Research Article https://doi.org/10.12973/eu-jer.12.3.1495

European Journal of Educational Research
Volume 12, Issue 3, 1495 - 1508.
ISSN: 2165-8714
https://www.eu-jer.com/

Similarities and Dissimilarities in Student Grades Distributions, Over Time and by Gender

Pedro Ferreira* University of Lisbon, PORTUGAL
Luísa Canto e Castro University of Lisbon, PORTUGAL
Carina Silva University of Lisbon, PORTUGAL

Received: December 7, 2022 ▪ Revised: March 4, 2023 ▪ Accepted: April 18, 2023

Abstract: The focus of this article is to analyze the distribution patterns of student grades over time for different subjects and by gender. Specifically, we examined the final term grades of upper secondary students in Portuguese public schools across four subjects (Mathematics, Portuguese Language, Philosophy, and Physical Education) from the academic years 2013-2014 to 2017-2018. These grades reflect the teachers' perceptions of the students' knowledge gained throughout the academic year. We expected to see some regularity in the grade distributions over time for a particular subject. However, we found that the similarity of grades across subjects and time was so striking that differences were barely noticeable by visual inspection. Due to the very large sample sizes (in the order of tens of thousands), the quantification of similarities and dissimilarities was done through distribution's proximity statistics and not by classic statistical methods, like Chi-Square or comparison of means tests. Additionally, we applied a methodology of multiple equivalence tests to globally compare the relative frequencies of each of the grades in pairs of independent samples. Our analysis showed that there was a high level of similarity in grades for the same subject over time, but we also found differences between subjects and between genders.

Keywords: *Distribution's proximity statistics, equivalence testing, gender disparity, student grades.*

To cite this article: Ferreira, P., Canto e Castro, L., & Silva, C. (2023). Similarities and dissimilarities in student grades distributions, over time and by gender. *European Journal of Educational Research, 12*(3), 1495-1508. https://doi.org/10.12973/eu-jer.12.3.1495

Introduction

The data for this study were provided by the Portuguese agency responsible for collecting and processing education data, the General Directorate for Education and Science Statistics. The data analyzed were grades obtained by upper secondary students enrolled in non-vocational education and training programs (non-VET) in Portuguese high schools. The programs in these schools consist of three curricular years, the 10th, 11th, and 12th, with three school terms per year. At the end of each term, students receive a grade for each subject they are enrolled in. These grades are given as an integer value from 1 to 20, and the final term grades represent a summary of the grades for the three terms and reflect the teachers' perceptions of the students' overall achievement.

This study analyzed the final term grades for four subjects - Mathematics, Portuguese Language, Philosophy, and Physical Education - over the last five academic years (2013-2014 to 2017-2018). The data were stratified by curricular year and by gender, and the sample size included around 70,000 grades per year for subjects that all students took and around 10,000 grades for optional subjects. The number of boys and girls in the sample was similar. It is important to mention that the total database for this study covered a longer period of academic years (2007-2008 to 2017-2018) and a wider range of subjects (including History, Biology-Geology, Physics-Chemistry, Economy, Descriptive Geometry and Drawing), but this analysis only considered the four subjects mentioned above.

Based on these data, the distribution of the students along the grading scale (from 1 to 20) was plotted for each academic year and subject. Figures 1 and 2 below show the regularity of the distribution patterns when we hold the subject constant.

* **Corresponding author:**
Pedro Ferreira, Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), Portugal. ✉ fc57220@alunos.fc.ul.pt

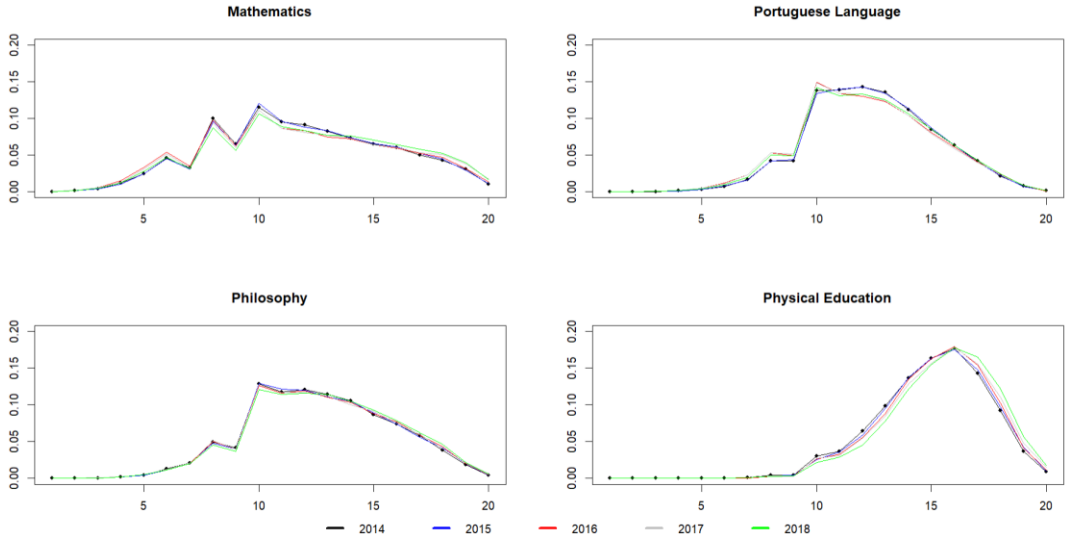

Figure 1. Grade Distributions From 2014 to 2018, for the Four Subjects of the 10th Curricular Year.

Figure 1 shows the grade distributions for the 10th curricular year for Mathematics, Portuguese Language, Philosophy, and Physical Education for the academic years 2014 to 2018. It is important to note that the students represented in these graphs are different in each year, and some of the teachers may also be different, although it is likely that many remain the same. The regularity observed in the grade distributions for each subject over the years raises some questions. Is the pattern in the distributions defined by the teachers beforehand? Or is it the result of students adapting to the nature of the subjects? It is even possible that this regularity could be predicted, but it is currently unknown. Across the years, not only does the mean remain approximately constant, but all other distributional characteristics such as dispersion, bias, and local peaks also remain unchanged and are not visibly different to the naked eye.

Initially, we might expect that the pattern of grade distributions would be similar across subjects. However, as shown in Figure 1, this is not the case (for example, the grade distributions for Mathematics and Philosophy are quite different). This suggests that there is a regularity in the way that students solidify their knowledge in different subjects, but that this process depends on something intrinsic to the subject itself. At first glance, this seems to be the most plausible hypothesis.

It is worth noting that it is highly unlikely that this regularity would be observed with the grades from a single test or exam. The final term grades represent the teacher's perception of the student's overall level in the subject taught throughout the year, and this perception is likely to be more reliable than a single test. In fact, another finding from the analysis of these data is that teachers as a group show a consistently strong ability to evaluate students over time. This may be due to the fact that many teachers in Portugal have over twenty years of experience, which gives them a clear understanding of evaluation standards.

The pattern of grade distributions by gender is another level of regularity that warrants further discussion and consideration. As shown in Figure 2, which depicts the grade distributions for students in the 10th curricular year in Portuguese Language and Physical Education stratified by gender and comparing the academic years 2014 to 2018, there is a regular pattern over time for each gender within the same subject. However, it is important to mention that there is a gender disparity in these distributions. This raises questions about potential factors that may contribute to this difference and the potential consequences of this discrepancy.

