An Analysis of the Performance of Italian Schools in Bebras and in the National Student Assessment INVALSI

Carlo Bellettini
carlo.bellettini@unimi.it

Violetta Lonati
violetta.lonati@unimi.it

Mattia Monga
mattia.monga@unimi.it

Anna Morpurgo
anna.morpurgo@unimi.it

Università degli Studi di Milano
Dept. of Computer Science
Via Celoria 18, 20133 — Milan, Italy

Abstract

This paper analyzes the results of the Bebras Challenge on Informatics and Computational Thinking held in Italy in the last three years and it compares them to the overall performance of Italian schools in the national INVALSI assessment of the standardized levels reached by students in Italian, Mathematics, and English. The main research question is if the mean regional performance at INVALSI tests can predict the performance of schools of the same region in the Bebras challenge. The answer is positive at the grossest level: macro regional areas with INVALSI results below the national average tend to perform worse also in the Bebras challenge. At regional level, a high correlation between Bebras and INVALSI was found among the regions whose results differ significantly.

1 Introduction

The Bebras International Challenge on Informatics and Computational Thinking (`http://bebras.org`) is a yearly contest organized since 2004 [Dag10, HCD11]. In 2018 almost three million participants from 54 countries took part to one of the locally organized events. The contest, open to pupils of all school levels (from primary up to upper secondary), is based on tasks rooted on core informatics concepts and computational thinking, yet independent of specific previous knowledge such as for instance that acquired during curricular activities. In fact Bebras tasks avoid the use of jargon and are especially aimed at a non-vocational audience, focusing on that part of informatics that should become familiar to everyone, not just computing professionals. The tasks are supposed to provide an entertaining learning experience, and they are designed by the Bebras community to be moderately challenging and solvable in a relatively short time. The setting of the contest is slightly different in each country, but in general participants have to solve a set of about 10-15 tasks in an average time of three minutes for each. In Italy, the Bebras is open to teams of 3 or 4 pupils, divided in five age groups: I (grades 4–5, ages ≈9–10), II (grades 6–7, ages ≈11–12), III (grade 8, age ≈13), IV (grades 9–10, ages ≈14–15), V (grades 11–13, ages ≈16–18). In the last three editions we had 36,018 teams, from schools located in all the 20 administrative regions Italy is subdivided into (see Table 3). Besides being used during the contests, Bebras tasks are an opportunity for educational activities [DS16, LMM^{+}17, CAC^{+}18]. Moreover, Bebras was used to measure improvements of students' attitude to computational thinking [SBS17]. The study examined 21 schools (children aged 9–11) which participated in "Code Clubs". The primary outcome measure was a set of Bebras tasks, which 317

In: I. Fronza, C. Pahl (eds.): Proceedings of the 2nd Systems of Assessments for Computational Thinking Learning Workshop (TACKLE 2019), co-located with 14th European Conference on Technology Enhanced Learning (EC-TEL 2019), 17-09-2019, published at http://ceur-ws.org.

pupils completed at baseline and endpoint. We wonder, instead, if the performances in Bebras follow the general level of competencies of the schools participating to the contest. In order to answer this question, one should have a measure of the curricular achievements of the schools (or even the classes) involved, but unfortunately these data are not publicly available. In fact, one of the Bebras' goals is to spread the acquaintance with informatics and computational thinking among every school population, even (or maybe especially) those not naturally attracted by computing. To this end, we avoid any participation fee and we try to keep the competition at a level such that nobody should feel ashamed to participate: Bebras should be perceived as an opportunity to have fun and learn something, not to show off the performances of the schools. For example in Italy, although every teacher receives ranking data about their teams, only the very top of the ranking is published (the best eight teams in each age group, with at most one team per school). Thus, we do not want to ask teachers about the marks of their pupils in the curricular activities or other proxies of their academic success. Instead, we tried to understand if the results in the Italian Bebras contest were somewhat correlated with the general school performances in the same territory. For this, we resort to INVALSI data, the national student assessment program, similar to OECD's Programme for International Student Assessment (PISA) or IEA's Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS).

Since school year 2005/6, all the pupils of the Italian school system at the end of grade 2, 5, 8, and 10 are evaluated by an INVALSI standardized test, aimed at measuring their proficiency in Italian, Mathematics, English listening and English reading. According to the 2018 INVALSI report, the performances of the twenty Italian regions differ in a significant way, at least from grade 8 and up. Thus, we set up a study aimed at understanding if these differences are reflected in the results we see in the Bebras contest. The number of Bebras teams is much smaller than the number of students involved in the INVALSI assessment (even by considering that their public data are based on a sample, see below), moreover Bebras participation depends on teachers' interest, while INVALSI is mandatory. Nevertheless, we wanted to understand if Bebras data reflect the general geographic pattern of the wider population of Italian schools.

The paper is organized as follows: in Section 2 we formalize our research questions, in Section 3 we describe our approach, in Section 4 we report our analyses, and finally in Section 5 we draw some conclusions.

2 The research questions

The 2018 INVALSI assessment [INV18] tested 29,520 grade 5 classes (562,635 pupils), 29,032 grade 8 classes (574,506 pupils), and 26,361 grade 10 classes (543,296 pupils). In order to guarantee data quality, a sample of students was observed directly during the test: the data reported publicly is based on this direct analysis of 29,371 grade 5 students, 31,300 grade 8 students, and 48,664 grade 10 students. In grade 5 the test is paper based and manually marked, while in the other grades the test is computer based and automatically marked. Results are separately assessed for four areas of competence: 'Italian', 'Mathematics', 'English listening', and 'English reading' (in 2018, grade 10 was not tested for English). Public data cover all the twenty Italian regions (Trentino-Alto Adige is actually divided into two autonomous provinces, since in the region live communities with different mother-tongues, no aggregated regional data are provided). Results are provided at two levels of aggregations:

1. five geographic macro-areas: North-West, North-East, Center, South, South-Islands;
2. 21 administrative regions (19 regions and 2 autonomous provinces).

