Universality, Correlations, and Rankings in the Brazilian Universities National Admission Examinations

Roberto da Silva[1], Luis C. Lamb[2], Marcia C. Barbosa[1]

1-Instituto de Fisica, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500 - CEP 91501-970, Porto Alegre, Rio Grande do Sul, Brazil E-mail:rdasilva,marcia.barbosa@if.ufrgs.br

2-Instituto de Informatica, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500 - CEP 91501-970, Porto Alegre, Rio Grande do Sul, Brazil E-mail:lamb@inf.ufrgs.br

Abstract

We analyze the scores obtained by students who have taken the ENEM examination: The Brazilian High School National Examination which is used in the admission process at Brazilian universities. The average high schools scores from different disciplines are compared through the Pearson correlation coefficient. The results show a very large correlation between the performance in the different school subjects. Even thought the students' scores in the ENEM form a Gaussian due to the standardization, we show that the high schools' scores form a bimodal distribution that cannot be used to evaluate and compare students performance over time. We also show that this high schools distribution reflects the correlation between school performance and the economic level (defined as the average family income) of the students. The ENEM scores are compared with a Brazilian non standardized exam, the entrance examination at the Universidade Federal do Rio Grande do Sul. The analysis of the performance of the same individuals in both tests showing that the two tests not only select different abilities, but also lead to the selection of different sets of individuals. Our results indicate that standardized tests might be an interesting tool to compare performance of individuals over the years, but not of institutions.

1. Introduction

The selection of part of the population to access higher education is a challenging process that has implications for the future of the nation. China was the first country in the world to face such a challenge. The imperial examination, created in 605 during the Sui Dynasty, was a civil service test system. During the Imperial China this process aimed at selecting candidates for the state bureaucracy. This system lasted until 1905 [1].

Recognizing that the use of a standardized test to select the elite would play a key role in the future of the United Kingdom, a test was introduced into Europe in the early 19th century by the British consul in Guangzhou, China, Thomas Taylor Meadows [2]. Then, in 1806 the United Kingdom implemented the selection of public servants through this examination.

In the higher education system, standardized tests were first employed by Napoleon who created le baccalauréat or simply le bac. In the same period, the General Certificate of Secondary Education (GCSE) was implemented in the United Kingdom. Even thought le bac was created first, it was from Britain that standardized testing spread, not only throughout the British Commonwealth, but also to Europe and then to America. In the United States two systems dominate the university admission system: the Scholastic Aptitude Test (SAT) and the American College Testing (ACT) created in 1926 and 1959, respectively.

What **current standardized tests** in the United States, European Countries and Asia have in common is that they are organized in such a way that the scores follow a normal distribution [3, 4], $f(x)$, which is characterized by the mean $\langle x \rangle$ and standard deviation σ. The results of a particular candidate in one test, x_i, becomes universally comparable by the regular transformation $z_i = (x_i - \langle x \rangle)/\sigma$. These tests are based in the Item Response Theory. It treats the difficulty of each item as information to be incorporated in scaling items. It is based on the idea that the probability of a correct response to an item is a mathematical function. Then it becomes natural that the resulting scores follow into a normal distribuiton.

Even thought they are quite appealing due to their simplicity, the use of standardized tests in university admissions is not free from criticisms [5, 6]. Because the exams are t ested on a biased population, minorities and foreigners often present difficulties in understanding the cultural subtleties [7, 8]. In addition it is not clear that results from a given year or from a given test can be compared with the results from other years or with other tests. Results have even shown that students performance at high school is a good predictor of college performance [9].

In the particular case of the United States, the university admission process is multidimensional since it takes into account not only the SAT or ACT scores, but also the students' performance at high school, recommendation letters and extra curricular activities. Thus, limitations and criticisms of the standardized test method possibly have a lower impact in the north american admission process when compared with countries in which a standardized test score is the only criteria used in the university admission process. In addition, several countries use a number of competing standardized tests which also guarantees that the admission process does not become hostage of one particular evaluation method.

A proper analysis of the standardized tests that would answer the criticisms from [9] to such methods is not possible, since the scores of all these standardized exams are not available for a detailed analysis.

Until the end of the 20th century, Brazilian universities organized their own individual admission tests. Even though this method **guarantees diversity** in the admission process, it made student mobility across the country rather difficult. In addition, differently from the United States and some European Union countries, Brazil does not use an university admission system based on historical or annual tests of high school students, but only on an entrance examination. Typically, the university entrance examination is composed of several multiple choice tests which encompass all high school subjects. In some cases, the student is also required to write a short essay.

Over the last decade, the Brazilian government has introduced a standardized university entrance examination known as ENEM (*pt*: Exame Nacional do Ensino Médio - *en*: High School National Examination). This new exam is applied across the country, thus allowing for a probable increase in students mobility from one state to another, it employs a methodology **that allows for** comparison of the scores obtained in one year with the scores of the previous years and it is elaborated in a centralized form by the Ministry of Education. The major drawback of using one particular exam to admit the students is that the system becomes dependent on the type of analysis. Additional problems are the following. The exam is too ample: it covers a very large number of questions, and many students are not able to finish the (long) exam questions in the allowed exam time. This means that rarely the student solves all the questions. Then, the system has to compare students that solved not the same questions but questions belonging to different samples. Thus, candidates with partial knowledge of the high school subjects potentially can have the same opportunity (and perform similarly to) a candidate with a comprehensive background. Extensive, unclear and redundant question statements take too much time to read and grasp and do not explore relevant knowledge of the students; rather, understanding a question statement has affected the students performance.

It is also important to mention that the Item Response Theory is implemented in a very obscure way. In addition the brute scores are not made available for external analyses by the independent scientific community. Finally, no changes have been made in the exam methodology since its inception; changes in the exam could lead to improvements in the quality of the test questions.

Despite the many criticisms about the contents of the ENEM's questions [10, 11], the process has its merits. If it is managed and carried out properly it would lead to an interesting mechanism to tackle biases and distortions towards bringing a larger contingent of state owned high school students to public, research universities.

However, before it becomes a unique tool to evaluate every student in Brazil, ENEM has to be evaluated and tested against other existing local exams. As far as we know, such an analysis was never done with other standardized tests, maybe with the exception of the SAT, whose performance has been checked against high school grades, but only for a small number of students [9].

In this paper, we make use of statistical physics tools to understand the universal aspects of ENEM. Our methodology is not new and has been used to analyze high school performance [12, 13, 14]. The scores that the students obtained in the different disciplines in ENEM examination are analyzed. In addition, the scores obtained by the different high schools in the same exam are also evaluated. Finally, a comparison between the performance of a selected number of students at the ENEM and at a local admission examination from one specific university - the Universidade Federal do Rio Grande do Sul (UFRGS) - during three consecutive years is also described providing a unique tool to identify what differs in the profile of the students selected by each method.

The remaining of the paper goes as follows. In Section 2 we describe the data set employed our analyses; in Section 3 describe the technical results; Section 4 concludes the paper and outlines further research directions.

