ABSTRACT
To test and adopt the methodology of future PISA surveys in Computer Science, we have utilized the Bebras Contest as a test field. We performed latent trait analysis of the questions of the German Bebras contest 2009 to find sets of questions that might measure joint psychometric constructs (competencies). After having identified several sets, we have tested the fit of the monoparametric Rasch Model and evaluated the distribution of person parameters.

Categories and Subject Descriptors
K.3.2 [Computers and Education]: Computer and Information Science Education–computer science education

General Terms
Measurement, Human Factors

Keywords
Competencies, item response theory, large scale studies, rasch model

1. INTRODUCTION
There is still a long way until CSE will be generally accepted as a compulsory school subject on the same level as mathematics, physics or foreign languages. One of the necessary preconditions for this acceptance is that research in CSE will reach the same methodological level compared to other subjects. Regarding mathematics, science or language understanding, this research level was substantially raised by the well-known international PISA (Programme for International Student Assessment) studies that have been conducted by the Organisation for Economic Co-operation and Development (OECD) since 2000. The scientific community had to acknowledge that the PISA studies follow a sophisticated, well-founded methodology [9], which had a groundbreaking impact on the whole field of empirical education research at least in the participating school subjects. In regard of this impact, we assume that a PISA survey of computer science competencies could advance the research methodologies of Computer Science Education in a pioneering way and might finally lift this field on the level of educational research in traditional school subjects. However, we have to realize that this would require considerable prerequisite work. Among others, we need to test and adopt the methodology of large scale investigations to fit the context of CSE. For this purpose, the methodology of Item Response Theory [7] has to be tested and validated in a sufficiently large scale. Hence, we had to look for a suitable test field. The annual Bebras online contest (www.bebras.org), founded by V. Dagiene [3], had started 2004 in Lithuania with 3470 participants and has grown to 523,319 participants in 21 countries in 2013, which represents a large scale very similar to PISA. The German issue of Bebras (Informatik-Biber, www.informatik-biber.de), performed in all German federal states and in all types of secondary schools, is the largest regarding the number of participants (206,430 in 2013). Whereas the Bebras Contest does not intend to be a test, we asked ourselves what the Bebras contest would measure if it were a test. For this, we took the Bebras questions of a certain contest as given and investigated by explorative statistical means, if there were subsets of questions among them that were measuring a common psychometric construct. In this case, we call those subsets homogenous. As an indicator, we investigated, if the observed responses on these subsets could be predicted satisficingly by a unidimensional Rasch Model [7] if there were such homogenous subsets, we could hypothesize that these questions might measure a common psychometric construct, which eventually might be a certain, unknown competency. In this case we could define this competency by this set of tasks according to [8]. We have named this methodology “competency mining”. For a detailed description, see [4].

2. THE EVALUATION
We started our evaluation on the data of German competition No 33, which was conducted in October 2009 in 4 age groups: AG1 (age 10-12), AG2 (12-14), AG3 (14-16), AG4 (16-18). The contest for each age group poses 18 multiple choice questions in three levels of difficulty, which determine the score values. The test is performed online. 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. We transformed the original score values to a dichotomous scale, by scoring correct answers with 1, incorrect with 0 and deleting all participants with any “no answer” value on any question. Additionally, we deleted all participants that worked in pairs, ending up with 55088 singles.

According to the Item Response Theory, it is assumed that the responses of the students to a certain set of questions can be described by a certain psychometric model, for example by the monofactorial Rasch Model [7] with one parameter. In this case, the probability of correct answers is considered to depend on the manifestation of this construct in the following way [7] with one parameter:

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

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. Under this assumption, one can estimate the person and item parameters for all \(k\) and \(i\) from the re-
solved this problem by a brute force approach: calculating requires an a priori defined set of questions to be tested. We compared the differences, a compared to the actually measured pattern frequencies. 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 meet 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, in AG2 four, in AG3 none and in AG4 three. Afterwards, we have performed a set of standard tests for the fit of the RM. First, we applied different latent trait models on the pattern matrix, using the packages ltm and eltm in Gnu R. We applied the several Rasch Models with 1-2 factors and 1-2 parameters.

Next, we calculated the standard deviation of the difficulty parameters and the discrimination parameters (slope) applying a 2-parameter model. 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). Next, 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 simplest model was not fitting substantially worse than the others.

Finally, we performed several tests for the fit of the Rasch Model: Likelihood-Ratio-Test according to Andersen [1] with the splitting criteria median (respectively mean), Martin-Löf-Test (see [5]) with the splitting criterion median (respectively mean) on the level of question combinations and the Wald-Test (see [10]) with the splitting criteria median (respectively mean) and gender on the level of single questions. 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 included questions that produced p-values below 0.05. According to the Wald test on median/mean, we found in every combinations one or more questions that would have to be excluded, except in the following combination AG1-1 (German title in parentheses, see www.informatik.biber.de):

- Question 156 (Finde den Startpunkt),
- Question 162 (Wäscheleinen im Garten),
- Question 164 (Biber am Fluss),
- Question 187 (Mehrinformatikerhaus),
- Question 184 (Versprengte Quadrate).

For this combination, we have calculated Spearman’s rank correlation of its score sums and the total score (over all 18 questions) of each participant as a measure for the predictive quality of this subset. This correlation turned out to be very strong (0.78).

Then we compared the mean scores of different groups of participants: singles and pairs, girls and boys (see table 1.). Considering the scale properties, the proper significance test for the differences is the 2-side approximate Gauss Test [2]. 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).

<table>
<thead>
<tr>
<th>Table 1. Differences in score means.</th>
</tr>
</thead>
<tbody>
<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>

*Significant difference for α = 0.05.

3. REFERENCES


