sustainability

MDPI

Article

Assessment Patterns during Portuguese Emergency Remote Teaching

Carlota Rodrigues [1], Joana Martinho Costa [2,*] and Sérgio Moro [2]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal; carlota_aveiro@iscte-iul.pt
[2] ISTAR, Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal; sergio_moro@iscte-iul.pt
* Correspondence: joana.martinho.costa@iscte-iul.pt

Abstract: COVID-19 certainly brought more negative aspects than positive ones to education. On the one hand, new gaps and challenges emerged from the lockdowns worldwide. On the other hand, we have been witnessing the increased relationship between technology and education, which created an opportunity for education to evolve and enhance the use of digital tools in classes. During several lockdowns worldwide, due to the pandemic crisis, millions of students and teachers were forced to continue the process of teaching and learning at home and experienced Emergency Remote Teaching (ERT), which led to new challenges on the process of students' assessment. To understand what assessment challenges teachers face during the ERT and their patterns for evaluation, we performed a survey in Portugal where the ERT lasted several months in the last two years. The survey was validated and conducted in the first semester of 2021. We found two main patterns: (i) the group of teachers that prefer oral discussion and dialogue simulation and display disbelief towards traditional tests and educational games; and (ii) the group of teachers that tend to prefer oral simulation and display greater disbelief about educational games, dialogue simulation and peer work and review. From the survey analysis, we also found that teachers considered their students to be more distracted and less engaged in online classes. They were negatively affected both in their learning and evaluation process. Using digital tools to collect and validate data and creating patterns between collected data is essential to understand what to expect in future crises. The presented analysis should be correlated with other studies to extract patterns of knowledge from data and to be able to obtain conclusions about how to move education forward.

Keywords: emergency remote teaching (ERT); COVID-19; evaluation; online assessment; assessment patterns

Citation: Rodrigues, C.; Costa, J.M.; Moro, S. Assessment Patterns during the Portuguese Emergency Remote Teaching. *Sustainability* **2022**, *14*, 3131. https://doi.org/10.3390/su14053131

Academic Editor: David González-Gómez

Received: 31 January 2022
Accepted: 1 March 2022
Published: 7 March 2022

1. Introduction

The appearance of the COVID-19 pandemic in the middle of a school year forced social isolation and caused a radical transformation in education. The changes to the teaching practice, due to the pandemic in 2020, have considerably altered how teachers recognize themselves in virtual environments, especially in terms of technological instruments and tools [1]. "Students also appear to be unable to concentrate well and are constantly distracted—both during online lessons and afterward—by all their social media" [2] (p. 9). This fact generated a need for teachers to manage and address what used to be an internal issue in a classroom externally and use new, dynamic, and efficient assessment tools. There were studies that investigated the use of online assessment tools during the pandemic and came to the conclusion that what most affected online evaluation was the students not being able to conclude tests because of their short duration, the responses on online quizzes being too limited and inflexible, or the internet connection failing during a quiz and causing the non-completion of the test. These are

some examples of things that went wrong and can be used to improve digital and distant education [3].

It was only when the COVID-19 pandemic was declared a calamity in Portugal that distant teaching was implemented in high and elementary schools for the first time. Until then, there had never been a time in which remote teaching was mandatory or even possible.

Nowadays, there are no studies concerning teachers' opinions about the modes of evaluation during the ERT emphasizing Portugal, a country where the lockdowns lasted several weeks between 2020 and 2021. Accordingly, Portugal is limited in terms of scholarly articles about the education context during ERT, whereas other countries, such as Germany, for example, discovered, through similar investigations, details that could really improve distant learning—such as the use of multiple monitors, that can relieve the teaching load, as teachers can control both the classroom and the presentation, considering more breaks or shorter lessons, or even having another instructor able to assist. These were some of the proven working techniques for teachers' success in the classroom [4]. Therefore, the main goal of this study is to find the best evaluation patterns in terms of modes of teaching and evaluation during the several ERT mandatory periods throughout the COVID-19 pandemic.

ERT is a temporary situation. However, this does not mean that it will not happen again. On the contrary, it is expected that, given to the evolution of humankind, it will happen more often. Hence the relevance of this study. "The special feature of emergency remote education is that it is an unplanned practice, with no option than to use any kind of offline and/or online resources that may be at hand. Stemming from this situation, researchers from across the globe have started to investigate a broad variety of topics related to teaching and learning during the pandemic including studies on, for example, how educators' and students' acceptance of digital formats changed in the context of COVID-19, and how this potentially affects higher education in the long-term [5], experienced instructors' views on online teaching and advice [6] or the relation between digital readiness and the social-emotional state of students [7]." [3] (p. 2). It is important to have defined evaluation methodologies that have been investigated and proven to be successful. Throughout this study, we intend to help teachers in the future to choose the most suitable assessment techniques and how to evaluate their students more successfully, using the most efficient tools and strategies, in the case of having to change from a face-to-face education to the ERT again.

