score values of Bebras (see table III) for our evaluation, because the difficulty of the questions is just estimated by their authors. In fact it turned out later that the authors had failed in the assessment of 33% of the questions, assuming that Medium would cover the range between 0.33 and 0.66 of descriptive item difficulty, represented by the proportion of correct solutions. Hence, we opted to transform the values to a dichotomous scale (of value 0 or 1). For this, several different transformations are possible. In all of those, the correct answers are represented by 1 and the incorrect ones by 0. In the case that no answer was given (represented by a 0 originally), we chose to delete all participant rows with any “no answer” value. Additionally, we deleted all participant rows that did contain any non-valid entry for one of the 18 responses of the respective age group, due to missing data sets in the original table of results. Nevertheless, due to the large scale of the contest, a quite satisfying numbers of 38,873 participants remained over the four age groups: 8,221 in AG1, 15,547 in AG2, 11,672 in AG3, and 3,433 in AG4.

III. METHODOLOGY

In this section, we will provide an exemplary walk-through of our evaluation process of the Bebras contest. The process consists of the following steps, which are explained in more detail below:

- Identify sets of questions that seem to measure a common psychological trait by exploring all possible subsets of questions.
- Compare the fit of other possible psychometric models (with more parameters or dimensions) to the data of the sets of questions identified in step 1.
- Using the best fitting model, analyze the student performance with the help of the model parameters.
Perform qualitative data analysis on the Bebras questions.

To detect which skills, knowledge, abilities, attitudes, or other personality traits would be required to respond correctly to the Bebras questions, the most important question is whether or not there are subsets of questions that are measuring joint (combinations of) psychometric constructs and - if so - which construct(s) this might be. We will call such sets of questions homogenous from now on. Yet, we have to keep in mind that each of these constructs can itself be composed of one or more components.

Traditionally, classical explorative factor analysis is applied for the purpose of detecting subsets of questions that measure joint personal abilities. Yet, as our score format is dichotomous, this is not applicable. The reasons are explained in detail in [15]. Additionally, we were looking for a method that is better fitting to the IRT principles. Thus, we chose the methodology of latent trait analysis (LTA) as presented in Chapter 8 of [15].

A. Latent trait analysis (LTA)

According to this methodology, it is assumed that the responses of the students to a given set of questions can be described by a certain psychometric model, for example by the monofactorial RM (RM1) as explained in section II.A. Under this assumption, one can estimate the person and item parameters for all k and i from the results of the contest. Based on this estimation, by calculating the probability P in equation (1) of section II.A, the expected number of occurrences $E(r)$ of all possible response patterns r can be calculated. For p dichotomous questions, we have $2^p$ response patterns (i.e. combinations of 0s and 1s with the length p). The expected frequencies are then compared to the actually occurring pattern frequencies $O(r)$. Based on the differences, two different test statistics are calculated that describe the deviation of the expected and the measured values. First, we applied the log-likelihood test statistic $G^2$ (5) and second, a $X^2$ statistic (6):

$$G^2 = 2 \sum_{r=1}^{2^p} O(r) \ln \frac{O(r)}{E(r)}$$

$$X^2 = \sum_{r=1}^{2^p} \frac{(O(r) - E(r))^2}{E(r)}$$

As both statistics are approximately $\chi^2$ distributed, we can estimate the goodness of fit of RM1 with df degrees of freedom, where

$$df = 2^p - p(q + 1) - 1$$

As a precondition for this calculation, there has to be a sufficient number of datasets. According to [15], it has to be large enough to ensure that the frequency of each pattern has an expectation value of more than 5. In the case of 6 questions for example, this results in a minimum of 320 data sets. For testing all 18 questions of an age group, we would need more than 1,310,720 participants.