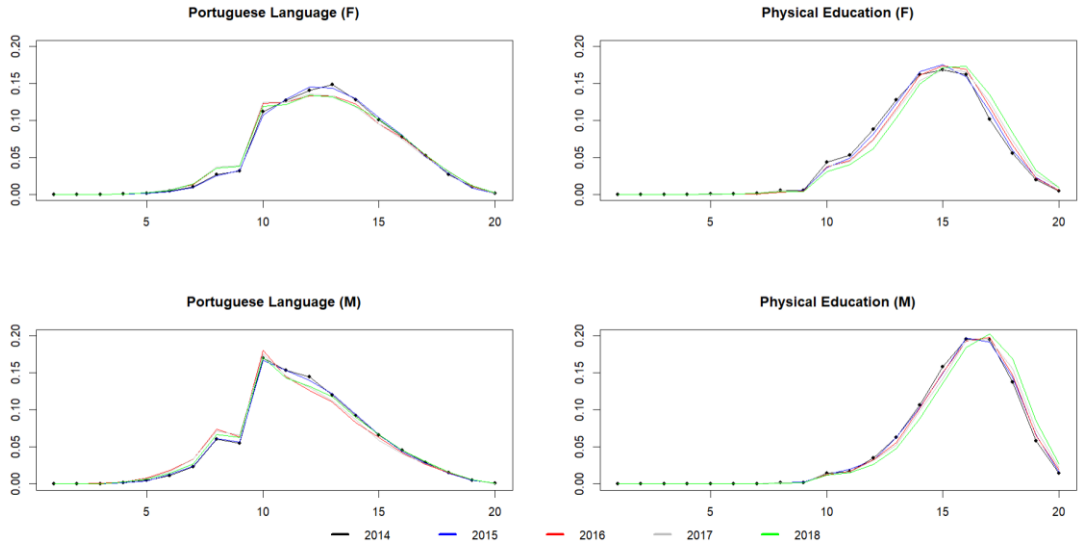

Figure 2. Grade Distributions, by Gender (Female (F) and Male (M)), by Academic Years 2014 to 2018, for Portuguese Language and Physical Education of the 10th Curricular Year.

When we examine the grade distributions separately for boys and girls, we can see that the pattern is repeated within each subgroup but not between them. In fact, for most subjects, the pattern is similar to that shown in Figure 2 for Portuguese Language: there is a higher incidence of grades in the range of 12 to 15 values in the subgroup of girls compared to boys. For boys, there is a greater incidence of the grade 10 (the first positive mark) and a notable decrease in the following grades. This suggests that there may be underlying factors that contribute to these differences in grade distributions by gender. One consequence of the higher weight of grades in the range of 12 to 15 compared to the lower weight of grades in the range of 10 and 11 is that the average final term grades are almost always higher in the subgroup of girls compared to the subgroup of boys. The exception is Physical Education, where boys have, on average, better results.

The phenomenon of regularity in the distribution of subject grades has been noted in previous studies, such as Ma (2001), which analyzed the stability of each student's grades throughout the school year. The stability of exam averages by subject is also highlighted in reports by the Portuguese agency responsible for conducting national exams, the Institute for Educational Assessment (IAVE). Gender disparities in grade distributions have also been studied, such as in the work of Meinck and Brese (2019). Additionally, Lewin (2021) recently published a study on the development of parametric models for the distribution of student grades.

In the field of education, the analysis of how student grades are distributed along the evaluation scale has focused primarily on two themes: grade inflation and its consequences for equity in educational progression and access to higher-paying professions, as in Griffin and Townsley (2021), and gender discrepancies, specifically differences by discipline and their reflection in subsequent career paths, as examined by Workman and Heyder (2020) and in the extensive literature review on the topic included therein. Regarding gender differences, O'Dea et al. (2018) analyzes the amplitude and variability of differences with the goal of finding an explanation for the fact that fewer women than men pursue careers in science, technology, engineering, and mathematics (STEM) despite girls' better performance in relevant subjects as students.

The data we will use in this study allows us to quantify gender differences in each of the four disciplines considered (Mathematics, Portuguese, Philosophy, and Physical Education), enabling comparison with other published results. However, the phenomenon we wish to highlight here, which to our knowledge has not yet been addressed for school summative evaluations, is the regularity of distribution patterns year after year, discipline by discipline.

To address some of the questions raised by the observed patterns in grade distributions, it is necessary to understand whether there are indeed patterns in the distribution of student grades over the years for the same subject, whether there are subjects where girls perform better than boys and vice versa, and which group of subjects follows the same distribution pattern. A statistical comparison of grade distributions can be a powerful tool to answer these questions. It can help to identify any trends or patterns in the data and to determine whether any observed differences are statistically significant. This can provide insight into the factors that may be influencing the grades and can inform efforts to improve student performance and address any disparities that may exist.

There are several statistical approaches that can be used to compare distributions, such as using statistical tests like the Chi-Square test for population homogeneity or tests based on the comparison of means. However, these classic statistical tests are not useful in this case because the sample sizes are too large and even very small differences can lead to the rejection of the null hypothesis. Instead, we used similarity measures and developed a procedure based on a statistical

test called an equivalence test, which is able to declare the absence of a meaningful effect. This approach allowed us to quantitatively assess the similarity or difference between the grade distributions being compared.

Methodology

Research Design

This is a large-scale exploratory observational study, whose main objective is to quantify the degree of similarity between pairs of distributions of grades assigned by teachers in four of the Portuguese curriculum subjects over a period of five academic years, from 2013-2014 to 2017-2018. More specifically: (a) given the subject and curricular year, the academic year 2013-2014 is used as a reference and each of the subsequent year's distributions is compared to that year's distribution; (b) given the curricular year (10th, 11th, or 12th) and academic year, the distribution of grades for each subject is compared to that of the others; (c) given the subject and academic year, the 10th curricular year is used as a reference and each of the subsequent curricular year's distributions (11th and 12th) is compared to that curricular year's distribution; (d) given the subject, curricular year, and academic year, the distribution of grades obtained by girls (reference) is compared to the distribution of grades obtained by boys.

Sample and Data Collection

The data involved in this study represents the population of students enrolled in non-VET courses in public upper-secondary education in Portugal during the academic years 2013-2014 to 2017-2018 and was collected by the General Directorate for Education and Science Statistics. More precisely, it encompasses all students, with the exception of those enrolled in isolated subjects. In each academic year, the sample involves the grades of about 200,000 students, distributed among each of the three curricular years that make up secondary education in Portugal (approximately 73,000 students in the 10th year, approximately 66,000 in the 11th year, and approximately 63,000 students in the 12th year). The range of disciplines considered here is limited to the three that are mandatory for all students (Portuguese Language, Philosophy, and Physical Education) and the Mathematics discipline, which is mandatory only for students enrolled in the Science and Technology and Socio-Economic Sciences courses. The grades will be analyzed in terms of how they are distributed along the evaluation scale, which in Portugal ranges from 1 (the lowest) to 20 (the highest).

Analyzing of Data

There are several possible statistical approaches to compare distributions. Parametric methods, such as, comparison of means or nonparametric methods, such as, two sample Kolmogorov-Smirnov tests are not very useful here due to sample sizes and the multiplicity of tests given the number of possible comparisons to be made. This is why the quantification of the degree of proximity between the distributions of student grades will be conducted using the following three measures: (a) Hellinger distance; (b) Overlapping index (OVL) and (c) Area Under the ROC Curve (AUC), which will be described in detail below. Additionally, in order to be able to declare that the relative frequencies of a certain grade are sufficiently close in pairs of distributions, we will use the Two One-Sided Tests (TOST) methodology. Since each case involves the application of 20 TOST equivalence tests, the global assessment will use one of the usual procedures in multiple testing.