This research aims to analyze teachers' perceptions about the evaluation of their students during the ERT in the 2020/2021 academic year. It was conducted to determine the extent to which the ERT impacted the evaluation of compulsory education during the COVID-19 pandemic and its benefits. To achieve the proposed objectives, we developed and validated a survey specifically designed for this research. The main objective in its development was to keep it simple and easy for the respondents. In this way, the constructed instrument allows for the collection of the needed information without burdening the participants [8].

This research aims to compare and interpret teachers' responses and perceptions, extracted from the survey performed, about online evaluation and gain helpful insights from their interpretation. It is essential to analyze teachers' opinions based on their experiences, since they are the ones who experience the assessment first-hand. There are many advantages in studies that analyze teachers and students. They allow us to learn more about how students are being evaluated and the difficulty in implementing each technique during ERT.

About this specific research, the following advantages must be highlighted:

- Learn about how students are evaluated, and the credibility of the online evaluation methods used in online evaluation during ERT;
- Learn about processes of online evaluation during ERT, whether the students understood all steps of the learning process, the delivered evaluation products and

what should be improved;
- Create patterns to help teachers to decide the best modes of evaluation in ERT;
- Facilitate future ERT states for teachers and students;
- A well-constructed evaluation process is one of the first steps for the student to grow and learn successfully.

Challenges and Impacts of ERT on Portuguese Schools

The pandemic crisis brought a significant change to teachers' and students' roles. Teachers were faced with an unpredictable, mandatory change in teaching and had to adapt every element of their teaching, including the assessment, and still coordinate their other daily tasks. "Without time to prepare, they suddenly had to teach in ways they had never taught before, with no experience, with minimal equipment, little to no support, and so on" [2] (p. 7). Students also had to adapt to this new form of learning and studying at home, which may not be the most suitable environment for learning.

According to Kirschner and Mirjam [2], students are not yet capable of managing their time remotely as supposed, which is due to not having the correct and necessary tools and knowledge to do so. Therefore, their learning is being affected. It is unquestionable that only education based solely on technology is not enough for student-centered learning and teaching. Presential teaching is also effective, but the truth is that technology makes learning and teaching more flexible [9].

To restructure the educational system during the ERT, when this first occurred, a set of critical governmental recommendations were transmitted to Portuguese schools, from elementary (1st grade to 9th grade) to high (10th grade to 12th grade) schools. The recommendations included: redefinition of curricular goals, elucidate the role of the teacher in effectively supporting student learning; guarantee support for the most vulnerable students and families, and the implementation of a communication system, adapted to each student, to closely monitor their learning [10].

The interruption of face-to-face teaching posed the challenge of adapting to this new digital era of teaching in a new educational model, based on online education methodologies that make use of digital technologies [10].

It is still premature to evaluate the number of damages, holdups, or even progresses or impacts in education after the several ERT states throughout this pandemic period. However, there is a need to redesign learning and pedagogy to obtain a better and more adaptive, transformative, and inclusive education [10,11]. Technology is arguably one of the essential parts of this new teaching and learning process, since it is what connects students and teachers and allows us to manage every teaching task online, including assessment.

Keeping in mind that the use of technology in education is not enough and cannot guarantee engagement or success in teaching and learning, the pedagogical competence of the teacher in digital education is also important and not quite so measurable [9].

2. Methodology

The primary purpose of our study was to analyze the extent to which students in the first year of elementary schools kept their grades during the ERT, how difficult it was for teachers to assess students in different educational levels, and the evolution of the learning process during the ERT, from the teachers' perspective.

According to Macdonald and Headlam [12], unless a study follows an appropriate methodology, it is implausible that the collected data and their value for science will generate a solid basis for research or even for an evaluation. Our research will be mainly quantitative, using a survey as an instrument. A survey is a suitable tool to collect attitudes and opinions from a population sample, as teachers [13]. Moreover, it presents several advantages related to the economy of the design, the simplicity of data collection, and the identification of factors of a population from a small group of participants [13].

The survey used for this research was cross-sectional—the data collected were extracted in a single period, treated, and presented to the participants [13].

2.1. Procedures

Since there was no previously validated instrument to assess the evaluation patterns during the ERT state, we first developed and validated a survey to answer the research questions. After confirming the validation and reliability of the survey, we collected responses from social media and email to include teachers from elementary to high schools.

Following the data collection, we treated and processed the results. The data analysis was divided into two main phases: descriptive statistical analysis, where we describe the results of each item of the survey, and cluster analysis, where we identify the assessment paths through the responses.

2.2. Sample

We collected 103 answers. Each of the answers was categorized by level of teaching, subject field, age, and years of experience. One answer was considered invalid, and the remaining 102 were considered valid. The distribution of the answers by demographic characteristics and teaching field can be analyzed in Table 1.

Table 1. Teacher sample categorization.