Table 1 shows the mean performance by area; the average is set at 200 (with a standard deviation of 40).

According to the INVALSI report [INV18], the differences among the areas at grade 5 are small[1]. Instead, the differences are considered increasingly significant in the higher grades. Overall they are claimed to match similar results in PISA assessment (surveyed internationally every three years) with the North part of the country performing better than the national average, and the South part worse than the national average; the Center instead reflects the national average. The report also mentions that the Northern part of the country has better than average results in recent TIMMS assessments.

The regional data are more detailed, since they report also the standard deviation of the distributions, not only the means. The data are shown in Table 2.

In this study, our goal is to understand if this variability is reflected in the results of the Italian Bebras. We have homogeneous data for the last three editions (2016, 2017, 2018). The total number of partecipating teams is reported in Table 3.

Bebras data involve a smaller number of schools with respect to INVALSI (which aims at being "universal" in the Italian school system: the participation

[1] Grade 2 has even smaller differences; it was not considered here, since the Italian Bebras involves pupils from grade 4 up to grade 13

area	*grade*	*Italian*	*Mathematics*	*English listening*	*English reading*
Center	5	204	204	207	205
North-East	5	202	203	203	204
North-West	5	203	202	203	203
South	5	195	197	192	194
South-Islands	5	192	191	192	191
Center	8	205	204	204	205
North-East	8	206	211	214	210
North-West	8	207	207	214	209
South	8	190	188	184	188
South-Islands	8	189	186	178	184
Center	10	200	201		
North-East	10	210	213		
North-West	10	210	212		
South	10	192	189		
South-Islands	10	185	182		

Table 1: INVALSI results by macro area. The standardized national mean is 200. English was not tested at grade 10.

is mandated by law. In the past it was also used to mark students at grade 8, but the 2018 edition was not used for this purpose). Nevertheless we would like to use them to try to answer the following research questions.

RQ1

Is there any correlation between the average ability of Bebras teams in a specific region and the regional performance in INVALSI tests?

RQ2

Is there any correlation between the average ability of Bebras teams in a geographic macro area and the area performance in INVALSI tests?

RQ3

Is the overall performance trend at INVALSI tests, with Northern schools performing better than the national average and Southern schools performing worse, reflected also in Bebras results?

3 Methodology

We estimated the ability of the Bebras teams by fitting an Item Response Theory (IRT) [HS85] model with two parameters. IRT is routinely used to evaluate massive educational assessment studies like OECD's PISA, and it has already been applied to Bebras and other informatics competitions [KVC06, HM14, BLM$^+$15]. Moreover, a similar IRT model is behind the INVALSI data as described in [Des18].

IRT models each solver with an *ability* (θ) 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* (δ) of a task and its *discrimination* (α). Difficulty locates the response function: if the ability of the solver is greater than the difficulty of a task, the probability of solving it 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 great impact on the probability of solving the task; a discrimination = 0 defines a task in which the ability of the solver does not matter at all. Figure 1 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. In order to be comparable with INVALSI data, we deviated from the common practice [GH06] of assuming that, overall, ability has mean = 0 with respect to an arbitrary reference point and standard deviation = 1. Instead, we assumed a mean ability = 200 and a standard deviation = 40.

In order to estimate the difficulty and discrimination of each task, we implemented the probabilistic model with Stan [Sta16]. Stan is a software tool which, given a statistical model, uses Hamiltonian Monte Carlo sampling (a very efficient form of Markov chain Monte Carlo sampling) to approximate the *posterior* probability of the parameters of interest.

$$P(\theta_i|Y) \quad i \in \text{teams} \tag{1}$$

where θ_i is the ability of team i. The statistical model sampled is a hierarchical one, with the following *prior* distributions:

region	area	grade	Italian	σ	Mathematics	σ	Eng. listening	σ	Eng. reading	σ
ABRUZZO	South	5	203	40	202	40	197	39	198	39
BASILICATA	South-Islands	5	204	39	211	40	201	40	202	41
CALABRIA	South-Islands	5	192	41	192	41	189	39	191	40
CAMPANIA	South	5	189	41	193	41	189	42	188	41
EMILIA-ROMAGNA	North-East	5	203	39	201	40	202	37	204	39
FRIULI-VENEZIA GIULIA	North-East	5	204	39	208	42	205	38	205	39
LAZIO	Center	5	202	40	201	38	207	41	204	40
LIGURIA	North-West	5	201	39	201	39	200	39	201	39
LOMBARDIA	North-West	5	204	39	202	40	205	40	205	39
MARCHE	Center	5	206	39	208	39	204	37	204	39
MOLISE	South	5	210	41	220	46	214	45	208	41
PIEMONTE	North-West	5	202	39	203	41	198	38	201	39
PUGLIA	South	5	202	41	202	40	195	38	200	39
SARDEGNA	South-Islands	5	194	40	188	37	187	37	194	39
SICILIA	South-Islands	5	190	40	189	39	193	43	189	43
TOSCANA	Center	5	207	39	207	39	208	39	207	40
TRENTINO-ALTO ADIGE[a]	North-East	5	205	38	208	39	223	41	211	39
UMBRIA	Center	5	206	38	207	40	210	38	206	38
VALLE D'AOSTA	North-West	5	203	37	198	38				
VENETO	North-East	5	202	37	203	38	202	35	203	37
ABRUZZO	South	8	201	38	200	38	198	38	199	38
BASILICATA	South-Islands	8	195	39	189	37	183	35	187	40
CALABRIA	South-Islands	8	185	40	181	36	170	41	177	41
CAMPANIA	South	8	185	42	183	38	179	39	183	42
EMILIA-ROMAGNA	North-East	8	207	40	211	41	215	35	210	38
FRIULI-VENEZIA GIULIA	North-East	8	208	35	213	39	219	33	214	35
LAZIO	Center	8	205	39	201	38	203	38	204	38
LIGURIA	North-West	8	205	38	204	37	210	35	207	37
LOMBARDIA	North-West	8	209	39	210	40	218	37	212	37
MARCHE	Center	8	208	38	209	40	210	33	208	36
MOLISE	South	8	202	38	202	40	194	36	197	39
PIEMONTE	North-West	8	202	39	203	39	206	35	203	38
PUGLIA	South	8	195	39	192	38	186	38	192	39
SARDEGNA	South-Islands	8	198	37	192	35	190	36	192	39
SICILIA	South-Islands	8	187	39	185	36	177	39	183	41
TOSCANA	Center	8	203	39	207	38	204	37	205	36
TRENTINO-ALTO ADIGE[a]	North-East	8	207	37	214	39	218	34	213	37
UMBRIA	Center	8	207	37	210	38	207	37	205	38
VALLE D'AOSTA	North-West	8	209	36	209	37	214	33	208	34
VENETO	North-East	8	205	37	211	40	211	33	209	35
ABRUZZO	South	10	199	39	200	40				
BASILICATA	South-Islands	10	196	38	196	37				
CALABRIA	South-Islands	10	181	42	176	35				
CAMPANIA	South	10	189	43	186	38				
EMILIA-ROMAGNA	North-East	10	207	38	210	40				
FRIULI-VENEZIA GIULIA	North-East	10	209	35	214	38				
LAZIO	Center	10	198	38	196	37				
LIGURIA	North-West	10	205	37	206	39				
LOMBARDIA	North-West	10	213	35	215	39				
MARCHE	Center	10	204	42	208	43				
MOLISE	South	10	194	42	195	40				
PIEMONTE	North-West	10	206	37	207	38				
PUGLIA	South	10	193	38	191	37				
SARDEGNA	South-Islands	10	183	44	178	34				
SICILIA	South-Islands	10	187	41	184	34				
TOSCANA	Center	10	200	38	203	39				
TRENTINO-ALTO ADIGE[a]	North-East	10	215	33	219	37				
UMBRIA	Center	10	205	39	207	42				
VALLE D'AOSTA	North-West	10	208	33	204	35				
VENETO	North-East	10	213	36	216	37				

[a] The data refer only to the autonomous province of Trento.

Table 2: INVALSI results by region. The standardized national mean is 200. English was not tested in grade 10 and data about English in Valle d'Aosta at grade 5 are not available. Trentino-Alto Adige is divided into two provinces and no aggregated datum is available.

area	*region*	*Grade 5*	*Grade 8*	*Grade 10*
Center	LAZIO	13,466	2,617	4,787
Center	MARCHE	1,572	1,746	1,976
Center	TOSCANA	3,034	2,470	1,342
Center	UMBRIA	1,771	234	713
	Total	**19,843**	**7,067**	**8,818**
North-East	EMILIA-ROMAGNA	4,079	4,117	4,891
North-East	FRIULI-VENEZIA GIULIA	708	1,357	4,588
North-East	TRENTINO-ALTO ADIGE	60	1,774	713
North-East	VENETO	11,197	7,403	9,623
	Total	**16,044**	**14,651**	**19,815**
North-West	LIGURIA	2,118	1,485	6,373
North-West	LOMBARDIA	30,416	14,125	13,878
North-West	PIEMONTE	6,235	4,814	3,167
North-West	VALLE D'AOSTA	672	954	0
	Total	**39,441**	**21,378**	**23,418**
South	ABRUZZO	2919	660	823
South	CAMPANIA	18,764	6,420	3,673
South	MOLISE	576	971	375
South	PUGLIA	14,822	7,426	1,423
	Total	**37,081**	**15,477**	**6,294**
South-Islands	BASILICATA	101	727	310
South-Islands	CALABRIA	1,759	917	1,105
South-Islands	SARDEGNA	684	1,730	514
South-Islands	SICILIA	3,169	3,005	1,581
	Total	**5,713**	**6,379**	**3,510**

Table 3: Total numbers of Bebras teams by region (data cover editions 2016, 2017, 2018)

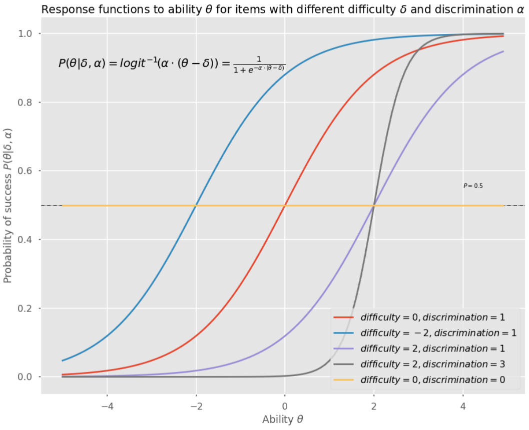

Figure 1: Logistic response functions

$$\overline{\delta} \sim Cauchy(200, 5), \sigma_\alpha, \sigma_\delta \sim Cauchy(0, 5),$$
$$\theta \sim Normal(200, 40),$$
$$\delta \sim Normal(\overline{\delta}, \sigma_\delta), \qquad \alpha \sim LogNormal(0, \sigma_\alpha),$$
$$y \sim BernoulliLogit(\alpha \cdot (\theta - (\delta + \overline{\delta}))/40).$$

In this model we assumed a Cauchy weakly informative prior distribution on hyper-parameters $\overline{\delta}$ — the mean difficulty used as a reference point in the logistic —, σ_δ, and σ_α — the standard deviation respectively of difficulty and discrimination —. The ability is then supposed to be normally distributed with mean = 200 and standard deviation = 40, the difficulty normally distributed with mean = 0 and standard deviation = σ_δ, and the logarithm of discrimination is normally distributed with mean = 200 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 [GH06, Sta16]). We sampled the Stan Monte Carlo model for 2,000 iterations, throwing away the first 1000 results (50% warm-up iterations). The results have all the typical properties of converging models, in particular the $\hat{R}$ statistics is close to 1 for every parameter of interest (a necessary, but unfortunately not sufficient, condition for convergence). Results are indeed sensible, with descriptive statistics consistent with score data, therefore we are rather confident that our model is plausible and useful to infer latent parameters.