Hellinger Distance

The Hellinger distance between two discrete probability distributions measures the distance between them in a common space. This measure can assume values between 0 and 1. However, for a certain distribution, the distance between itself and the most distant distribution isn't reflected in a Hellinger value equal to 1. Hellinger (1909) introduced this measure and defined the distance between two distribution, $p(x)$ and $q(x)$, as:

$$H(p,q) = \frac{1}{\sqrt{2}} ||\sqrt{p} - \sqrt{q}||_2$$

where $||.||_2$ is the L2 norm and $\frac{1}{\sqrt{2}}$ ensures that $H(p,q) \leq 1$, for any pair of distributions. This measure is usually associated with the Bhattacharyya coefficient (Cieslak et al., 2012):

$$H(p,q) = \sqrt{1 - BC(p,q)}$$

where $BC(p,q) = \sum_{x \epsilon X} \sqrt{p(x)q(x)}$ is the Bhattacharyya coefficient.

Overlapping Index (OVL)

According to Pastore and Calcagni (2019), the overlap can be defined as the area of intersection of two or more probability functions and suggests a simple way to quantify the difference, or similarity, between samples or populations, which are described in terms of distributions. Intuitively, two populations (or samples) are similar when their distribution functions overlap. The simplicity of the overlap concept makes the use of this index particularly suitable for

many applications, such as the comparison of probability distributions, exploring the amount of common area in the same domain. Furthermore, overlap can be used as a measure to estimate distances between clusters or, alternatively, to measure similarities between datasets. The measure used to describe overlap is the overlap index (OVL). This coefficient can take values between 0 and 1, where values close to 1 mean a large area of overlap. For this reason, the OVL cannot be considered a distance, since, considering two points x and y, when $x = y$, the OVL has the value of 1, while a distance would have a value of zero. The OVL can detect differences between two distributions, not only by differences in mean value, but also by differences in variance. Another advantage is that the overlap index does not change when there are monotonous scale transformations of the variables. Weitzman (1970) proposed the following expression to determine the OVL between two distributions, $p(x)$ and $q(x)$:

$$\text{OVL} = \int_C min[p(c), q(c)]\, dc.$$

The equation above is intended for continuous distributions. For discrete distributions it is necessary to replace the integral by summation:

$$\text{OVL} = \sum_{x \epsilon X} min[p(x), q(x)].$$

This index is highly versatile and applicable in a range of practical scenarios. Notably, it does not rely on any strict distributional assumptions, which makes it an excellent tool for measuring differences between samples or populations that are characterized by various types of distributions.

Area Under the ROC Curve (AUC)

The Receiver Operating Characteristic (ROC) analysis began to be developed in the 50s, during the 2nd World War, based on statistical decision theory. This methodology was created to evaluate the detection of signals in radars and later in several other areas, such as medicine, for the analysis of the discriminative performance of diagnostic tests. To apply this analysis to the comparison of two distributions we start with the definition of the ROC curve. Given two distribution functions, $F(x)$ and $G(x)$, where F is the reference and G is the distribution function that we want to compare with F, Jensen et al. (2000) defined the ROC curve by

$$ROC(x) = G(F^{-1}(x)), 0 \leq x \leq 1$$

where F^{-1} is the inverse function of F and the probability quantile, x, of the F function is defined as:

$$F^{-1}(x) = inf\{y \, \epsilon \, S(F) : F(y) \geq x\}, 0 \leq x \leq 1$$

where

$$S(F) = \{x \, \epsilon \, R : 0 < F(x) < 1\}$$

is the support of F. When F and G coincide, we get $ROC(x) = x$, $0 \leq x \leq 1$, which corresponds to the diagonal of the unit square.

Graphically, the ROC curve will be above the diagonal when F assumes bigger values than G, that is to say, when the values of the underlying random variable of F are shifted to the right when comparing to the values of the underlying random variable of G and ROC curve will be under the diagonal, otherwise. The area under the ROC curve (AUC) is the most used index in the ROC methodology to describe the ability to discriminate a model, a diagnostic test or two populations. This index ranges from 0.5 to 1. However, in this case we'll also consider values between 0 and 0.5, since for the purpose of this work they are useful to the conclusions we want to get. So, values close to 0 or close to 1 mean a large difference between distributions. AUC values around 0.5 are obtained for identical distribution functions and greater (or lower) values than 0.5 mean that there is a shift to the right (or left) of F density when compared to the density of the reference G. For example, an AUC of 0.55 tells us that the dissimilarity between the distribution of sample 1 and the distribution of sample 2 (reference) is small and that sample 2 (reference) has higher values than sample 1. However, the distance between the distributions are the same if the AUC is 0.45, we just swapped the reference sample, that is: if the AUC < 0.5, the sample 1 appears to have higher values than the sample 2; if AUC > 0.5, sample 2 seems to have higher values than sample 1. There are several approaches for the determination of AUC, both through parametric and non-parametric methods. Non-parametric methods do not need distributional assumptions, however, they have the effect of losing efficiency. In this work was considered a parametric approach to estimate the AUC using the binormal model. Let's assume that X and Y are independent and that $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$. The parametric estimator of AUC obtained through this model is given by Faraggi and Reiser (2002):

$$\widehat{AUC} = \Phi\left(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right).$$

An advantage of the binormal AUC is that this model has the power of smoothing the ROC curve when variables are ordinal or discrete, not overestimating the value of the AUC.

Equivalence Test (TOST)

The Two One-Sided Test (TOST) is an equivalence test that is based on the classical t test, used to test the hypothesis of equality of means. Based on two independent samples, it is intended to prove that there is a statistical difference between the samples means. The interest is to see if the means of two samples are really equivalent, that is, if they differ at most by a certain margin, M. Let X and Y be two independent variables. We assume that n_x observations belong to sample 1 and n_y to sample 2. Schuirmann (1987) defined the equivalence test hypotheses as follows:

$$H_0: (\mu_x - \mu_y) < -M \; or \; (\mu_x - \mu_y) > M$$

$$vs$$

$$H_1: -M < (\mu_x - \mu_y) < M$$

that is, the non-rejection of the null hypothesis points to a significant difference (depending on the margin) between the samples, which differs from a classic test of equality of means. This joint null hypothesis originates two one-sided hypotheses:

$$H_{0_1}: (\mu_x - \mu_y) < -M$$

$$vs$$

$$H_{0_2}: (\mu_x - \mu_y) > M$$

If both one-sided tests are rejected at a given significance level α, the null hypothesis H_0 is rejected as well. This strong statistical decision to reject implies that the two means can be considered equivalent, meaning that their difference is smaller than the specified equivalence margin.

In the application of TOST there is a subjective component, which is the equivalence margin (M) - according to Lakens et al. (2018) it is called smallest effect size of interest (SESOI). In this paper, as we are going to compare 20 pairs of proportions (the proportion of students that obtained the grade "x" in subject T and in subject C, for x = 1, 2, ..., 20), in each of the comparisons we will set a SESOI that involves a classic test of proportions comparison.