2. Data Set Analysis

The first part of our data set supplies the average scores of $14,715$ high schools from Brazil in 2013 by considering: School percentage (participation) rate of their students and the income/economic level (average family income) of the school.

The ENEM examination is composed by five different school subjects: Writing, Language, Human Sciences, Natural Sciences, and Mathematics. **The income/economic level of the schools are divided into 7 different levels**: very high, high, high average, average, low average, low, and very low. We attributed $3, 2, 1, 0, -1, -2, -3$ respectively for these levels.

The second data set supplies ENEM and UFRGS entrance examination scores of the students who have taken both exams. We have analyzed students from three consecutive years 2011, 2012 and 2013. Here we have cleaned the data by extracting students that have score *zero* in one or more school subjects. For example for 2011 after data cleaning we have $11,515$ students who took the Writing school subject test at UFRGS. From these students only $10,315$ scored above grade zero, which is then the minimum size sample used in our work. These data are in used all comparison tests (Pearson correlation and ranking tests). This also means that we also work with larger datasets when we consider the crossovers between every pair of school subjects combinations: UFRGS/UFRGS or ENEM/ENEM or even UFRGS/ENEM. This guarantees good significance of our calculations. For example, we find in 2013 more than $25,000$ students who took the Maths school subject in both UFRGS and ENEM with score greater than zero.

The high school subjects analyzed by the UFRGS entrance examination are: Writing, Language, Mathematics, Geography, History, Physics, Chemistry and Biology. In the case of ENEM the subjects are: Writing, Language, Mathematics, Humanities and Natural Sciences. Then for comparison purposes we adopted the following strategy. The scores in the Geography and in the History tests at UFRGS were compared with the scores in the Humanities test at ENEM while the scores in the Chemistry, in the Biology and in the Physics tests at UFRGS were compared with the scores in the Natural Sciences test at ENEM.

3. Results

3.1. ENEM Scores in the Brazilian High Schools

First, the correlations between the scores at different subjects obtained by all the high schools were computed. Figure 1 illustrates the comparison between these scores. Visually, these diagrams show a strong linear correlation between the scores of different subjects. This indication can be quantified by a single number, the Pearson correlation coefficient given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \langle x\rangle)(y_i - \langle y\rangle)}{\sqrt{\sum_{i=1}^{n}(x_i - \langle x\rangle)^2}\sqrt{\sum_{i=1}^{n}(y_i - \langle y\rangle)^2}} \quad (1)$$

where x_i and y_i represent the scores of two different subjects associated to i-th institution, with $i = 1,...,n$. The values of r range from -1 when the two data sets are negatively correlated, to 0 when they are uncorrelated up to 1 when they are positively correlated. Since r is computed averaged over all the $n = 14,715$ schools, it gives a robust indication of the correlations between the performance of the schools in the different topics.

Table 1 illustrates the values of r for the different pairs of subjects. We can observe a high correlation among the different schools which is not a surprise indeed, since the schools scores are more representative as they represent averages over many students. However some particularities must be mentioned. All subjects are more correlated with Language and Humanities (or Human Sciences) than with Writing. This is quite surprising since, in principle, one would expect that Natural Sciences and Mathematics would show a less evident correlation with Language or Humanities. Language and Human Sciences are slightly more correlated with Writing than Natural Sciences and Maths. Although the biggest correlations are found in the somewhat more intuitive cases, e.g. between Language and Human Sciences and between Natural Sciences and Maths, $r = 0.9554$ and $r = 0.9531$, respectively, we also found $r = 0.9523$ between Human and Natural Sciences and $r = 0.9408$ between Human Sciences and Maths, which are not expected results, as the analysis was made with correlations for the different schools. The last row of the table (in bold) corresponds to correlation coefficients between each school subject and the average final score of the **schools which is quite strong**. This indicates that either the schools in Brazil present no specific strength in any subject, or the exam is unable to capture the difference in the performance of the schools in different areas of knowledge.

Writing	Language	Human Sciences	Natural Sciences	Math
Writing	0.8878	0.8899	0.8624	0.8555
–	Language	0.9554	0.9250	0.9243
–	–	Human Sciences	0.9523	0.9408
–	–	–	Natural Sciences	0.9531
0.8918	**0.9694**	**0.9823**	**0.9791**	**0.9801**

Table 1: Pearson correlation coefficients, r, between two subjects scores in the ENEM 2013. The last row corresponds to the coefficient between each subject score and the average score of the school.

One of the claims of standardized exams is that they would make it possible for students coming from disadvantaged areas and schools to access the best universities in the country. In order to test this hypothesis, two different parameter **were computed**: (a) the scores as a function of the social-economic level of the schools and (b) the score as a function of the number of the students' participation at ENEM, namely the ratio between the number of students that effectively took the examination and the total number of students that were eligible to take the examination.

Figure 2 (a) shows the scores as a function of the socioeconomic level of the high school. It is clear that the socioeconomic level is quite relevant for good school performance. In particular, it is important to observe the large slope after the "high average level". The small error bars indicate reliable results. Figure 2 (b) shows that the score increases with the percentage of the school participation, showing a linear correlation namely $Score = 360(2) + 2.05(2) \cdot \rho$.

We compute the distribution of the scores in order to test to check if the scores at different disciplines are correlated. Here this calculation is done in terms of the normalized value given by

$$z = \frac{\sqrt{n}(score - \langle score \rangle)}{\left\langle (score - \langle score \rangle)^2 \right\rangle}.$$

Figure 3 shows the normalized score distributions for the different school subjects (points) in mono-log scale. The continuous curve represents the average score distribution. The inset plot is depicted to facilitate observation from the traditional linear scale point of view. It is important to highlight that all the different subjects follow the same distribution of scores fluctuation and this distribution is not Gaussian since in mono-log scale we do not observe a second degree polynomial behavior.

What would the distribution of the scores be? In order to respond this question, a few distributions shown in Table 2 have been used to fit the scores of the schools. First, the standard two-parametric statistical distributions (normal and log normal) were checked. In this case x_c and σ were the free parameter for the fit. Next, other more complicated asymmetrical distributions based on three or four parameters were also checked.

The quality of the different fits performed here is checked by the following procedure. Given the original data set $y_1,...,y_n$ with n values and the fit of these values given by the functions values $f_1,..,f_n$ the **quality** of the fit is given by

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{tot} - SS_{res}}{SS_{tot}} \tag{2}$$

where SS_{res}, known as residual sum of squares is calculated by

$$\begin{aligned} SS_{res} &= \sum_{i=1}^{n} (y_i - f_i)^2 \\ SS_{tot} &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \, . \end{aligned} \tag{3}$$

In a general form, R^2 can be related to the unexplained variance, since the second term compares the unexplained variance (variance of the model's errors) with the total variance (of the data). Since $0 \leq R^2 \leq 1$, with $R^2 = 1$ when original data and fit are identical, it gives a good measurement of how far the fit is from the original data. It is also important to mention that in the linear least squares regression, R^2 is equal to the square of the Pearson correlation coefficient given by Equation 1.