4 Data analysis

In order to answer the research questions posed in Section 2, we start by identifying which variations among Bebras data are indeed significant. Ideally, we would like to filter out the differences due to statistical fluctuations. In fact, even the INVALSI 2018 report warns the readers that the differences in grade 5 results are too small to be considered a true assessment of the local competencies [INV18]. Unfortunately the report does not give enough details to replicate the significance test they used. We used a t-test between each pair of areas and regions, and we considered as significant those in which the t-test has a p-value $< 1 \times 10^{-4}$ (*i.e.,* the "null" hypothesis that the two generating distributions have the same mean is less probable than $\frac{1}{10000}$). Table 4 collects the significance of the results grouped by macro-area: only a few differences are significant at grade 5, but the overall significance increases with grades 8 and 10.

A similar pattern is also found when the results are grouped by regions, as reported in Table 5.

4.1 Analysis at the regional level

When one considers Bebras and INVALSI results grouped by region, the correlation among the rankings of the means is rather low. Tables 6,7, and 8 give the Kendall rank correlation coefficients respectively for grade 5, grade 8, and grade 10. The correlation increases with grades, but several inversions among the rankings remain.

In order to also appreciate the impact of the standard deviation of the results, we give the pictures of the distributions too (see Figures 2, 3, and 4, respectively grade 5, 8, and 10), approximated with a Gaussian with the same mean and standard deviation.

We also investigated if, whenever the difference in Bebras results between two regions is considered significant (see Table 5), the difference is in the same "direction" of the difference in INVALSI (please note, however, that we do not have detailed enough data to test if the difference in INVALSI results is also significant). For example, VENETO and CAMPANIA have a significant difference in Bebras results: VENETO performed better than CAMPANIA, and the same is true with respect to INVALSI tests.

For grade 5, we found 10 significant differences between regions, the differences have the same direction for 4 pairs. In the other 6 pairs, the directions differ: Bebras difference has the same direction of 'English reading' in 5 cases, of 'English listening' in 4 cases, of 'Italian' in 4 cases, of 'Mathematics' in 4 cases; thus, 17 cases out 24 are in the same direction.

For grade 8, we found 31 significant differences between regions and the differences have the same direction for all.

For grade 10, we found 78 significant differences between regions, the differences have the same direction for 63 pairs. In the other 15 pairs, the directions differ: Bebras difference has the same direction of 'Italian' in 1 case, in all other 29 cases the direction of Bebras difference is opposite of the difference in Italian and Mathematics, which instead are consistent between them.

All in all, we believe we have preliminary evidence that the answer to RQ1 is somewhat positive: at least when the difference is significant, the difference in Bebras mostly matches INVALSI differences.

4.2 Analysis at the level of macro-areas

With the exception of grade 5 (see Table 9, but at this grade, as noted above, the differences are mostly not significant), the correlation among the rankings of the means grouped by macro-areas is rather high. Tables 10 and 11 give the Kendall rank correlation coefficients respectively for grade 8 and grade 10.

Thus, also for RQ2 we believe we have evidence to answer positively, at least for the grades 8 and 10, where the differences between the results of the macro-areas are considered significant.

4.3 Analysis at the grossest level

The INVALSI 2018 report claims that the overall INVALSI results generally match PISA results: the Northern part of Italy performs better than the national average, while the Southern part performs worse. This pattern, with the best mean results in the two Northern macro-areas and the worst mean results in the two Southern macro-areas, is found also in Bebras. According to Bebras data, the Center macro-area performs slightly below the national average.

Thus, RQ3 seems also positively supported by our data.

4.4 Threats to validity

The 2018 INVALSI report does not give the details about the significance tests used to mark the differences at grade 5 as not significant, while at grades 8 and 10 they were considered so. Also, no pairwise (at both regional and macro-area levels) significance was reported. Since the Bebras sample is much smaller, we used a rather tight criterion: a t-test with a p-value threshold $< 1 \times 10^{-4}$. The underlying statistical model is the same in INVALSI and Bebras (2-parameter IRT), but we do not know the fitting ap-

Area	Center	North-East	North-West	South	South-Islands
Center	—	10	5	8	8 10
North-East	10	—	10	5 8 10	8 10
North-West	5	10	—	5 8 10	5 8 10
South	8	5 8 10	5 8 10	—	10
South-Islands	8 10	8 10	5 8 10	10	—

Table 4: Significance of the difference in Bebras results by macro-area, measured by a t-test. Cells show the grades in which the p-value is less than 1×10^{-3}, the threshold we used to reject the hypothesis that the two distributions have the same mean.