The Chi-Square statistic, in situations of moderate-sized samples, does not detect small differences and, therefore, we chose to set this small difference as the one that the Chi-Square statistic does not identify in samples of size 100. In fact, it can be proven that, at a significance level, α, Chi-Square test only detects differences such that $|(\hat{p} - p)| > \sqrt{\chi_\alpha \frac{p}{n}}$. Thus, we have defined the equivalence margin, M, as follows:

$$M = \sqrt{\frac{\chi_\alpha p_i}{n}}$$

where p_i is the proportion of the category i in the joint sample and χ_α is the critical point of the distribution χ^2_α for an α to be defined. Consequently, the lower margin is $-M = -\sqrt{\frac{\chi_\alpha p_i}{n}}$.

This means that the equivalence margins, M, will vary throughout the tests that will be carried out for each category.

As said, we will apply the previous test to the proportions (relative frequencies) of each grade, using a significance level of 10%. Then, for each comparison we will summarize the conclusions of the multiple tests by showing the total number of rejections. As a rule of thumb, we can consider that:

- above 4 non-rejections of the null hypothesis in the multiple tests, the distributions are significantly apart;
- above 17 rejections of the null hypothesis, we consider that the distributions are quite similar.

In order to facilitate the analysis of multiple tests, we used the Benjamini-Hochberg procedure.

Multiple Testing: Benjamini-Hochberg Procedure

According to Benjamini and Hochberg (1995), as input we have the significance level, α, and the p-values, p_1, ..., p_n, of the individual tests. The procedure is as follows:

1. order the p-values: $p_{(1)} \leq ... \leq p_{(n)}$;
2. let $\hat{k}$ the largest $k \geq 1$ such that: $p_{(k)} \leq \frac{\alpha k}{n}$;
3. reject $H_{0,(1)} \leq ... \leq H_{0,(\hat{k})}$ and accept all other hypothesis.

Note: the index $\hat{k}$ is the last time that $p_{(k)} \leq \frac{\alpha k}{n}$.

Results

We present below the comparison analysis performed using grade data from Portuguese upper secondary public schools. These data include the frequencies of each grade obtained (ranging from 1 to 20) for four subjects (Mathematics, Portuguese Language, Philosophy, and Physical Education), academic years (2008/2009 to 2017/2018), curricular years (10th, 11th, or 12th), and by gender. All the analysis was conducted using the R software (4.2.0). To capture the overall proximity of each pair of grade distributions, we used the Hellinger distance measure, which does not consider differences in location but allows us to rank the degrees of similarity.

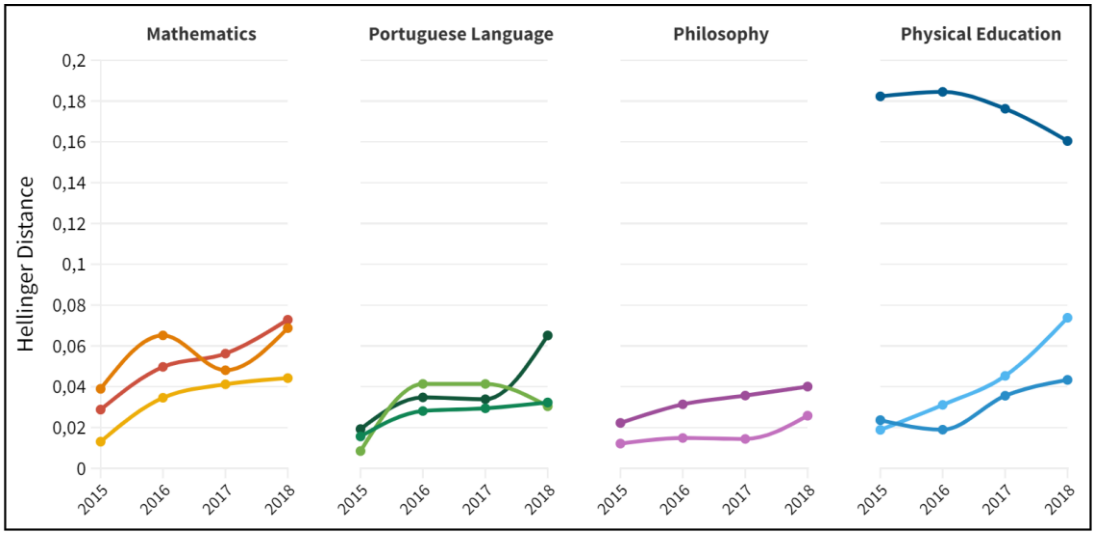

Figure 3. Comparison of Grades Distributions Over Time via Hellinger Distance. The Reference Academic Year Is 2014 and Colors Get Darker as the Curricular Year Progress (10th, 11th and 12th).

The values shown in Figure 3 represent the Hellinger distances between the grade distributions for the four subjects in 2014 and the subsequent years (2015 to 2018). The smaller the values, the more similar the grade distributions are. We used different gradients of the same color to illustrate the differences between curricular years (10th, 11th, and 12th) in the figure.

The subject that shows the strongest statistical regularity over time is Philosophy, particularly for the grades given by teachers in the 10th curricular year. The Hellinger distance in this subject never exceeds 0.04, even though it increases with time. It's important to note that the Hellinger distance values for other subjects are also close to zero, with the exception of Physical Education in the 12th grade. In this case, the Hellinger distance is about 0.2 when comparing student grades in 2014 to those in subsequent years. This can be explained by a policy measure that took effect in 2015, which stopped the grade in Physical Education from counting towards the average grade for higher education access. As a result, teachers started giving slightly lower grades in this subject.

Table 1. Comparing Between Subjects Using Hellinger Distance in Academic Year of 2018 (10th Curricular Year)

	Mathematics	**Portuguese**	**Philosophy**	**Phys. Educ.**
Mathematics	0	0.207	0.170	0.408
Portuguese		0	0.083	0.412
Philosophy			0	0.344
Phys. Educ.				0

As shown in Table 1, the Hellinger distance has much higher values when comparing different subjects than when comparing the same subject over time. In fact, it can be more than twenty times higher when comparing Portuguese Language and Physical Education, or just five times higher when comparing Portuguese Language and Philosophy. These patterns were similar for the other two curricular years and academic years. This indicates that there is a stable subject fingerprint reflected in the distribution of student grades on the evaluation scale. When looking at grades by gender, subject stability is also maintained, but the Hellinger distance can reach values more than ten times higher when comparing grades of boys and girls, as shown in Figure 4.