Even thought the **Gaussian distribution appears more promptly in most problems in the natural sciences**, the multiple parameters approach is observed in the particles movement in random media [16], noise in semiconductor devices [17], and stochastic aspects of soccer scores [18]. For all the tested distributions exemplified in Table 2 the parameter R^2 was computed. Here the tested distributions are the normal or Gaussian (N), the log normal (LN), the Exponentially Modified Gaussian (EMG), Gram-Charlier peak function (GC) and Edgeworth-Cramer peak function (EC). Then, the fits using these distributions were performed by the Levenberg-Marquardt method [19] for non-linear fits.

In the case of the EMG, GC, and EC distributions, two approaches have been employed. First, two parameters were estimated with a statistical measure and the others were fitted. For example, in the case of the EMG, GC, and EC distributions, x_c was fixed with the average of the scores. This procedure yields $x_c = \langle x \rangle = 513.36$. Then, σ is computed in the standard deviation of the original data. This gives $\sigma = \left\langle x^2 \right\rangle - \langle x \rangle^2 = 52.1$. With x_c and σ fixed, the only free parameters for the EMG, GC, and EC distributions become (w,t_0), (a_3,a_4), and (b_3, b_4) respectively. In addition to the fit with two parameters, a fit with all parameters was performed. The comparison between the value of R^2 (see Eq. 2) obtained using these two fitting methods is illustrated in Table 2.

Figure 4(a) illustrates the comparison of the original data with the N, LN, EMG,GC and EC, the last

Dist.	Formula	parameters	R^2_{two}	R^2_{all}
N	$f(x) = \frac{e^{-(x-x_c)^2/2\sigma^2}}{(2\pi\sigma^2)^{1/2}}$	2: x_c and σ	0.868	0.868
LN	$f(x) = \frac{e^{-\frac{1}{2\sigma^2}\ln^2(x/x_c)}}{(2\pi\sigma^2)^{1/2}x}$	2: x_c and σ	0.899	0.899
EMG	$f(x) = \frac{1}{t_0}e^{\frac{1}{2}(\frac{w}{t_0})^2-\frac{x-x_c}{t_0}}\int_0^{\frac{x-x_c}{w}-\frac{w}{t_0}} e^{-t^2}dt$	3: x_c, w, and t_0	0.811	0.971
GC	$f(x) = \frac{e^{-z^2/2}}{(2\pi\sigma^2)^{1/2}}\left[1+\frac{a_3}{3!}(z^3-3z)+\frac{a_4}{4!}(z^4-6z^3+3)\right]$ with $z=(x-x_c)/\sigma$	4: x_c, σ, a_3, and a_4	0.945	0.970
EC	$f(x) = \frac{e^{-z^2/2}}{(2\pi\sigma^2)^{1/2}}\left[1+\frac{b_3}{3!}(z^3-3z)+\frac{b_4}{4!}(z^4-6z^3+3)\right.$ $\left.\frac{10b_3^2}{6!}(z^6-15z^4+45z-15)\right]$ with $z=(x-x_c)/\sigma$	5: x_c, σ, b_3, and b_4	0.958	0.979

Table 2: Functions used to fit the distribution of scores, x, of the schools. The last two columns show the determination coefficient R^2 by using, respectively, two and all parameters of the considered functions. For the computation of R for EMG, GC, and EC with only two parameters, the parameters x_c and σ were fixed by the average and standard deviation estimated from the original data

three employing a two parameters fit. The visual inspection of the graphs support the results of the determination coefficient R^2 [20] shown in Table 2 indicate that when using two parameters, EC is the best fitting distribution. In Figure 4 (b) the original data is compared with the results for the distributions EMG, GC and EC but using all the parameters for the fit. In this case, the EC performance is the most efficient and it is more efficient than when the adjustment is done with only two parameters.

Even though the ENEM is constructed to give a standardized score of individuals, this is not the case for the score of the schools. The distribution shows a region with a peak at the score 500 and another peak at the score 550 which presents two distinct score evolutions. This observation is supported by Figure 2 which shows an abrupt change in the slope of the averaged scores with the increase of economic level of the school. It is important to point out that since the schools' scores are not Gaussians, the schools' score evolution over time is not a reliable measure, since the score of one year cannot be compared with the score of the subsequent year, simply because they are not standardized.

3.2. ENEM and UFRGS Students' Scores

Next, the performance of the students is analyzed. In order to check how the ENEM selection differs from the traditional methods employed by the Brazilian Universities in the past, we analyze the performance of the students that have done both the ENEM and the UFRGS entrance examinations. It is important to emphasize that here we are comparing the performance of the same group of people in both exams.

Table 3 shows for the years 2011, 2012 and 2013 the correlations, r, between the scores obtained by the students in the different subjects at the ENEM tests.

Year	Writing	Language	Human Sciences	Natural Sciences	Maths
2011	Writing	0.349	0.343	0.313	0.232
	–	Language	0.710	0.668	0.599
	–	–	Human Sciences	**0.772**	**0.619**
	–	–	–	Natural Sciences	0.723
	Writing	Language	Human Sciences	Natural Sciences	Maths
2012	Writing	0.362	0.360	0.345	0.261
	–	Language	0.744	0.673	0.575
	–	–	Human Sciences	**0.773**	**0.647**
	–	–	–	Natural Sciences	0.725
	Writing	Language	Human Sciences	Natural Sciences	Maths
2013	Writing	0.463	0.477	0.445	0.378
	–	Language	0.769	0.675	0.597
	–	–	Human Sciences	**0.745**	**0.652**
	–	–	–	Natural Sciences	0.766

Table 3: Correlation coefficients, r, between scores in the different exam subjects at the ENEM of the students who have also taken the entrance exam of UFRGS in the years 2011, 2012 and 2013.

It is interesting to observe that the correlation between the students' scores in all subjects is large with the exception of Writing. It is particularly intriguing the high correlation between the scores on Human and Natural Sciences and Math, usually topics at school in which the performance of the students can differ a lot. One possible explanation for this phenomena is related to the fact that the ENEM questions are quite long with the addition of a contextualization usually absent in problem-solving texts in hard sciences and Maths. It is important to mention that such behavior is the same for all the years we have analyzed in our work. The low correlation between the Writing test and the other subject tests can be understood because this is the only part of the exam that is not manipulated by the standardized method.

In order to check if the high correlation between scores is a characteristic of the standardized procedure employed by ENEM or it is due to the students' profile, the same analysis was performed for the scores at the entrance exam at UFRGS.