Region	LAZIO	MARCHE	TOSCANA	UMBRIA	LIGURIA	LOMBARDIA	PIEMONTE	VALLE D'AOSTA	EMILIA-ROMAGNA	FRIULI-VENEZIA GIULIA	TRENTINO-ALTO ADIGE	VENETO	ABRUZZO	CAMPANIA	MOLISE	PUGLIA	BASILICATA	CALABRIA	SICILIA	SARDEGNA
LAZIO	—		10		10	5 10	5 10		10			10			10			10	8 10	
MARCHE		—	10			10	10		10	10		10							8 10	
TOSCANA	10	10	—		10						10			8 10	10	10	10	10	8 10	10
UMBRIA				—					10						10				10	
LIGURIA	10		10		—	8 10	10	8	10	10		10		10					10	
LOMBARDIA	5 10	10			8 10	—	10		10				5	5 8	10	5 8 10		8 10	8 10	
PIEMONTE	5 10	10			10	10	—			10	10	10	5 10	5 8 10	10	5 10	10	10	8 10	10
VALLE D'AOSTA					8			—						8		8		8	8	
EMILIA-ROMAGNA	10	10		10	10	10			—	10	10	10	10	8 10	10	8 10	10	8 10	8 10	10
FRIULI-VENEZIA GIULIA		10			10		10		10	—				8	10	8 10		8 10	8 10	
TRENTINO-ALTO ADIGE			10				10		10		—			8				8	8 10	
VENETO	10	10			10		10		10			—		5 8 10	10	5 8 10	10	8 10	8 10	
ABRUZZO						5	5 10		10				—						10	
CAMPANIA			8 10		10	5 8	5 8 10	8	8 10	8	8	5 8 10		—	10			10	10	
MOLISE	10		10	10		10	10		10	10		10		10	—					
PUGLIA			10			5 8 10	5 10	8	8 10	8 10		5 8 10				—				
BASILICATA			10				10		10			10					—			
CALABRIA	10		10			8 10	10	8	8 10	8 10	8	8 10		10				—		
SICILIA	8 10	8 10	8 10	10	10	8 10	8 10	8	8 10	8 10	8 10	8 10	10	10					—	
SARDEGNA			10				10		10											—

Table 5: Significance of the difference in Bebras results by region, measured by a t-test. Cells show the grades in which the p-value is less than 1×10^{-4}, the threshold we used to reject the hypothesis that the two distributions have the same mean.

	Italian	Mathematics	Eng. listening	Eng. reading	*Bebras*
Italian	1.00	0.65	0.67	0.77	*0.10*
Mathematics	0.65	1.00	0.57	0.56	*0.01*
Eng. listening	0.67	0.57	1.00	0.89	*0.21*
Eng. reading	0.77	0.56	0.89	1.00	*0.23*
Bebras	*0.10*	*0.01*	*0.21*	*0.23*	1.00

Table 6: Kendall τ for grade 5 INVALSI and Bebras results (regions)

	Italian	Mathematics	Eng. listening	Eng. reading	*Bebras*
Italian	1.00	0.75	0.81	0.82	*0.56*
Mathematics	0.75	1.00	0.86	0.88	*0.54*
Eng. listening	0.81	0.86	1.00	0.95	*0.58*
Eng. reading	0.82	0.88	0.95	1.00	*0.53*
Bebras	*0.56*	*0.54*	*0.58*	*0.53*	1.00

Table 7: Kendall τ for grade 8 INVALSI and Bebras results (regions)