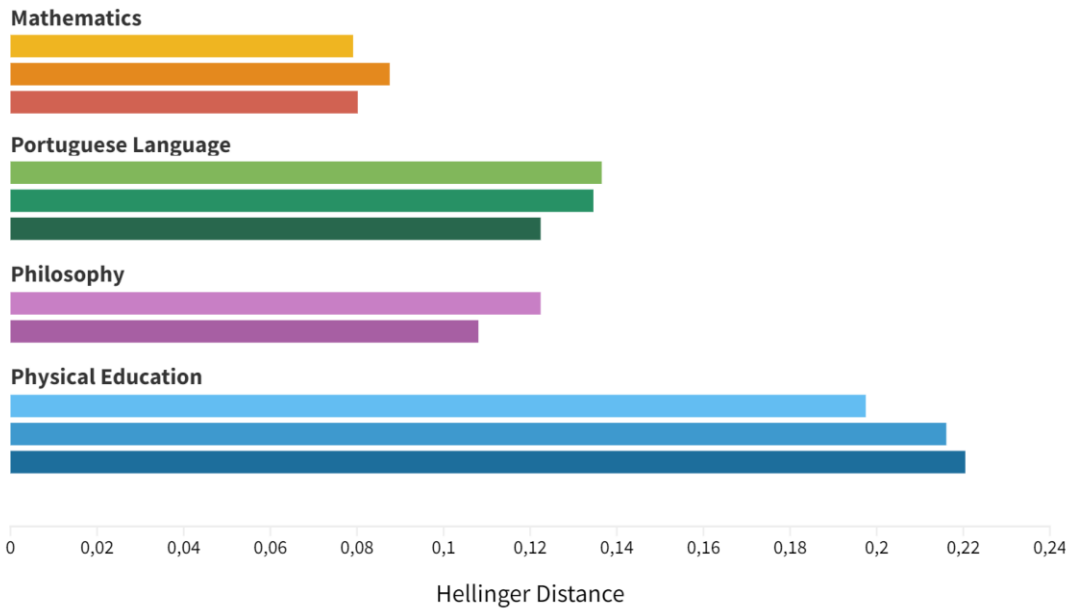

Figure 4. Comparison of Grades Distributions by Gender Via Hellinger Distance. The Reference Gender Is Female and Colors Get Darker as the Curricular Years Progress (10th, 11th and 12th).

Mathematics is the subject with the smallest gender disparity, regardless of the curricular year (Hellinger distance below 0.1), while Physical Education is the subject with the highest gender difference (Hellinger distance above 0.2). Gender disparity in Portuguese Language and Philosophy is also notable, though it narrows as the curricular years progress.

The next graphical representations, known as Arrow Plots (Silva et al., 2020), provide additional information by combining the degree of proximity, indicated by the overlapping index (OVL), with the direction of deviation in the distribution of grades, indicated by the area under the ROC curve (AUC). The OVL is plotted on the horizontal axis and the AUC is plotted on the vertical axis. As mentioned earlier, identical distributions have an OVL of 1 and an AUC of 0.5, so points near (1, 0.5) represent very similar distributions. Points below the horizontal line AUC=0.5 represent pairs of distributions where the reference distribution is shifted towards lower grades, while points above the line represent pairs of distributions where the reference distribution is shifted towards higher grades.

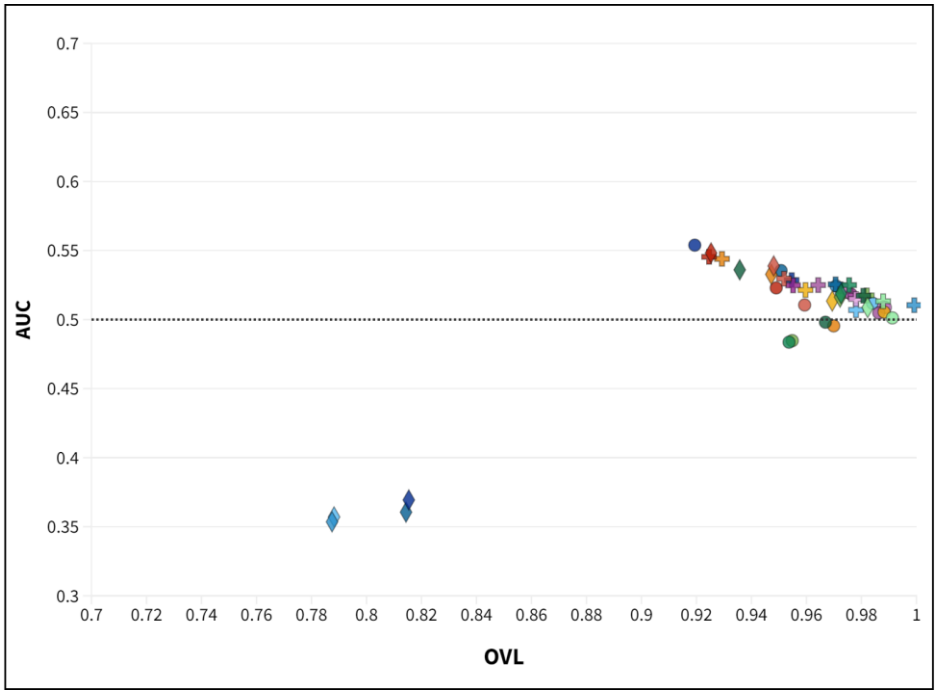

Figure 5. Arrow Plot Over Time. The Reference School Year Is 2014. Colors Are Set by Subject (Orange: Mathematics; Green: Portuguese Language; Violet: Philosophy; Blue: Physical Education) and Get Darker as Academic Years Progress. Point Shapes Vary by Curricular Year (Circle: 10th; Cross: 11th; Diamond: 12th).

Figure 5 provides additional information to what was shown in Figure 3. Each point represents a comparison between the grade distributions in 2014 and those in subsequent years (2015 to 2018) for the four subjects under analysis: Mathematics is represented in orange, Portuguese Language in green, Philosophy in violet, and Physical Education in blue. The circles represent grades given in the 10th curricular year, the crosses represent grades given in the 11th curricular year, and the diamonds represent grades given in the 12th curricular year. To show the evolution over time, the colors gradually change in intensity.

The main observation is that most points are above the line AUC=0.5 and have an OVL higher than 0.9, indicating that for each fixed subject and curricular year, the distributions in the period 2015-2018 are similar to those of 2014 but with slightly better results. The exceptions are in Portuguese Language (10th grade) and, to a much greater extent, in Physical Education in the 12th curricular year (the four diamonds in the lower left of the graph).

A similar analysis can be conducted to compare the distribution of student grades in different subjects. Figure 6 shows the Arrow Plot for grades obtained in 2018 in the 10th and 11th curricular years. For each fixed subject and fixed curricular year, comparisons were made with each of the other three subjects. All points are represented as circles and use the same color code as before: Mathematics is represented in orange, Portuguese Language in green, Philosophy in violet, and Physical Education in blue. In this case, the change in darkness reflects the curricular year. It can be seen that results in Physical Education were clearly better than those in the other three subjects, and that results in Philosophy were slightly better than those in Portuguese Language and Mathematics. The points corresponding to other comparisons between subjects are closer to the vertex (AUC=0.5 and OVL=1), indicating that the differences between them are less expressive.