Table 4 illustrates the correlation between the students' scores at different subjects at the entrance at UFRGS during the years of 2011, 2012 and 2013. The division in subject areas in the UFRGS tests is a little different from the ENEM tests. In the case of UFRGS, Natural Sciences is divided into Physics, Chemistry and Biology, while Human Sciences is divided into History and Geography. It is interesting to notice that the correlation between Human Sciences and Natural Sciences is much lower than the correlation observed in the ENEM and clearly a high correlation is present only between the Physics, Chemistry and Maths as

Year		Writing	Math	Phys	Chem	Bio	Geo	Hist
2013	Writing	0.381	0.327	0.366	0.369	0.319	0.372	
			Math	0.744	0.731	0.652	**0.576**	**0.583**
				Phys	0.697	0.634	**0.548**	**0.552**
					Chem	0.671	**0.559**	**0.587**
						Bio	**0.575**	**0.600**
							Geo	0.587
2012	Writing	0.366	0.335	0.330	0.323	0.323	0.363	
			Math	0.781	0.744	0.638	**0.594**	**0.557**
				Phys	0.750	0.663	**0.585**	**0.570**
					Chem	0.649	**0.564**	**0.544**
						Bio	**0.534**	**0.543**
							Geo	0.606
2011	Writing	0.307	0.308	0.319	0.322	0.314	0.319	
			Math	0.736	0.732	0.611	**0.587**	**0.524**
				Phys	0.757	0.662	**0.595**	**0.571**
					Chem	0.687	**0.608**	**0.556**
						Bio	**0.597**	**0.562**
							Geo	0.602

Table 4: Correlation, r, between the scores of the different subjects at the UFRGS examinations in the years 2011, 2012 and 2013

traditionally is observed at high schools. As in the case of ENEM, **Writing** shows a very small correlation with other subjects. In the case of the UFRGS examination, the Writing test is not use for student elimination purposes, but for student classification which means that this ability is used to discriminate between people equality apt to enter the university testing for better communication skills.

Such difference between the correlations among subjects in the two exams make the need to directly compare the scores in the same subjects quite clear. Table 5 illustrates such a comparison. This table shows that the Writing test not only is not correlated with other subjects within the same exam, but also that it is not correlated with the performance in the other exams. In addition, the correlation between the scores in other subjects when two exams are compared is not high, except for Maths.

Although correlations are high, we would expect even higher correlations between the two examinations if they intend to admit good candidates to the university (UFRGS). Let us observe that the University has been able to admit good students and the institution has achieved high rankings in all evaluations carried out by the Ministry of Education over the last decades. UFRGS is consistently ranked among the top 5 universities in Brazil for both research and education. It is important to mention and note the small

UFRGS-ENEM	Writing	Math	Human/Geo	Human/Hist	Phys/Nat	Chem/Nat	Bio/Nat
2011	0.313	0.700	0.627	0.628	0.643	0.668	0.653
2012	0.313	0.728	0.654	0.687	0.684	0.676	0.641
2013	0.384	0.759	0.613	0.673	0.679	0.692	0.681

Table 5: Correlation between the scores at specific subjects, r, at the ENEM and at the UFRGS examinations in the years 2011, 2012 and 2013.

Skewness	**2011**	**2012**	**2013**	**Kurtosis**	**2011**	**2012**	**2013**
Writing	-0.056	$-8.7.10^{-4}$	$+0.16$	Writing	-0.28	-0.17	-0.42
Language	-0.54	-0.60	-0.35	Language	$+0.98$	$+1.21$	$+0.33$
Humanities	-0.46	-0.24	-0.31	Humanities	$+0.56$	$+0.35$	$+0.057$
Natural Sciences	-0.35	-0.031	-0.098	Natural Sciences	$+0.31$	$+0.19$	-0.34
Math	-0.26	-0.36	-0.29	Math	-0.11	-0.18	-0.12

Table 6: Skewness and kurtosis of the ENEM's score distributions in the years 2011, 2012 and 2013.

correlations between the Writing test between two exams.

The differences between the two exams was also checked by comparing the distribution of the scores for the Maths test. Figure 5 illustrates the ENEM's and the UFRGS's distributions for Maths for the years 2011, 2012 and 2013. The ENEM's distributions are visually Gaussian forms while the UFRGS exam are distinct when compared with a Gaussian. These similarities and differences can be computed by two quantities: skewness and kurtosis. Skewness is a measure of lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. A symmetrical distribution has a skewness of zero, while an asymmetrical distribution with right(left) tail has a positive(negative) skew. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. A Gaussian distribution has a kurtosis of 0, while a flatter distribution has a negative kurtosis and a very peaked distribution has a positive kurtosis. Table 6 shows the kurtosis and the skewness of the ENEM score distributions in the years 2011, 2012 and 2013 while Table 7 shows the kurtosis and the skewness of the UFRGS score distributions for the same period.

The tables show that there is a negative skewness for ENEM's Maths scores, but positive in the UFRGS's scores in the analyzed years. The same occurs, now shown here for simplicity, for Natural Sciences (ENEM) when compared with Physics, Chemistry and Biology (UFRGS) and Humanities (ENEM) when compared with History and Geography (UFRGS). For the kurtosis, for example, we have opposite signals for the writing test in ENEM and UFRGS for Writing and Humanities.

Such differences can be observed for a particular case, i.e. the Maths test. We can see the deviation from normal of the UFRGS examination which is not observed for the ENEM examinations. This corroborates

Skewness	2011	2012	2013	Kurtosis	2011	2012	2013
Writing	−0.19	−0.33	−0.17	Writing	0.14	0.64	0.20
Geo	0.34	0.18	0.15	Geo	−0.28	−0.44	−0.24
Hist	0.28	0.25	0.25	Hist	−0.36	−0.41	−0.38
Math	0.88	0.82	0.77	Math	0.13	−0.14	−0.27
Phys	0.85	1.03	0.90	Phys	0.26	0.58	0.69
Chem	0.89	0.99	0.81	Chem	0.22	0.59	0.25
Bio	0.64	0.81	0.63	Bio	0.11	0.62	0.074

Table 7: Skewness and kurtosis of the UFRGS's score distributions in the years 2011, 2012 and 2013.

the results found in Tables 6 and 7.

These results suggest that the exams rank the students in a different order. In order to check this hypothesis, the following strategy was employed. The differences between the rankings of students according to their scores in the two exams was obtained by denoting by $r_{ENEM}(i)$ the rank of the i-th student in the ENEM examinations and denoting by $r_{UFRGS}(i)$ the rank corresponding to the same student at UFRGS. Then, we define the following

$$d_i = r_{UFRGS}(i) - r_{ENEM}(i), \tag{4}$$

that measures the difference between the ranks established by the two exams for a specific school subject.

Then the average difference in the α ranking index becomes

$$\alpha = \frac{\langle |d_i| \rangle}{N_{total}} = \frac{\sum_{i=1}^{N_{total}} |d_i|}{N_{total}^2} \tag{5}$$

where N_{total} is the total number of analyzed students in which we choose to represent in percentages. It measures the average ranking difference between the two exams. In the data, we excluded students with score zero in one of the analyzed exams for a fair comparison. In Table 8 it is possible to observe the differences determined by the two rankings considering two subjects, Maths and Writing.

It is important to observe that the ENEM ranking does not match that of UFRGS. We can observe that a maximum difference $\max_i |d_i|$ (in Table 8) is close to the maximum possible difference (N_{total}). The histogram of the rank differences, i.e., d_i, $i = 1, ..., N_{total}$, can be observed in Figure 6.