Unfortunately, this method is confirmatory in nature and therefore requires an a priori defined set of questions to be tested. We solved this problem by a brute force approach: calculating both statistics $G^2$ and $X^2$ for all possible combinations of $p = 3, 4, 5,$ and 6 out of the total set of 18 questions per age group. Following this, we selected those combinations where RM1 has shown a sufficiently good prediction of the observed results. More precisely, we have selected all combinations of p questions where both $G^2$ and $X^2$ did not exceed the $\chi^2$ limits for the respecting values of df (see equation 8 and table IV).

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\chi^2$-Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3.8</td>
</tr>
<tr>
<td>4</td>
<td>14.1</td>
</tr>
<tr>
<td>5</td>
<td>32.7</td>
</tr>
<tr>
<td>6</td>
<td>68.7</td>
</tr>
</tbody>
</table>

It turned out that a lot of 3-question combinations (more than 30), many 4-question combinations (10-20), only a few (0-4) 5-question combinations and no 6-question combinations met the requirements of this Likelihood analysis. Driven by the goal to find preferably large combinations, we decided to focus on the 5-question combinations from this point on. In AG1, we found 3 combinations (see table V), in AG2 four, in AG3 none and in AG4 three.

<table>
<thead>
<tr>
<th>Comb.</th>
<th>Questions</th>
<th>$G^2$</th>
<th>$X^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG1-1</td>
<td>156 162 164 184 187</td>
<td>16.33</td>
<td>16.47</td>
</tr>
<tr>
<td>AG1-2</td>
<td>156 164 184 187 189</td>
<td>30.72</td>
<td>30.68</td>
</tr>
<tr>
<td>AG1-3</td>
<td>156 164 184 187 194</td>
<td>28.60</td>
<td>29.09</td>
</tr>
</tbody>
</table>

For those combinations, we have calculated Spearman’s rank correlation of the scores of the combination and the total score of each participant as a measure for the predictive quality of the regarded subset of questions. The latter was calculated as the sum over the dichotomous scores of all questions, ignoring the weights according to the estimated question difficulties. This correlation turned out to vary between 0.57 and 0.78.

To visualize the psychometric structure of the questions of this contest over all age groups, we aggregated a graph of all those questions that belonged to any of the selected homogenous sets. An edge between two questions is drawn if those two questions belong to the same homogenous set. Hence, the edges show which pairs of questions might measure the same construct. The resulting graph for AG1 is displayed in Figure 1.

Regarding the result for all 33 questions of the total contest over all 4 age groups, it turned out that 17 of the 33 questions were quite closely connected - see Figure 2. Thus we can
hypothesize that these questions are measuring more or less one common psychometric construct or at least the same combination of constructs.

A. Rasch Model Tests

Although our LTA already suggests that the mono-parametric and mono-factorial RM will fit quite well on our data, there still remain some uncertainties. Most important, LTA is focused solely on the item difficulties and parameters, neglecting the distribution of person parameters, as demanded by the precondition of specific objectivity (see section II.A). And although we have apparently found a good model, there may be an even better fitting one (e.g. BM). Therefore, we have performed a set of standard tests for the fit of the RM, which are presented in the following for one exemplary combination of questions selected from AG1.

![Psychometric combinations of the questions of AG1](image1)

First, we applied different latent trait models on the pattern matrix, using the packages ltm and eRm in Gnu R: We applied BM (see equation (3) in section II.A), the RM with 2 factors (RM2), and another model with two factors and interaction parameter. In consequence, we were able to use all parameter values that were estimated under all these models for the following calculations. Next, we calculated the standard deviation of the difficulty parameters and the discrimination parameters (slope) applying BM. In order to represent a good set of Rasch test items, the former would have to be large, allowing to measure the person parameters over a large scale, while the latter would have to be low, avoiding cross-overs of the Item Characteristic Curves (ICC), see figure 3. For the difficulty parameters, the standard deviation was found to be between 0.62 and 1.63, while for the discriminations parameters it turned out to be between 0.02 and 0.21. Finally, we performed an ANOVA comparison of all applied models, comparing the values for AIC, BIC and Log-Likelihood. The result was quite acceptable in all cases, indicating the RM1 model was not fitting substantially worse than BM or the two-factor models.