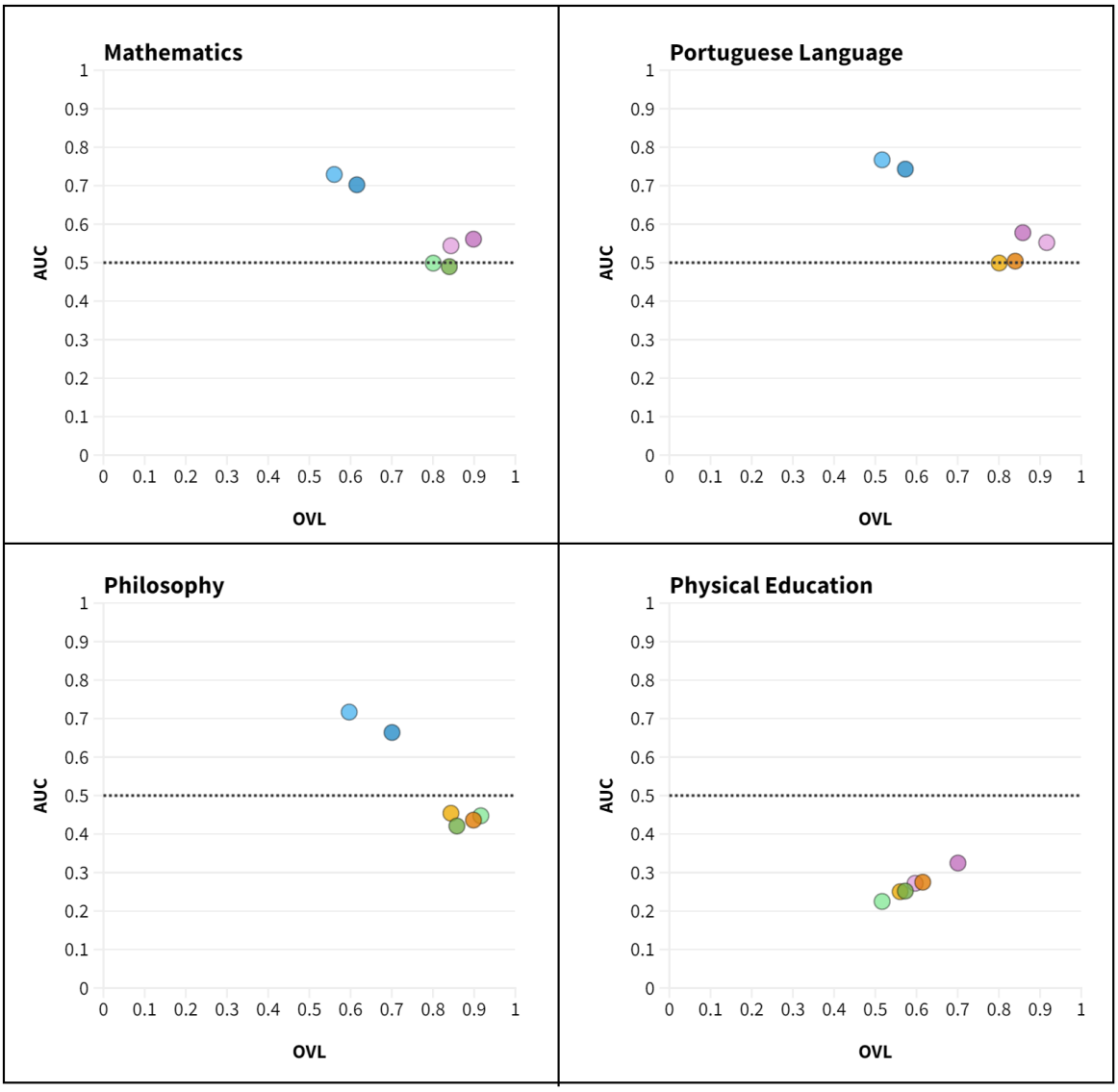

Figure 6. Arrow Plots. The School Year Is 2018 and the Reference Subject Is Specified in Each Chart Title. Colors Are Set by Subject (Orange: Mathematics; Green: Portuguese Language; Violet: Philosophy; Blue: Physical Education) and Get Darker as Curricular Years Progress (10th and 11th).

Finally, we arrive at the most striking comparisons. The subject and curricular year (10th, 11th, and 12th) are fixed, and comparisons are made between the distribution of grades between genders, with female gender as the reference. As before, in Figure 7, all points are represented as circles and the color code for subjects is maintained: Mathematics is represented in orange, Portuguese Language in green, Philosophy in violet, and Physical Education in blue. The change in darkness reflects the curricular year. The Arrow Plot shows that the blue points, representing Physical Education, are all above the horizontal line AUC=0.5, while the other points are below it, with some OVL values around 0.85. This indicates that gender disparity is significant in all academic grades (10th, 11th, and 12th) and favours boys in Physical Education. For the other three subjects, and mainly for Portuguese Language, the gender differences are also quite clear, now favouring girls.

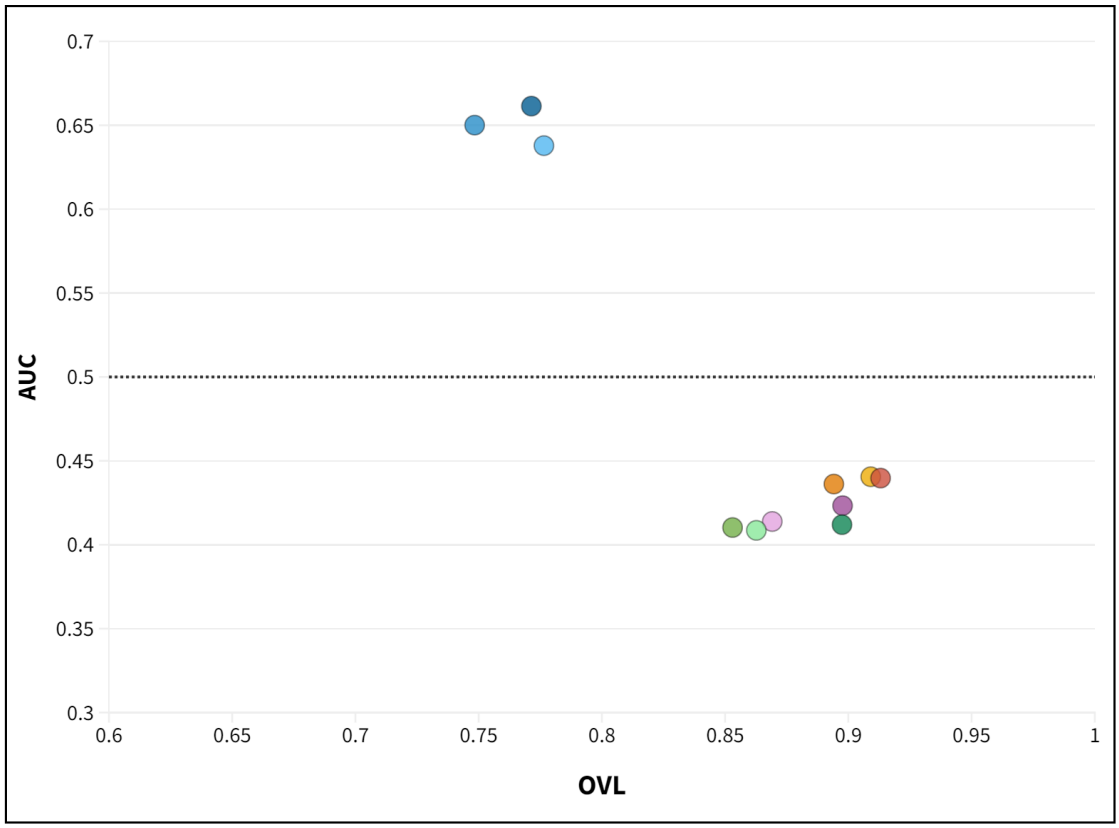

Figure 7. Arrow Plot. The School Year Is 2018 and the Reference Gender Is Female. Colors Are Set by Subject (Orange: Mathematics; Green: Portuguese Language; Violet: Philosophy; Blue: Physical Education) and Get Darker as Curricular Years Progress (10th, 11th and 12th).

We will now present statistical test confirmations of the similarities and differences revealed by the Hellinger distance, AUC, and OVL measures. As mentioned earlier, classical tests such as comparisons of means or nonparametric methods like two-sample Kolmogorov-Smirnov tests are not useful in this case because, with very large sample sizes, even small differences are declared as significant. The objective is to provide statistical evidence that the distributions are either sufficiently similar or that the differences are large enough to have practical impact. The TOST methodology, described in the "Methodology and Materials" section, has been widely used in health studies and aims to determine whether the means of two samples differ by no more than a certain equivalence margin. In our context, we have compared the proportion observed in each sample for each grade (1, 2, ..., 20) and recorded the respective p-value using the TOST procedure. Details about the choice of the equivalence margin are provided in the "Methodology and Materials" section.