Although the differences d_i are distributed around zero, we can observe that the standard deviation of $|d_i|$ is very large according to Table 8. The average difference in Maths, considering the three years for example is around $3,550$ positions which is a very large difference when one considers that ENEM will be used as a national exam. In order to understand the coefficient α we performed a simple numerical simulation. Basically we consider N_{total} numbers in ascending order. We build from this ordered list a

year	$\langle\lvert d_i\rvert\rangle$	$\langle d_i^2\rangle - \langle d_i\rangle^2$	$\langle d_i^2\rangle - \langle\lvert d_i\rvert\rangle^2$	$\max_i \lvert d_i\rvert$	N_{total}	α	β
Math							
2011	3641	4758	3063	19549	21510	16.9%	17%
2012	3501	4589	2964	20267	22651	15.4%	15%
2013	3628	4792	3131	20532	25023	14.5%	14%
Writing							
2011	2834	3581	2188	10315	10559	26.8%	33%
2012	2922	3685	2245	10761	10857	26.9%	37%
2013	3156	4010	2472	11868	12423	25.4%	34%

Table 8: Ranking deviation statistics between ENEM and UFRGS in Maths and Writing

partially randomized list by performing $\lceil \beta N_{total} \rceil$ swaps between pairs of numbers randomly chosen and independently on their positions. Now with this new list in hands we calculate $\langle |d_i| \rangle_{\text{rand}}$. The optimization method finds the best β such that $\langle |d_i| \rangle_{\text{rand}}$ is closest to the $\langle |d_i| \rangle_{\text{real}}$ corresponding to the ranking obtained by the experimental data between two exams (second column in Table 8). A pseudo-code of the algorithm used to find the optimal β, which we call Optimal_Beta, can be seen in Table 9. In this algorithm rand($idum$) is a pseudo-random number and *idum* is the respective seed used to generate the sequence of these numbers. The symbol */ denotes comments in the pseudo-code.

The $\beta-$values are shown in the last columns in Table 8. There is a clear correspondence between α and β which corroborates the definition used to measure the difference between the two rankings.

We claim that all factors previously raised with respect to ENEM, such as the size of question statements and the duration of the exam, allow for less prepared students to obtain similar scores to that of well prepared students who have more comprehensive knowledge. This is corroborated by the statistics related to score distribution: such statistics show an apparent homogenization of the evaluation system process when it actually should separate well-prepared students from the other candidates.

4. Conclusions

Standardized university entrance exams have been employed in many countries. They are typically characterized by a Gaussian score distribution. In this paper we analyzed one particular standardized test, the Brazilian's ENEM examination.

We found out that unlike the students' scores distribution, the schools' scores do not follow a Gaussian curve, but form a **distribution with a main peak followed by a lower "hump"** which is best fitted by an

Procedure Optimal_Beta ($\beta_{\min}, \beta_{\max}, N_{total}, \langle|d_i|\rangle_{\text{real}}, \Delta\beta$)

```
    input: βmin, βmax, Ntotal, ⟨|di|⟩real
    output: βopt
    Vector: v[i = 1, ..., Ntotal]
*/  Initializations:
    Δ = Ntotal^2 (or other big number of your choice)
    For i = 1, ..., Ntotal
      v[i] = i
    Endfor
*/  Span β from βmin up to βmax with precision Δβ:
    For β = βmin, βmax; Δβ
    For icount = 1, ⌈β Ntotal⌉
        i := rand(idum) ∗ Ntotal + 1
        j := rand(idum) ∗ Ntotal + 1
*/  Perform the swap!
        aux := v(i)
        v(i) := v(j)
        v(j) = aux
    EndFor
*/  Compute ⟨|di|⟩rand, i.e, the average distance between the
*/  randomized list and ordered Ntotal numbers;
     For i = 1, Ntotal
       ⟨|di|⟩rand = ⟨|di|⟩rand + |i − v(i)|
     Endfor
       ⟨|di|⟩rand = ⟨|di|⟩rand /Ntotal
       temp := |⟨|di|⟩rand − ⟨|di|⟩real|
          If (temp < Δ) then
             βopt := β;
            Δ := temp;
          Endif
    Endfor
    Return βopt
    Stop
    End
```

Table 9: Procedure for computing the β index

EC distribution. This reflects the fact that the average school scores increase linearly with the economic level of the school in two distinct regions with different slopes. This indicates that the exam is designed to identify skills that are more commonly developed in the economic advantaged schools of the country. One possible explanation is the that since the exam is very long, it demands that test takers should be trained to spend hours focusing on one specific topic, which is a kind of preparation that only more expensive schools are typically able to provide.

Since the schools scores distribution is not a Gaussian, it cannot be used to compare the schools performance over time since it is not a standardized measure. Next, the score of the students in the ENEM and UFRGS exams were compared. The correlation between different subjects in both cases was observed. Surprisingly, the correlation between Human Sciences and Natural Sciences and Maths is quite high in the case of the ENEM, which suggests that the exam is not assessing for the specific abilities in the different subjects.

Since the ENEM's scores and the UFRGS's scores follow very different distributions, the change from one standardized test to a more itemized exam both impacts and implies selecting a different student profile. In summary, we employed statistical methods to understand the characteristics of the selection by two exams, one standardized and a non-standardized exam. Our results indicate that there are differences in the selection/admission of students when each one of these exams is considered. Finally, it would be interesting to compare the effectiveness and the long term impact of the use of these exams. One could consider an analysis of the college performance of each group of students and perhaps to investigate the professional performance of future graduates who were admitted by each of these methods [21, 22].

Acknowledgment

We would like thank the Brazilian agencies CNPq, INCT-FCx, and CAPES for the financial support. We also thank the fruitful discussions with Prof. Fernando Lang da Silveira.

References

[1] P.B. Ebrey, *The Cambridge Illustrated History of China* (Cambridge: Cambridge University Press) (2010).

[2] Mark W. Huddleston, William W. Boyer, *The Higher Civil Service in the United States: Quest for Reform*, (Pittsburgh: University of Pittsburgh Press) (1996).

[3] Neil J. Dorans, *Recentering and Realigning the SAT Score Distributions: How and Why*, Issue Journal of Educational Measurement Journal of Educational Measurement **39**, 59 (2006).

[4] Daniel Koretz,*The Impact of Score Differences on the Admission of Minority Students: An Illustration*, National Board on Educational Testing and Public Policy Carolyn A. and Peter S. Lynch School of Education Boston College **1**, Number 5 (2000).

[5] G. Corrado, *Rethinking the SAT: The future of standardized testing in university admissions*, Teachers College Record **106**, 1526 (2004).