The following tests for specific objectivity (see section II.A) follow the joint assumption that the Likelihood of a well-fitting model should be nearly the same for any subgroup of participants (i.e. the model is independent of the particular set of participants that was chosen to estimate the model). To test this, the persons were split in subgroups according to different criteria: median (respectively mean), values of combination score, and gender. For the subgroups, a test-specific statistic, basically representing the Likelihood of this model given the estimated parameters, is calculated. Finally, the p-value for the hypothesis that the statistic would be equal for all subgroups is calculated. The hypothesis (and thus the model) is rejected if p < 0.05. We have applied three different tests, again using the eRm package in R:

- Likelihood-Ratio-Test according to Andersen [17] with the splitting criteria median (respectively mean), values of combination score (1, 2, 3, 4) and gender on the level of question combinations,
- Martin-Löf-Test [18] with the splitting criterion median (respectively mean) on the level of question combinations,
- Wald-Test [19] with the splitting criteria median (respectively mean) and gender on the level of single questions.

![Psychometric combinations among all 33 questions of the contest](image2)

**TABLE VI. P-VALUES OF THE APPLIED TESTS ON COMBINATION LEVEL**

<table>
<thead>
<tr>
<th>Group Comb</th>
<th>LR Median</th>
<th>LR Score</th>
<th>LR Gender</th>
<th>Martin-Löf</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG1-1</td>
<td>0.67</td>
<td>0.85</td>
<td>0.90</td>
<td>0.75</td>
</tr>
<tr>
<td>AG1-2</td>
<td>0.61</td>
<td>0.84</td>
<td>0.04</td>
<td>0.61</td>
</tr>
<tr>
<td>AG1-3</td>
<td>0.79</td>
<td>0.37</td>
<td>0.75</td>
<td>0.32</td>
</tr>
<tr>
<td>AG2-1</td>
<td>0.31</td>
<td>0.13</td>
<td>0.00</td>
<td>0.49</td>
</tr>
<tr>
<td>AG2-2</td>
<td>0.44</td>
<td>0.95</td>
<td>0.00</td>
<td>0.78</td>
</tr>
<tr>
<td>AG2-3</td>
<td>0.11</td>
<td>0.59</td>
<td>0.00</td>
<td>0.56</td>
</tr>
<tr>
<td>AG2-4</td>
<td>0.16</td>
<td>0.33</td>
<td>0.01</td>
<td>0.99</td>
</tr>
</tbody>
</table>
While the Martin-Löf-Test and the LR tests regarding median/mean and score were passed by all question combinations, only two combinations passed the LR-Test regarding gender. On the question level, the Wald test demonstrated the same problematic nature of the gender splitting, because all but two combinations (again AG1-1 and AG1-2) included questions that produced p-values below 0.05. According to the Wald test on median/mean, there were questions in the combinations AG2-1, AG2-3, and AG2-4 that would have to be excluded.

In summary, over all age groups only the two combinations AG1-1 and AG1-2 (of the originally 10) passed all tests without any problems. Thus, looking for a set of homogenous test questions, those would be the ones to consider. In the light of this, we will restrict our discussion in the following to AG1-1 and AG1-3.

Considering the ICCs of the two combinations, AG1-1 is clearly more suitable than AG1-3, due to its higher variation in difficulty (1.43 vs. 0.62) and its lower variation in discrimination (0.04 vs. 0.07). In Figure 3, the ICCs of AG1-1 of the RM1 model are displayed. Interestingly, both are very strongly correlated with the total score over all 18 questions (0.74 vs. 0.77). This means that both subsets have are very good predictors for the overall success in AG1.

Due to its quite acceptable homogeneity, we will conduct the exemplary student evaluation with AG1-1. In Figures 4-8, the questions are paraphrased in English (G: Given, Q: Question, German title in parentheses). The original test booklet is available at www.informatik-biber.de. The questions are ordered according to their empirical difficulty from easy to hard.