As expected, the sequence of TOST tests with the Benjamini-Hochberg procedure for multiple testing rejected the null hypothesis at a significance level of 10% for all comparisons involving the same subject over time (indicating that the differences of the proportion of students that obtained the grade ”x”, for x = 1, 2, ..., 20, are within a margin of non-detectability) are within a margin of non-detectability), with the exception of Physical Education in the 12th curricular year, where only 16 to 18 equivalent proportions were obtained. Table 2 shows the number of equivalent proportions obtained when applying the TOST multiple testing procedure in the case of different subjects. The grades in consideration are those obtained in 2018 by students enrolled in the 10th curricular year.

Table 2. Number of Equivalent Proportions With TOST Multiple Testing for Different Subjects: Mathematics, Portuguese Language, Philosophy and Physical Education in Academic Year 2018 (10th Grade)

	Mathematics	Portuguese	Philosophy	Phys. Educ.
Mathematics	20	11	16	6
Portuguese		20	19	7
Philosophy			20	8
Phys. Educ.				20

Using our classification rule, we can conclude that the distribution of Physical Education grades is significantly different from the ones of the other three subjects, and that there are also statistical differences between Mathematics and Portuguese Language (TOST multiple testing detected only 11 equivalent grade frequencies) and between Mathematics and Philosophy (16 equivalent frequencies). As noted earlier with other measures, the distribution of student grades in Portuguese Language is very similar to that of Philosophy (TOST multiple testing detected 19 equivalent grade frequencies). Finally, we present the results of the TOST multiple testing procedure for gender comparisons in Table 3.

Table 3. Number of Equivalent Proportions With TOST Multiple Testing for Gender Comparison by Grade Level (10th, 11th and 12th) in Mathematics, Portuguese Language, Philosophy and Physical Education (School Year 2018)

	10th	11th	12th
Mathematics	20	20	20
Portuguese Language	14	14	15
Philosophy	17	17	-
Phys. Educ.	11	10	11

The main finding is that significant gender differences were detected in all comparisons except in Mathematics. For the other three subjects, all values in the table are less than 20. However, the values for Philosophy do not indicate a large dissimilarity between genders, as they are 17. Gender disparity in the distribution of grades is very significant in Physical Education (only half of the grade frequencies were considered equivalent by the TOST multiple testing procedure), but significant differences were also detected in Portuguese Language and, to a lesser extent, in Philosophy.

Discussion

The importance and relevance of the results presented in this article can be seen from two angles: the revelation of similarities and dissimilarities in the distribution of grades over the years and the options made on the methodologies used to quantify and provide statistical evidence for the existence of these similarities and dissimilarities.

The finding and demonstration that there is statistical regularity in the distribution of grades awarded by teachers over the years, which reflect the teacher's perception of the knowledge acquired by the students throughout the school year, not only calls us to search for explanations for this regularity (such as the possibility of a predetermined pattern defined by teachers or a relationship with the nature of the subjects and the way students assimilate the content) but also has the additional effect of providing a reference point for studies on the impact of new educational measures or curricular restructuring. This information can help to assess the effectiveness of these interventions and inform future efforts to improve student performance.

While the observation of graphical representations of the data may be sufficient to raise questions and provoke discussion and reflection on the regularity phenomenon, it is also important to demonstrate that the similarities found are not the result of chance. The approach taken involved quantifying not only the degree of similarity between the distributions, but also the direction in which any differences manifested themselves, with the goal of creating a graphical representation (Arrow plot) that would quickly reflect the results of the comparisons. The next stage involved applying statistical tests to confirm or refute the homogeneity of the distributions, which is particularly relevant given the large sample sizes and the non-applicability of classic statistical tests. Instead, a methodology using a suitable sequence of multiple equivalence tests was used. This innovative approach allowed a robust statistical analysis of the data.

The bridges connecting the results obtained in this study and those recently reported in the scientific literature in the field of education are mainly limited to the issue of gender disparities. The grades achieved by girls showed a shift towards higher values compared to boys, in three out of the four subjects considered, namely, Mathematics, Portuguese, and Philosophy. This is aligned with previous studies, such as Workman and Heyder (2020), that analyzed the academic results of boys and girls in language and STEM subjects. In contrast, boys consistently showed better results than girls in Physical Education. It is worth noting that scientific literature has paid little attention to the gender gap in Physical Education grades (Svennberg & Högberg, 2018), focusing instead on the association between good results in this subject and lower levels of stress and better results in other subjects.

As noted by Brookhart et al. (2016), grades and their interpretation are complex and influenced by various factors, such as achievement and non-achievement data, external and contextual factors, and teachers' values and beliefs. The authors conducted a literature review and found that studies mainly focused on (a) assessing the reliability and predictive validity of grades, (b) exploring the composition of end-of-secondary school grades and their relation to other academic outcomes, (c) examining teachers' perceptions of grading and assessment practices, (d) investigating the correlation between end-of-course report card grades and accountability assessments on a large scale, and (e) grading practices in higher education.

In addition, several authors, such as Resh (2009) and Prøitz (2013), addressed the issue of differences in grading across subjects. Resh's study explored the views of Humanities, Mathematics, and Science teachers on just grade allocation and the weight given to each component, such as knowledge, performance, effort, attitude and participation. Meanwhile, Prøitz's study, based on interviews with 41 teachers from six schools, balanced across five subjects (Norwegian Language, Mathematics, Science, Arts & Craft, and Physical Education) sheds light on the differences in grades distributions across subjects, as it reveals the variety of references used by teachers. The author identified two axes: one based on Universal grading (knowledge and performance) *versus* Differential grading (effort, attitude, and participation), where Arts & Craft and Mathematics fall into the first pole and Physical Education falls into the second; and another axis based on contextual orientation, which intersects the first axis and assesses whether grading is based on a continually-negotiated or on a standardized basis. Norwegian Language and Arts & Craft are positioned more in the first pole, while Mathematics falls more into the second. Sciences are at the intersection of the two axes, with a slightly greater emphasis on knowledge and performance and standardized basis references.

As far as we know, there has been no investigation into large-scale analysis of the patterns of distribution of grades obtained by students in school summative evaluations. The specificities of the Portuguese grading system, where the scale ranges from 1 (lowest) to 20 (highest) and where the grade given to a student in the last term summarizes all aspects evaluated throughout the academic year (including test results, class participation, quality of homework, and behavior), led to the identification of a phenomenon of statistical regularity. This could be a first step towards similar analysis being conducted in other countries to clarify whether the phenomenon of regularity is mainly inherent to the subjects themselves or whether it reflects the culture and national history of the grading references used by teachers.

Just like in many other research fields, the observation of distributional patterns with statistical regularity can constitute a foundational basis for the development of the area and for the refinement of theoretical models that can better explain the phenomenon, make future predictions and evaluate the impact of interventions. This was the case in the natural sciences where the Gaussianity of distributions was crucial for the advancement of hypothesis testing methodologies and analysis of variance, but also in economics, where Pareto's laws reveal a good fit to the distribution of wealth, or in psychology, where the regularity of the distribution of scores on standardized questionnaires allows for the identification of patterns in personality traits. In the field of education, some phenomena of statistical regularity have been observed, although of a different nature than those revealed by the analyzed data. They mainly manifest in the form of correlations and trends. The positive correlation between parents' socioeconomic status and their children's academic results is an example of this.