[6] K. Yang *A Review of Foreign College Entrance Examinations Focusing on Scoring Procedures*, The Journal of Curriculum and Evaluation **10**, 145 (2007).
[7] Steven A. Culpepper and Ernest C. Davenport, *Assessing Differential Prediction of College Grades by Race/Ethnicity with a Multilevel Model*, Jornal of Educational Measurement **46**, 220 (2009).
[8] Kateryna Kokhan, *An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses*, Language Testing **30**, 467 (2013).
[9] S. M. R. Ting, *Predicting academic success of first-year engineering students from standardized test scores and psychosocial variables*, International Journal of Engineering Education **17**, 75 (2001).
[10] Fernando Lang da Silveira, Jurgen Stilck and Marcia C. Barbosa, *Manifesto sobre a qualidade das questõess de Física na Prova de Ciêcias da Natureza no Exame Nacional de Ensino Médio*, Cadernos Brasileiros de Ensino de Física **31**, 471 (2014).
[11] Fernando Lang da Silveira, Marcia C. Barbosa and Roberto da Silva, *Exame Nacional de Ensino Médio (ENEM): Uma análise Crítica*, Revista Brasileira de Ensino de Física **37**, 1001 (2015).
[12] M. Gligor and M. Ignat Ignat, *Scaling in the distribution of marks in high school*, Fractal-Complex Geometry Patterns and Scaling in Nature and Society **11**, 363-368 (2003).
[13] T. Leonard and J. S. J. Hsu, *Bayesian-Indference for a Covariance-Matrix*, Annals of Statistics **20**, 1669-1696 (1992).
[14] B. Neelon, A. E. Gelfand and and Marie Lynn Miranda, *A multivariate spatial mixture model for areal data: examining regional differences in standardized test scores*, Royal Statistical Society: Appl. Statistical **63**, 737-761 (2014).
[15] http://veja.abril.com.br/educacao/ranking-escolas-brasil-2013/ranking-escolas-brasil-2013.shtml
[16] R. da Silva, L. C. Lamb, E. C. Lima, J. Dupont, *A simple combinatorial method to describe particle retention time in random media with applications in chromatography*, Physica A, **391**, 1-7 (2012).
[17] R. da Silva, L. Brusamarello, G. Wirth, *Statistical fluctuations for the noise current from random telegraph signals in semiconductor devices: Monte Carlo computer simulations and best fits*, Physica A **389**, 2687-2699 (2010).
[18] R. Silva; S. R. Dahmen, *Universality in the distance between two teams in a football tournament*, Physica A,**398**, 56-64 (2013).
[19] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, *Numerical Recipes in Fortran 77: The Art of Scientific Computing* (1992).
[20] K. S. Trivedi, *Probability and Statistics with Realiability, Queuing, and Computer Science and Applications*, (New York: Wiley) (2002).
[21] Ana Carolina Zoghbi, Fabiana Rocha, Enlinson Mattos,*Education production efficiency: Evidence from Brazilian universities*, Economic Modelling, **31**, 94-103 (2012).
[22] H.M. Gupta, J.R. Campanha, F.R. Chavarette, *Power Law Distribution in Education: Effect of Economical, Teaching and Study Conditions in University Entrance Examination*, International Journal of Modern Physics C, **14**, 449-457 (2002).

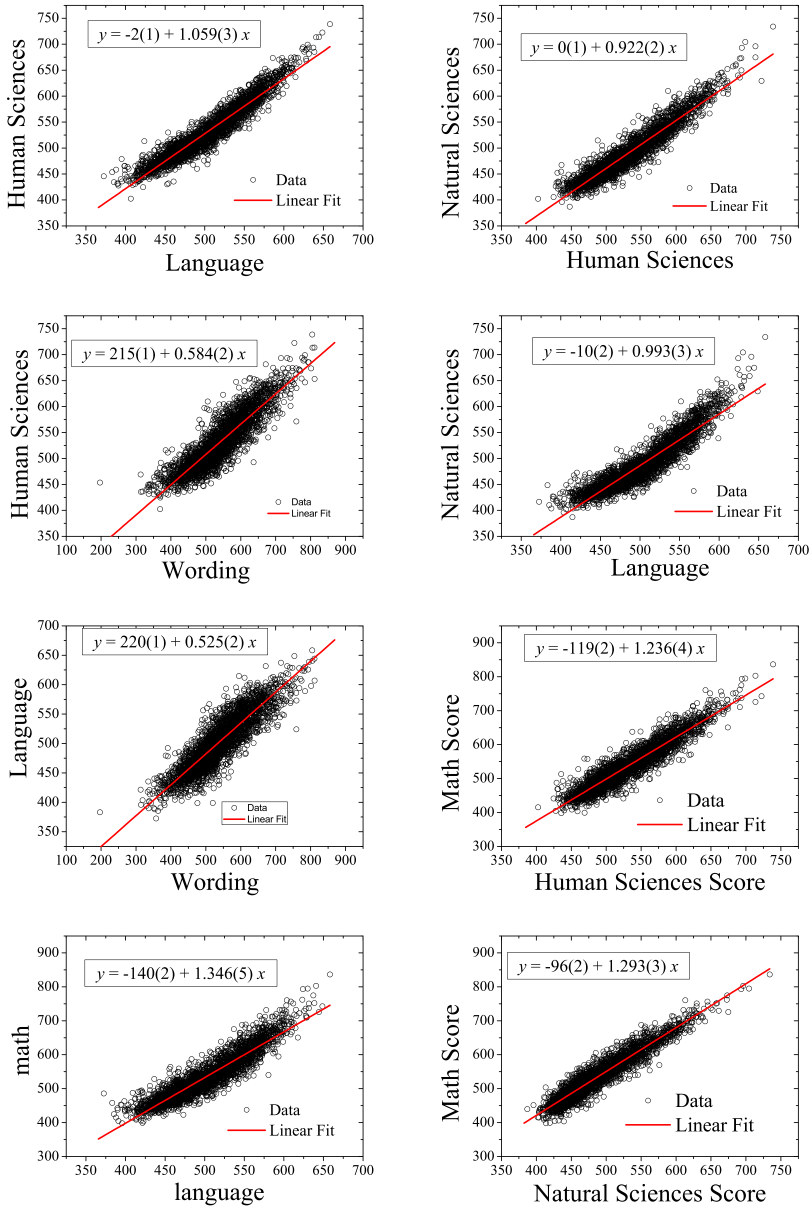

Figure 1: Scattering diagrams for different pairs of high school subjects. Visually, we can observe a good correlation (pairwise) between them.