### B. Evaluation of Person Parameters

To illustrate our methodology, we have conducted an exemplary evaluation of the most homogenous and suitable question combination AG1-1. Figure 9 displays the distribution of the person parameters θ, (“ability”), which looks quite reasonable.

---

<table>
<thead>
<tr>
<th>AG4-1</th>
<th>0.06</th>
<th>0.49</th>
<th>0.02</th>
<th>0.93</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG4-2</td>
<td>0.65</td>
<td>0.67</td>
<td>0.00</td>
<td>0.70</td>
</tr>
<tr>
<td>AG4-3</td>
<td>0.94</td>
<td>0.99</td>
<td>0.00</td>
<td>0.93</td>
</tr>
</tbody>
</table>

---

G: Grid of crossroads; Formalization rule for the choices at each crossing: L (left), R (right), S (straight) Formal path L-R-L-S, ending at the school.

Q: Where was the starting point?

---

G: Clotheslines, tied to poles have been transformed to post-situations by an unknown rule. Pre-situations p1, p2, p3 are shown with assigned post-situations. One more pre-situation p4 without post-situation

Q: How many lines have to be added to p4 according to the same rule that had transformed p1, p2, p3?

---

G: Graph, represented by 3 stones and tree trunks, crossing a river.

Q: Which stone has to be passed by any way over the river?

---

G: Different patterns of squares

Q: Which pattern does not allow to build a square from?

---

### TABLE VII. DESCRIPTIVE DIFFICULTY OF AG1-1 AND AG1-3

<table>
<thead>
<tr>
<th>Nr.</th>
<th>X164</th>
<th>X187</th>
<th>X184</th>
<th>X156</th>
<th>X194</th>
<th>X189</th>
<th>X162</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD</td>
<td>0.81</td>
<td>0.68</td>
<td>0.6</td>
<td>0.57</td>
<td>0.51</td>
<td>0.44</td>
<td>0.13</td>
</tr>
</tbody>
</table>

*a DD = Descriptive difficulty*
G: Graph, representing a finite state machine. The input are the letters of a name, the final states (numbers) are the levels in a building where a person with this name lives. 
Q: Which is the level where Jan lives?

Fig. 8. Question 187 (Mehrinformatikerhaus)

<table>
<thead>
<tr>
<th>TABLE VIII. DIFFERENCES IN SCORE MEANS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total scores</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>All boys – all girls</td>
</tr>
<tr>
<td>Singles – pairs</td>
</tr>
<tr>
<td>Single boys – single girls</td>
</tr>
<tr>
<td>Male pairs – fem. pairs</td>
</tr>
</tbody>
</table>

Then we compared the mean scores of different groups of participants: singles and pairs, girls and boys (see Table VII). Considering the scale properties, the proper significance test for the differences is the 2-side approximate Gauss Test [20]. The theta-values of person parameters were normalized according to the PISA scale, which results per definition in a mean of 500 and a standard deviation of 100 points (over all participants).

Second, we have ranked the German federal states (“Länder”) according to the performance of their students in grade 5 and 6 (see Figure 10). Again, these differences seem quite notable compared to the range of PISA results in Mathematics. While Brandenburg (540) would be in second place, Hamburg (430) would be ranked last but two among all OECD countries. The explanation for these differences might be the fact that the participation strategies for Bebras are very different in the states. In some of them, nearly all classes take part during school time, while in others only students are involved that work at home on the questions deliberately.

Apparently, the boys showed a significant better performance compared to the girls. Also, the pairs performed better than the singles. As the difference in the mathematical competence between boys and girls in PISA 2012 was only 11 points in the OECD average [21], these results seem quite considerable. Interestingly, the difference between male and female pairs (there were no mixed pairs in AG1) was not significant. This could be an indicator that girls collaborate better than boys.