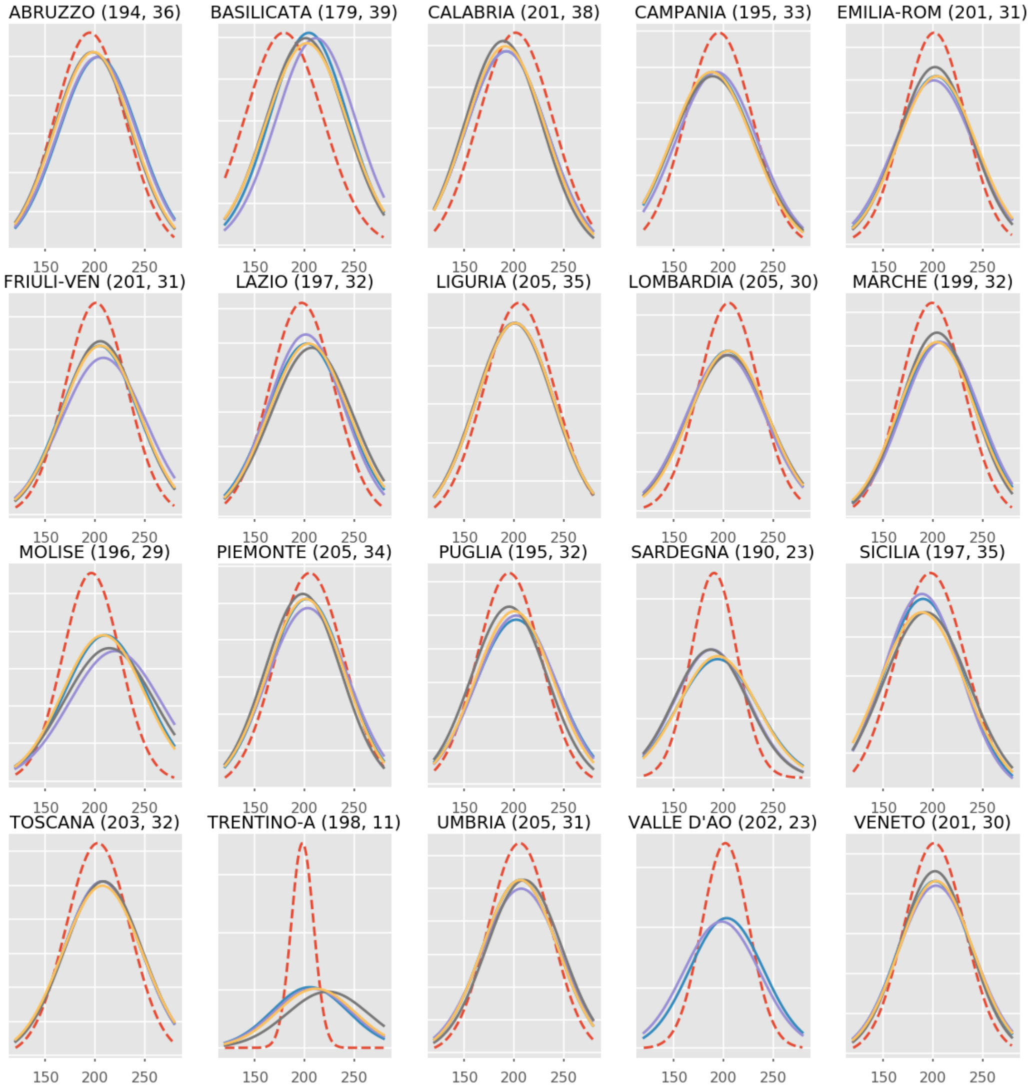

Figure 2: Grade 5 comparison between Bebras (dashed) and INVALSI (solid) results at regional level. The mean Bebras result and its standard deviation are also shown in the graph title in brackets.

	Italian	Mathematics	*Bebras*
Italian	1.00	0.90	*0.48*
Mathematics	0.90	1.00	*0.46*
Bebras	*0.48*	*0.46*	1.00

Table 8: Kendall τ for grade 10 INVALSI and Bebras results (regions)

proach used in INVALSI: to get numerically comparable results we used Normal distributions located in 200, with scale of 40. We adopted sensible prior parameter choices, common in the IRT literature, but we do not know if a difference considered significant in our model would be marked as such also by the INVALSI approach.

The main threat to validity, however, is the bias intrinsic in the Bebras sample. While INVALSI data cover every school in Italy and the sample surveyed in [INV18] was supposedly chosen with statistical goals in mind, we just used all the data of the teams who participated to the last three editions of the Italian Bebras and were able to ship a result with our online platform [BCL$^+$18]. Bebras pupils are thus drawn from the classes and schools with teachers interested in computational thinking and informatics (although this special interest is not necessarily shared by their pupils) and had the equipment and the logistic context suitable to participate. Also, while INVALSI tests individuals, Bebras is played in teams of 3–4 students.

For INVALSI we used the data as reported in [INV18], since we have no access to raw data. The data source is incomplete, for example no pieces of information are given about the numbers of sampled

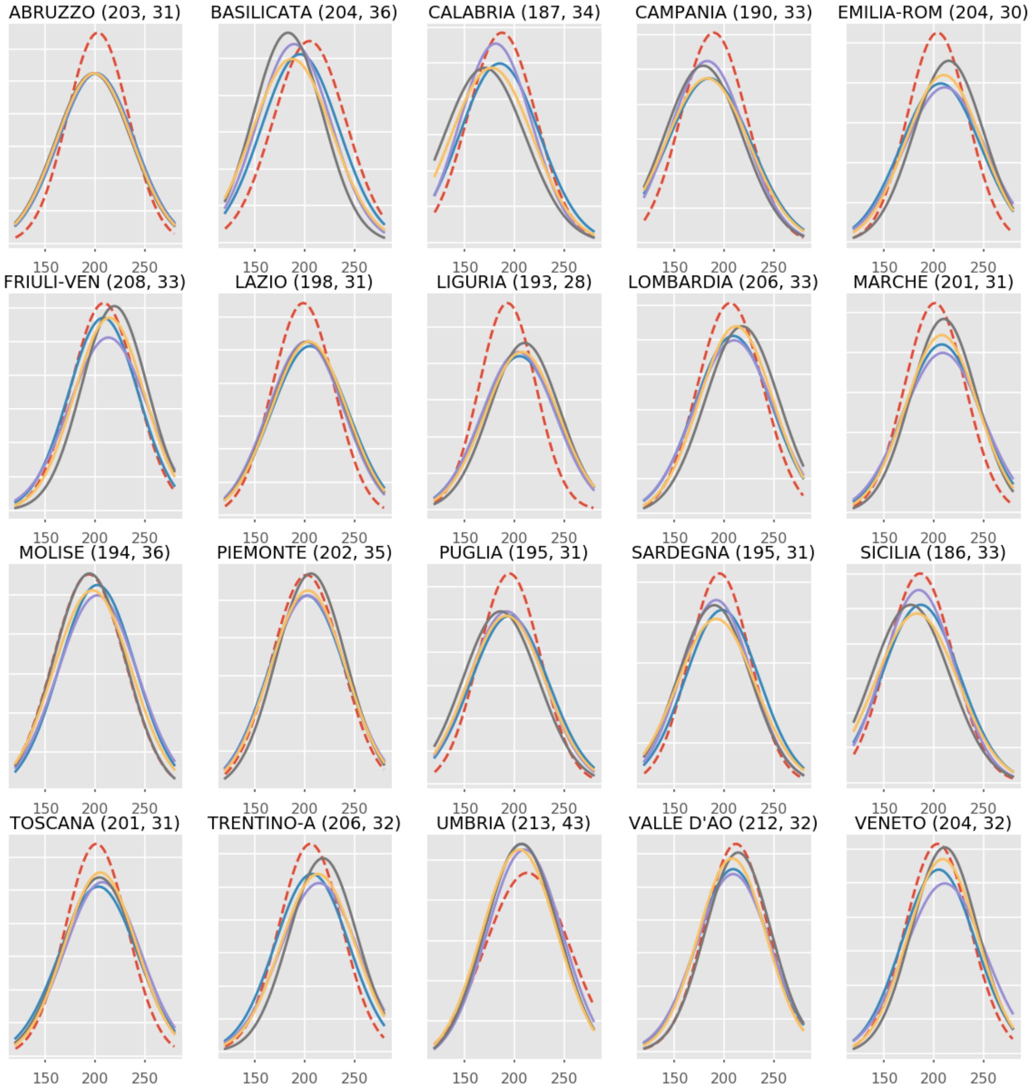

Figure 3: Grade 8 comparison between Bebras (dashed) and INVALSI (solid) results at regional level. The mean Bebras result and its standard deviation are also shown in the graph title in brackets.

students by region or even macro-area. This makes it impossible to aggregate data in different ways with respect to the ones given or to put together INVALSI data related to different school years.