This exploratory study points to the existence of a regular pattern in the distribution of grades obtained by high school students in certain subjects. Consolidating this evidence over a longer period of time and in other subjects has clear potential for application, namely as a reference base for comparative readings (by regions, by subgroups of students from different contexts, or by school typology), for the recalibration of evaluation scales and also to identify outlier situations, such as grade inflation.

Conclusions

The grades obtained by upper secondary school students in Portugal have consistently shown statistical regularity over the years. Not only do location measures such as the mean and median remain similar, but other distributional characteristics such as dispersion, bias, and local extremes also remain unchanged. However, the large sample sizes in these studies have made it difficult to use classical statistical methodologies like tests of comparison of means or tests of homogeneity of populations, as any difference, even small ones, are detectable. Therefore, alternative approaches have been necessary to compare the distributions and obtain interpretable results.

Our objective was to confirm situations of similarity and to rank, grade and indicate the direction of differences when they existed. To achieve this, we used three measures of similarity: the Hellinger Distance, to allow for the hierarchy of differences, the overlapping index, for its easy interpretation, and the Area Under ROC Curve (AUC), to combine the quantification of the difference with the direction of displacement of the compared distribution against the reference distribution. The joint analysis of these three measures, although exploratory, allowed us to confirm some observations based on the graphic representations of the distributions. For instance, we found very small differences when comparing the same subject over the years (except for Physical Education due to an educational policy measure), clear differences between subjects (especially between Physical Education and other subjects) and relatively significant differences by gender, with girls performing better in Mathematics, Portuguese and Philosophy, and boys performing better in Physical Education. To statistically confirm these observations, we used a multiple hypothesis testing methodology where a strong

decision to reject the null hypothesis at significance level alpha indicates that the relative frequency of a certain grade in one of the distributions is close enough to the relative frequency of that same grade in the other distribution. We conclude that cases in which all relative frequencies are equivalent are completely similar.

Recommendations

The data used in this article was recorded by Portuguese public schools, year after year, and it was only possible to structure and organize them in frequency tables because there is a single platform where schools enter all the information, both at the student and teacher levels. This type of platform is present in most countries and the data it contains is a valuable resource for new and interesting discoveries. However, the subject of classifications given by teachers in school internal assessments has received less attention compared to exam classifications, which have greater potential for benchmarking and value-added studies. The regular patterns identified in this article present opportunities for exploratory analysis in various subgroups. We observed regularity within each gender, but not between genders, and it would also be interesting to conduct similar analysis by socio-economic index, region, or the nature of the course attended (scientific-technological or humanistic).

Limitations

The work developed in this article was the result of a first exploratory approach to the distributional patterns of classifications obtained by upper secondary students and, in this sense, it is limited to exposing the reality revealed by the data and presenting some statistical tools that we consider appropriate for measuring and evaluating the similarities and differences of distributions in large sample contexts. It does not allow us to identify possible reasons for the observed statistical regularity (such as subject specificities, teachers' seniority or genetic factors) or to extrapolate to universes outside of Portuguese public schools, as we did not find any other publications that treated data similar to that of classifications given by teachers at the end of the school year. Additionally, there are still open questions in terms of data exploration, such as comparing the tails of the distributions, identifying situations of greater success or failure at school and examining the central zone that appears to be the main factor in differences in academic performance by gender.

Acknowledgements

This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020.

Authorship Contribution Statement

Ferreira: Contributed with data analysis / interpretation, drafting manuscript and statistical analysis. Canto e Castro: Contributed to concept and design, data acquisition, data analysis / interpretation, drafting manuscript, supervision and final approval. Silva contributed with critical revision of the manuscript, supervision and final approval.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, *86*(4), 803–848. https://doi.org/10.3102/0034654316672069

Cieslak, D. A., Hoens, T. R., Chawla, N. V., & Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, *24*, 136-158. https://doi.org/10.1007/s10618-011-0222-1

Faraggi, D., & Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, *21*(20), 3093–3106. https://doi.org/10.1002/sim.1228

Griffin, R., & Townsley, M. (2021). Points, points, and more points: High school grade inflation and deflation when homework and employability scores are incorporated. *Journal of School Administration Research and Development*, *6*(1), 1-11. https://doi.org/10.32674/jsard.v6i1.3460

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. [New foundation of the theory of quadratic forms of infinitely many variables]. *Journal für die Reine und Angewandte Mathematik*, *136*, 210–271. https://doi.org/10.1515/crll.1909.136.210

Jensen, K., Müller, H.-H., & Schäfer, H. (2000). Regional confidence bands for ROC curves. *Statistics in Medicine*, *19*(4), 493–509. https://doi.org/btw9pf

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lewin, D. R. (2021). What can we learn from exam grade distributions? *International Journal for the Scholarship of Teaching and Learning*, *15*(2), Article 7. https://doi.org/10.20429/ijsotl.2021.150207

Ma, X. (2001). Stability of school academic performance across subject areas. *Journal of Educational Measurement*, *38*(1), 1–18. https://doi.org/10.1111/j.1745-3984.2001.tb01114.x

Meinck, S., & Brese, F. (2019). Trends in gender gaps: Using 20 years of evidence from TIMSS. *Large-scale Assessments in Education*, *7*, Article 8. https://doi.org/10.1186/s40536-019-0076-3

O'Dea, R. E., Lagisz, M., Jennions, M. D., & Nakagawa, S. (2018). Gender differences in individual variation in academic grades fail to fit expected patterns for STEM. *Nature Communications, 9*, Article 3777. https://doi.org/10.1038/s41467-018-06292-0

Pastore, M., & Calcagni, A. (2019). Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, *10*, Article 1089. https://doi.org/10.3389/fpsyg.2019.01089

Prøitz, T. S. (2013). Variations in grading practice – subjects matter. *Education Inquiry, 4*(3), Article 22629. https://doi.org/10.3402/edui.v4i3.22629

Resh, N. (2009). Justice in grades allocation: Teachers' perspective. *Social Psychology of Education, 12*, 315–325. https://doi.org/10.1007/s11218-008-9073-z

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*, 657–680. https://doi.org/10.1007/BF01068419

Silva, C., Turkman, M. A. A., & Sousa, L. (2020). Impact of OVL variation on AUC bias estimated by non-parametric methods. In O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, C. M. Torre & Y. Karaca (Eds.), *Computational Science and Its Applications–ICCSA 2020* (vol 12251, pp. 173-184). Springer. https://doi.org/10.1007/978-3-030-58808-3_14

Svennberg, L., & Högberg, H. (2018). Who gains? Sociological parameters for obtaining high grades in physical education. *Nordic Journal of Studies in Educational Policy*, *4*(1), 48-60. https://doi.org/10.1080/20020317.2018.1440112

Weitzman, M. S. (1970). *Measure of the overlap of income distribution of white and negro families in the United States* (Technical paper 22). U.S. Department of Commerce. https://searchworks.stanford.edu/view/7507794

Workman, J., & Heyder, A. (2020). Gender achievement gaps: The role of social costs to trying hard in high school. *Social Psychology of Education*, *23*, 1407–1427. https://doi.org/10.1007/s11218-020-09588-6