IV. QUALITATIVE ANALYSIS OF SELECTED QUESTIONS

Aiming to investigate the nature of the psychometric concepts, we have performed qualitative data analysis on the tasks. As the result of an inductive coding process according to the methodology of Mayring [22], we obtained a category system that comprised the following three dimensions (with subdimensions in parentheses):

- Indication (Given Information, Representation Form),
- Question (Required Information, Representation Form),
- Required Operations.

The analysis was performed using the tool MaxQDA (www.maxqda.org). As the cognitive processes connected with the questions turned out to be very complex and difficult to analyze, we restrict the discussion in this paper to the most homogenous set of questions AG1-1. Figure 11 shows the MaxQDA table of the categories that have been coded in AG1-1.

Fig. 9. Distribution of Person parameters in AG1-1.
V. DISCUSSION

First, we have to admit once again that the Bebras Contest is not a test (like PISA). This might be the most probable reason for the low number of question combinations that are measuring some homogenous psychometric constructs in a suitable way. Further, compared to PISA, the participation is nearly totally uncontrolled. Thus, the sampling might provide serious biases, because at least in some regions only classes of very interested and motivated teachers might participate. Regarding the work on the questions, it is not clear which assistance the students had, e.g. by the teacher or other peers.

Fig. 11. Coded categories of the question set AG1-1.

However, despite all these deficits compared to a proper large scale study, our methodology has produced some remarkable results. First, it is amazing that there is a coherence between 17 questions that are contained in any of the combinations selected by the LTA over all age groups, while none of the remaining 12 shows up in any of these combinations. This suggests that those 17 questions are measuring some joint constructs, while the remaining 12 don’t show any commonality. Yet, the quality of the 17 questions seems not high enough to represent good test items which is easy to explain considering the goals of the Bebras contest. Additionally, the results we have found when analyzing the most homogenous set of questions AG1-1 and student performance in this, seem very reasonable compared to the PISA results in Mathematics.

Apparently, four of the five questions are focused on graphs and pathways on these. The only exception is question X184, which requires arranging different checked patterns to squares. Despite this difference, it is very visually oriented as all other questions of AG1-1. Thus, the most apparent common requirement of these questions is the ability to operate on graphical represented patterns, to find pathways across networks or to arrange checkers in various spatial dimensions. Therefore, it might be that these 5 questions are just measuring something similar to the visual-spatial intelligence factor of Gardner [23].

Nevertheless, the accordance of AG1-1 with the CSTA standards of the strand CT is obvious. The required cognitive operations are referring to all but one of the standards listed in table I: 1/3-6, 2/3-6, 1/6-9, 6/6-9, 8/6-9. For instance, it seems apparent that all of them require more or less Nr. 8/6-9: “Use visual representations of problem states, structures, and data”. This might inspire the community to develop tests for CT based on Bebras questions. For this purpose, after a close analysis of the questions of all contests up to now, some suitable sets could be assembled to form test booklets that indeed could measure the CT standards of the CSTA.

VI. CONCLUSION AND FUTURE WORK

In conclusion, we have presented an exploratory analysis of the questions of the Bebras contest regarding the homogeneity of the measured competencies. While the results are promising, our main goal was to show a methodology in form of a specific process of evaluation. This can then be taken and applied to other sources of possible test-items with the hope of finding a theory-based and empirically proven competency test for CS, which could then be incorporated in international large-scale assessments.

We will apply this methodology to the remaining available sets of German Bebras data up to 2013. Hopefully, this will yield more homogenous sets of questions. Furthermore, as already mentioned above, one of our next steps has to be a qualitative analysis of the cognitive demands of the selected questions, e.g. which CSTA standards are tested by them. We have to find out,
what the selected questions have in common regarding this issue. Eventually, this will allow us to describe the psychometric constructs that we have in terms of Computational Thinking.

REFERENCES


[13] CSTA and ISTE, Operational Definition of Computational Thinking for K-12 Education. CSTA