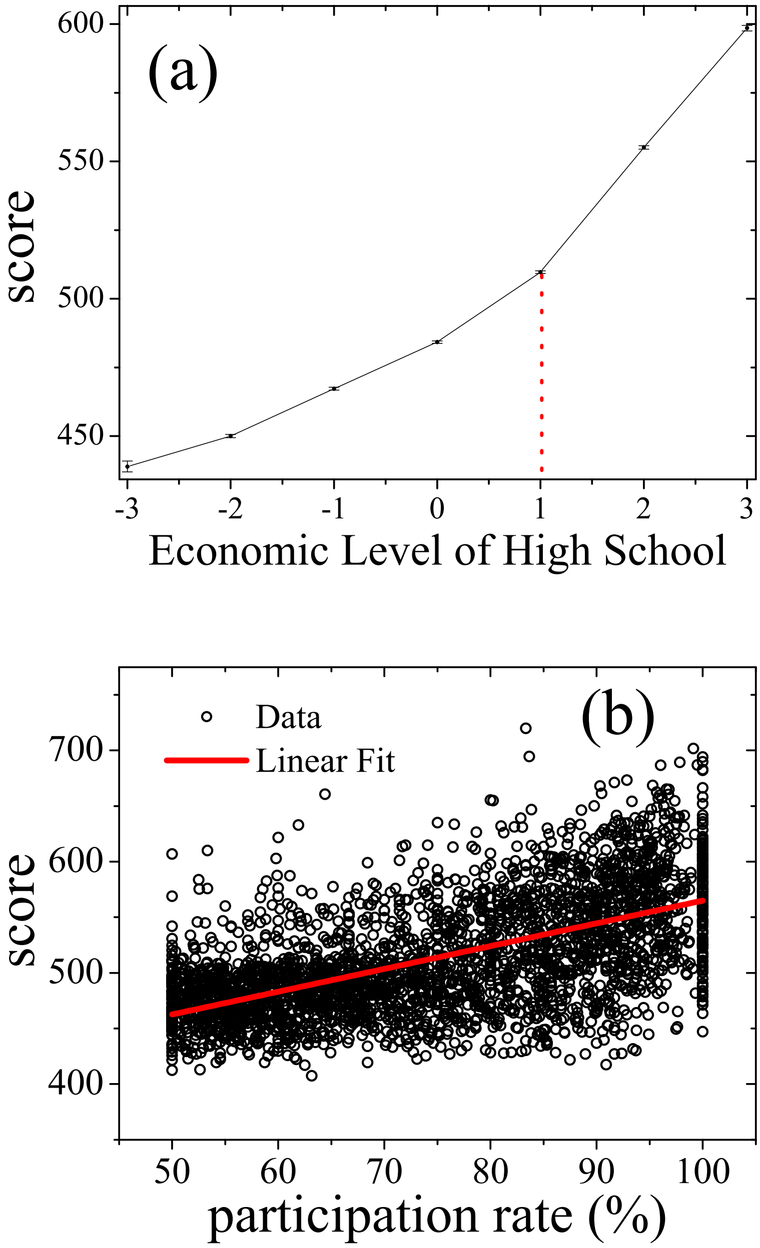

Figure 2: (a) Average score as a function of the school socioeconomic level from lower to upper level. (b) Score as function of students' participation.

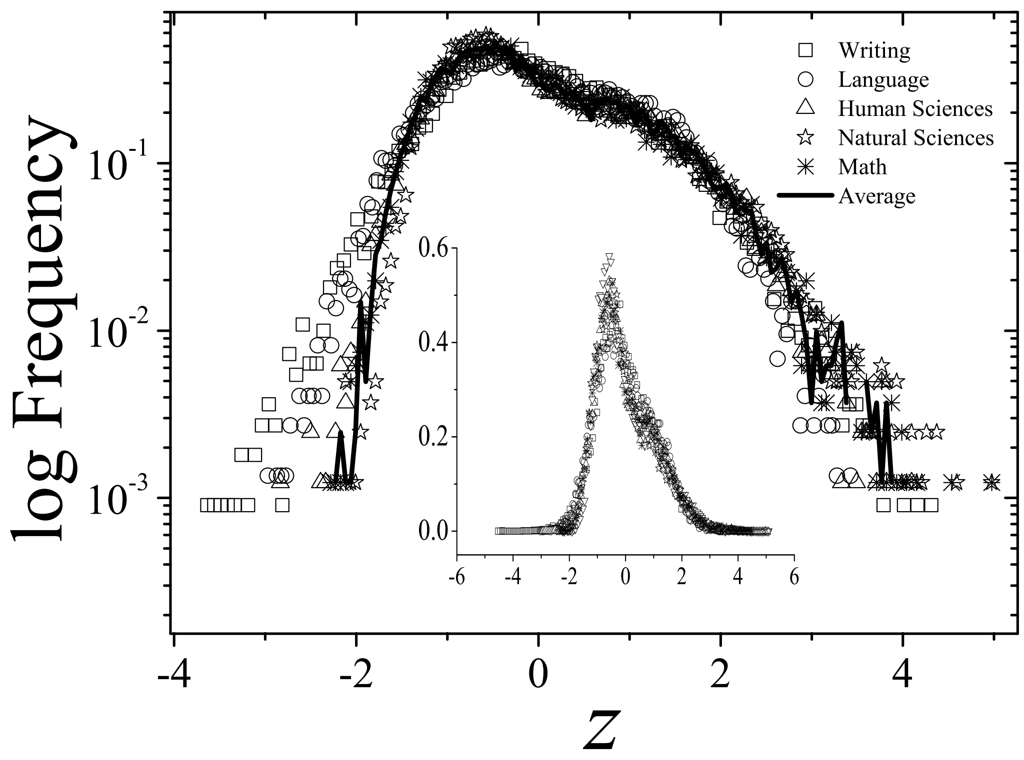

Figure 3: Distributions of the scores for each subject in mono-log scale. All the subjects follow in the same curve. The inset plot represents the same data in a linear scale

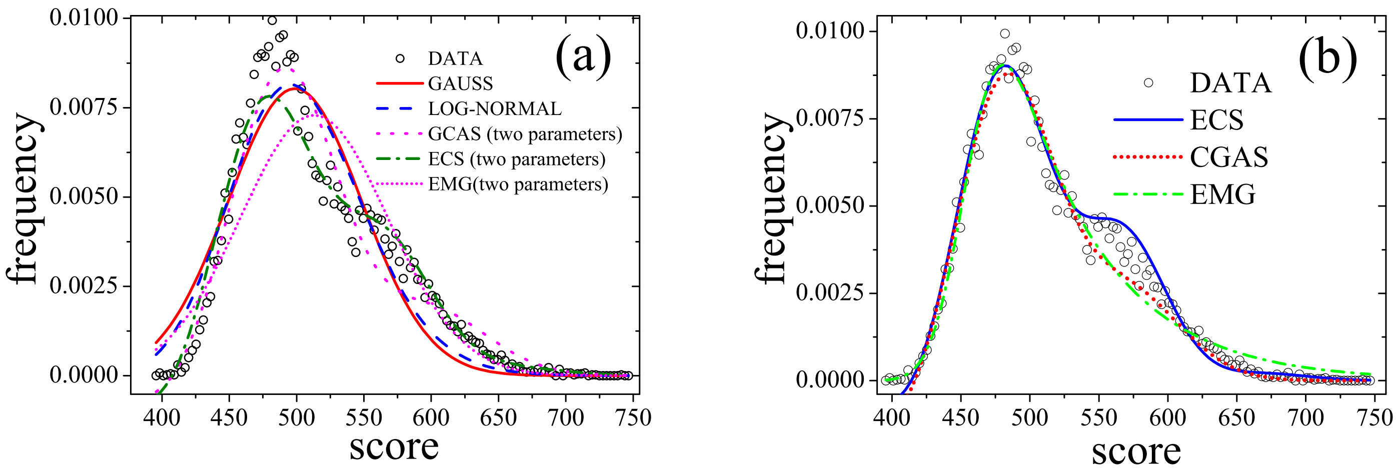

Figure 4: Fits of the data using (a) two parameters distributions and (b) three or four parameters distributions.