5 Conclusions

We can conclude that yes, the data of the last three editions of the Italian Bebras support the hypothesis that the general INVALSI national assessment of Italian schools can be used to predict the performance of students in the Italian edition of the Bebras International Challenge on Informatics and Computational Thinking. This result is not completely obvious, since Bebras avoids tasks based on curricular subjects and technical jargon and INVALSI assesses competencies in linguistic and mathematical areas, not directly addressed by Bebras. In fact, Italian schools do not have curricular informatics in grades 5 and 8. The national guidelines for primary and lower secondary schools somewhat mention computational thinking, but the adoption in school and its perception by teachers is rather discontinuous [CLN17a, CLN17b]. Even in grade 10, informatics appears only in vocational curricula and science oriented programs. A more coherent proposal is under discussion (see [FLL$^+$18]), but currently we can safely assume that informatics and computational thinking are not routinely faced by the general population of Italian schools. Nevertheless, the Bebras snapshot seems to reflect the general geographic trend of Italian schools, even if the participants come from schools with a special interest in computational thinking and informatics. This could be an important result, because Bebras data can be

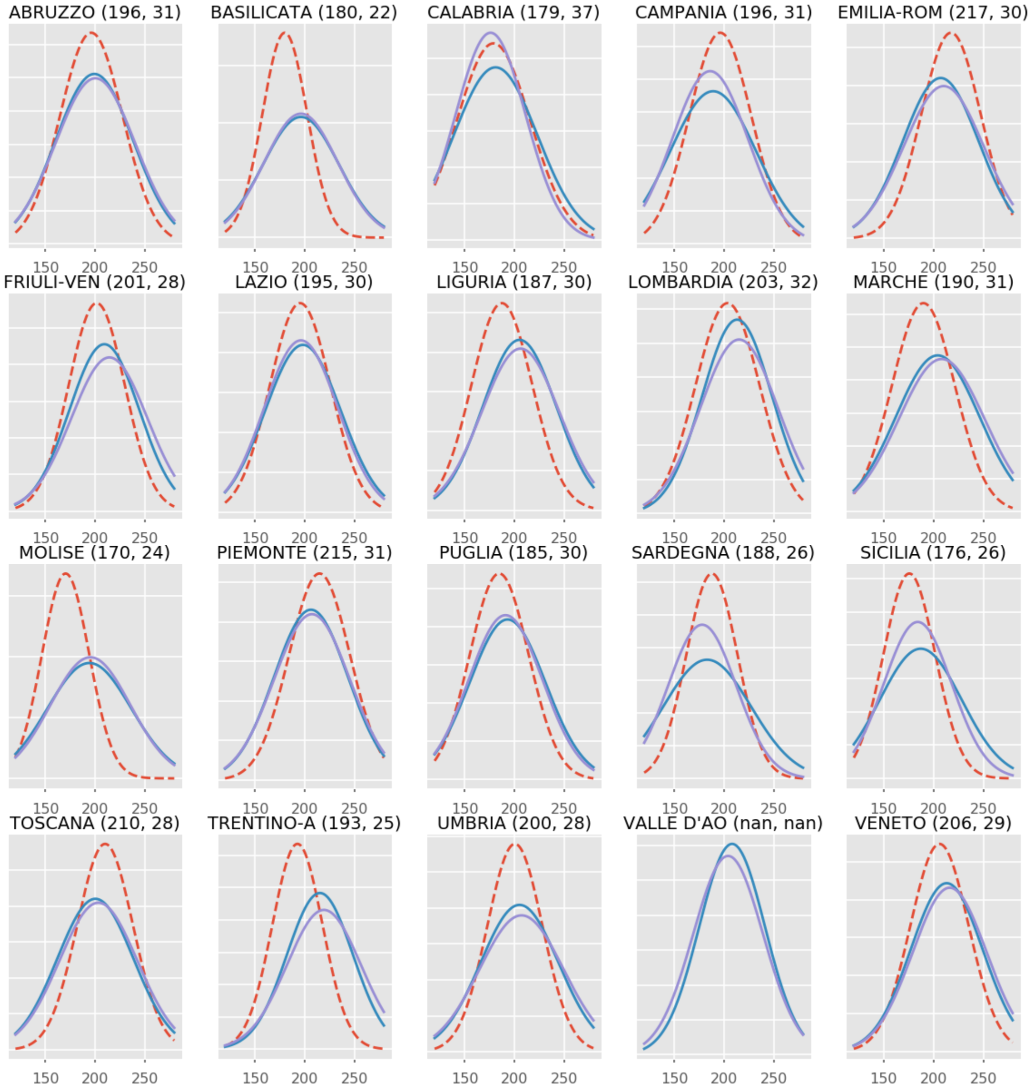

Figure 4: Grade 10 comparison among Bebras (dashed) and INVALSI (solid) results at regional level. The mean Bebras result and its standard deviation are also shown in the graph title in brackets.

used to assess the computational skills of the students and, according to our study, they have the potential to be generalized to a wider population.

Acknowledgments

The authors wish to thank Federico Pedersini and Massimo Santini for discussing early drafts of this paper.

References

[BCL+18] Carlo Bellettini, Fabrizio Carimati, Violetta Lonati, Riccardo Macoratti, Dario Malchiodi, Mattia Monga, and Anna Morpurgo. A platform for the Italian Bebras. In *Proceedings of the 10th international conference on computer supported education (CSEDU 2018) — Volume 1*, pages 350–357. SCITEPRESS, 2018. Best poster award winner.

[BLM+15] Carlo Bellettini, Violetta Lonati, Dario Malchiodi, Mattia Monga, Anna Morpurgo, and Mauro Torelli. How challenging are Bebras tasks? an IRT analysis based on the performance of Italian students. In *Proceedings of ITiCSE 2015*, pages 27–32, Vilnius, Lithuania, July 2015. ACM.

[CAC+18] Giuseppe Chiazzese, Marco Arrigo, Antonella Chifari, Violetta Lonati, and Crispino Tosto. Exploring the effect of a robotics laboratory on computational

	Italian	Mathematics	Eng. listening	Eng. reading	*Bebras*
Italian	1.00	0.80	0.89	0.80	*0.40*
Mathematics	0.80	1.00	0.89	1.00	*0.20*
Eng. listening	0.89	0.89	1.00	0.89	*0.45*
Eng. reading	0.80	1.00	0.89	1.00	*0.20*
Bebras	*0.40*	*0.20*	*0.45*	*0.20*	1.00

Table 9: Kendall τ for grade 5 INVALSI and Bebras results (macro-areas)

	Italian	Mathematics	Eng. listening	Eng. reading	*Bebras*
Italian	1.00	0.80	0.95	0.80	*0.80*
Mathematics	0.80	1.00	0.95	1.00	*1.00*
Eng. listening	0.95	0.95	1.00	0.95	*0.95*
Eng. reading	0.80	1.00	0.95	1.00	*1.00*
Bebras	*0.80*	*1.00*	*0.95*	*1.00*	1.00

Table 10: Kendall τ for grade 8 INVALSI and Bebras results (macro-areas)

	Italian	Mathematics	*Bebras*
Italian	1.00	0.95	*0.95*
Mathematics	0.95	1.00	*1.00*
Bebras	*0.95*	*1.00*	1.00

Table 11: Kendall τ for grade 10 INVALSI and Bebras results (macro-areas)

thinking skills in primary school children using the bebras tasks. In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'18, pages 25–30, New York, NY, USA, 2018. ACM.

[CLN17a] Isabella Corradini, Michael Lodi, and Enrico Nardelli. Computational thinking in italian schools: Quantitative data and teachers' sentiment analysis after two years of programma il futuro. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, pages 224–229. ACM, 2017.

[CLN17b] Isabella Corradini, Michael Lodi, and Enrico Nardelli. Conceptions and misconceptions about computational thinking among italian primary school teachers. In *Proceedings of the 2017 ACM Conference on International Computing Education Research*, pages 136–144. ACM, 2017.

[Dag10] Valentina Dagienė. Sustaining informatics education by contests. In *Proceedings of ISSEP 2010*, volume 5941 of *Lecture Notes in Computer Science*, pages 1–12, Zurich, Switzerland, 2010. Springer.

[Des18] Marta Desimoni. *Prove INVALSI 2018*, chapter Le prove carta e matita per la rilevazione nazionale degli apprendimenti INVALSI 2018: aspetti metodologici. INVALSI, 2018. `https://invalsi-areaprove.cineca.it/docs/2019/Parte_II_capitolo_2_aspetti_metodologici_P&P_2018.pdf`.

[DS16] Valentina Dagienė and Sue Sentance. It's computational thinking! bebras tasks in the curriculum. In *Proceedings of ISSEP 2016*, volume 9973 of *Lecture Notes in Computer Science*, pages 28–39, Cham, 2016. Springer.

[FLL^{+}18] Luca Forlizzi, Michael Lodi, Violetta Lonati, Claudio Mirolo, Mattia Monga, Alberto Montresor, Anna Morpurgo, and Enrico Nardelli. A core informatics curriculum for Italian compulsory schools. In Pozdniakov S. and Dagienė V., editors, *Informatics in schools. fundamentals of computer science and software engineering. ISSEP 2018.*, volume 11169 of *LNCS*, pages 141–153. Springer, Cham, 2018.

[GH06] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press, Cambridge, UK, 2006.

[HCD11] Bruria Haberman, Avi Cohen, and Valentina Dagienė. The beaver contest: Attracting youngsters to study computing. In *Proceedings of ITiCSE 2011*, pages 378–378, Darmstadt, Germany, 2011. ACM.

[HM14] Peter Hubwieser and Andreas Mühling. Playing PISA with Bebras. In *Proceedings* *of the 9th WiPSCE*, pages 128–129, New York, NY, USA, 2014. ACM.

[HS85] Ronald K. Hambleton and H. Swaminathan. *Item Response Theory: Principles and Applications*. Springer-Verlag, Berlin, 1985.

[INV18] INVALSI. Rapporto prove INVALSI 2018. Technical report, INVALSI, 2018. Only in Italian, available at `https://www.invalsi.it/invalsi/doc_evidenza/2018/Rapporto_prove_INVALSI_2018.pdf`.

[KVC06] Graeme Kemkes, Troy Vasiga, and Gordon V. Cormack. Objective scoring for computing competition tasks. In *Proceedings of 2nd ISSEP*, volume 4226 of *Lecture Notes in Computer Science*, pages 230–241, Berlin, Germany, 2006. Springer.

[LMM+17] Violetta Lonati, Mattia Monga, Anna Morpurgo, Dario Malchiodi, and Annalisa Calcagni. 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, Cham, CH, 2017. Springer International Publishing AG. To appear.

[SBS17] Suzanne Straw, Susie Bamford, and Ben Styles. Randomised controlled trial and process evaluation of code clubs. Technical Report CODE01, National Foundation for Educational Research, May 2017. Available at: `https://www.nfer.ac.uk/publications/CODE01`.

[Sta16] Stan Development Team. Stan modeling language users guide and reference manual version 2.19.0. `http://mc-stan.org`, 2016.