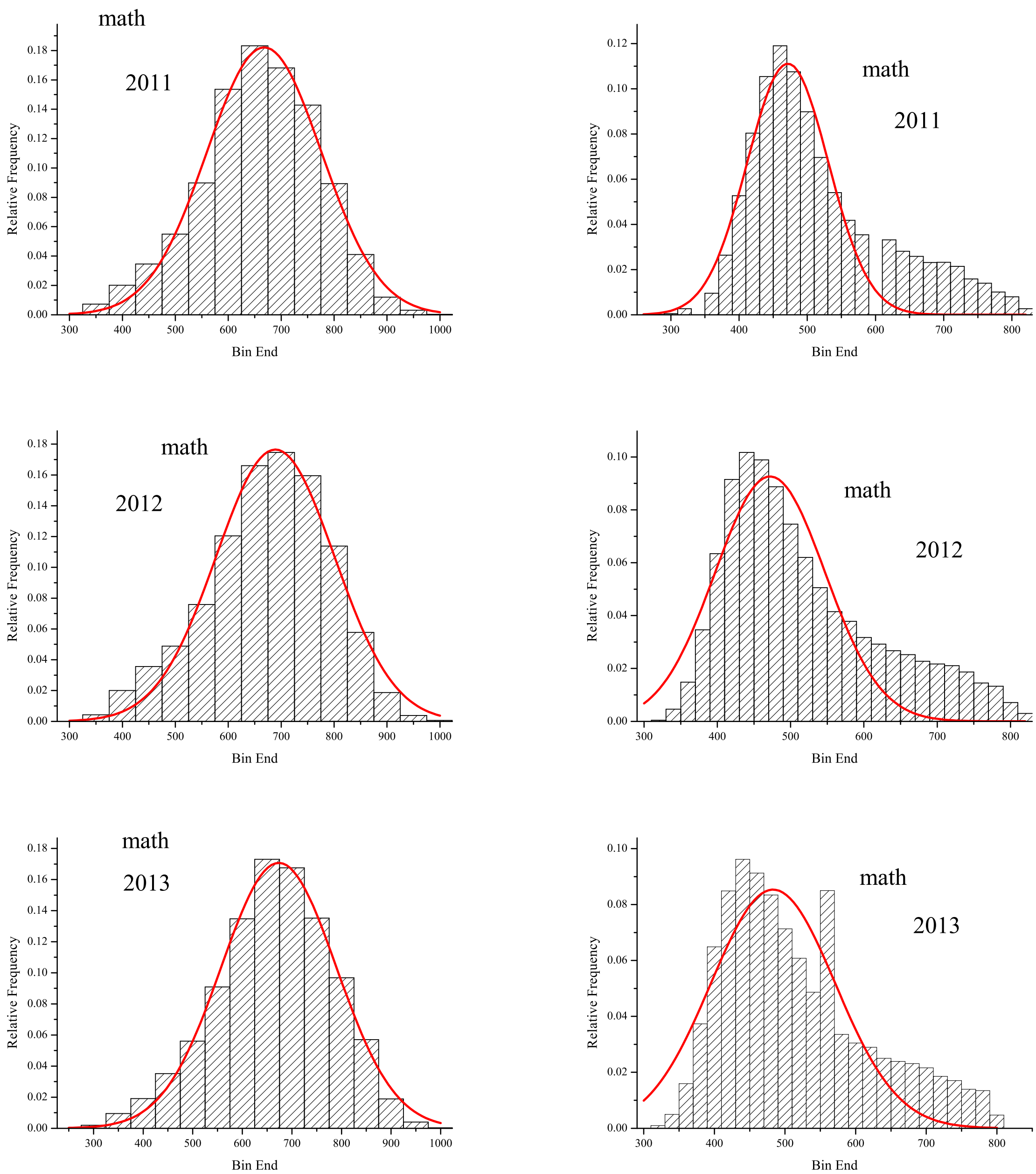

Figure 5: Score distribution of the same candidates in UFRGS and ENEM for Mathematics. The continuous curves correspond to Gaussian fits. We can observe a deep difference in the right histograms (UFRGS) in comparison with the left histograms (ENEM).

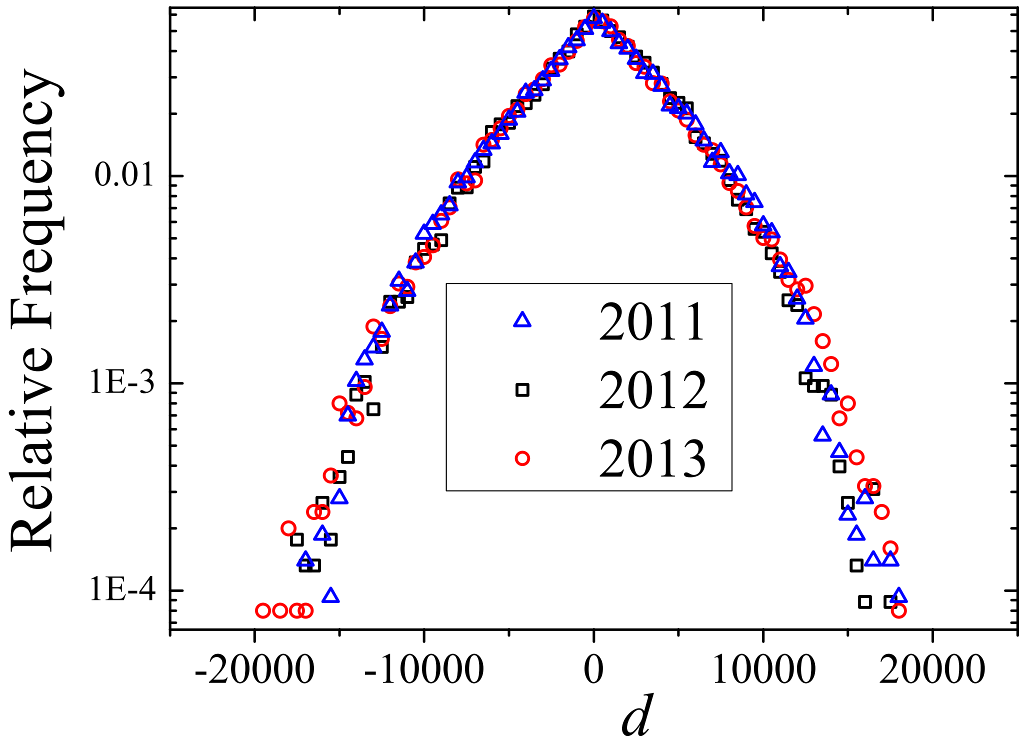

Figure 6: Histogram of ranking differences between UFRGS and ENEM for Maths in mono-log scale. A universality is observed under different years analyzed.