	Respondents' Characteristics	**Number Samples (102)**	**% Samples (100%)**
Age	Less than 30	6	5.88%
	30 to 39	16	15.68%
	40 to 49	36	35.29%
	50 to 59	35	34.31%
	More than 60	9	8.82%
Elementary School	Yes	60	58.82%
	No	42	41.18%
High School	Yes	42	41.18%
	No	60	58.82%
Subject Field	Elementary School—1st Years	28	27.45%
	Visual Arts, Visual and Technological Education	16	15.69%
	English	7	6.86%
	Mathematics	5	4.90%
	Informatics	5	4.90%
	Portuguese	6	5.88%
	Physical Education	5	4.90%
	Philosophy	5	4.90%
	History	2	1.96%
	Geography	1	0.98%
	Foreign Languages	4	3.92%
	Musical Education	1	0.98%
	Physical and Chemical	1	0.98%
	Economics and Accounting	1	0.98%
	Biology and Geology	1	0.98%
	Others (Law, Professional Courses, Health)	14	13.73%

2.3. Survey Development

The survey was created using Google Forms to save financial and time resources. Not all questions required a response: multiple choices and checkboxes were marked as mandatory, and all written responses were marked as optional to avoid respondents abandoning halfway through. The survey was divided into three parts. The first one aimed to categorize participants by age, gender, or type of education, considering the various ethical principles of research. The second part consisted of questions related to the transition from face-to-face to distance education. The third part consisted of statements that related to assessment during ERT.

Most of the questions were close-ended, as we were interested in obtaining a pattern. Based on a Likert scale presented as a classification table, some questions were designed to be answered as "strongly agree" or "strongly disagree". Others were asked based on multiple choices, tick boxes or open responses. The survey could be answered only one time by each participant. The responses were numbered to make it easier for the participants to read.

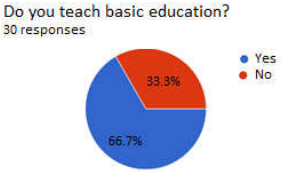
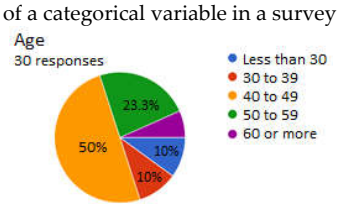
2.4. Data Collection and Analysis

We analyzed the answers and interpreted them based on the confidence interval we could estimate. According to Macdonald and Headlam [12], three factors can affect the confidence interval: the sample size, the percentage of answers, and the population's size.

Only valid responses were considered among the teachers' total responses to the survey (for example, "The students did not get involved in the classes; there was no real learning" and "Not all students had the necessary technologies for the ERT, so the learning process was not positive." were considered as valid answers). To deepen the analysis, we considered the answers by age group, gender, and type of teaching (elementary or high school).

We built a survey with well-structured questions, so that responses could be exploited in many ways. Several types of questions and variables could be considered and explored, but the most important ones for this research are the dichotomous, categorical, and latent ones (Table 2).

Table 2. Subjects for the sample.

Types of Variables	Description	Author
Dichotomous	It is used to obtain a clear view of the participants' opinion, as it consists in the choice of one from two possible answers to a single question. e.g., Yes/No. Example of a dichotomous variable in a survey question: 	(Jales, 2015)
Categorical	It is used to obtain a more descriptive answer without any measurement scale, e.g., educational. level. Example of a categorical variable in a survey question: 	
Latent	It is a hidden variable and, consequently, cannot be seen. It is normally extracted from the interpretation of other questions.	(Jales, 2015)

2.5. Data Treatment, Processing, and Validation (Instrument Analysis)

The survey gathered was exposed to a severe cleaning process that resulted in the exclusion of a few respondents, obtaining a total of 103 subjects for the sample.

The Alpha Coefficient, also known as Cronbach's Alpha, was used to validate the collected data, commonly used to measure data reliability [14].

According to Maroco and Garcia-Marques [15], the Cronbach's alpha can be classified as follows: a measure greater than 0.9 is the best possible result; between 0.9 and 0.8 is considered a good result; between 0.8 and 0.7 is considered a reasonable result; and results below 0.7 are considered weak.

The survey was administered to elementary and/or high school teachers. To validate the survey, we got 33 responses at this stage. After analyzing the responses of the pilot group, some questions were rephrased to facilitate understanding. In general, the questions were considered relevant and easy to answer. For the 33 responses analyzed in the pre-test, we obtained an Alpha of 0.773, indicating that it is reasonable to ensure reliability in the data collected, proceed with the study, and use the survey in the final data collection.

3. Results

3.1. Descriptive Statistical Analysis

We conducted a descriptive statistical analysis on the results to understand the collected data. Data can be ordered, and the differences between the various variables can be quantified [16]. These scales are designated values within one or more intervals, allowing to get information from the relationship between two or more variables of the same type [17].

Several quantitative variables collected from the final questionnaire were used to achieve the proposed goal. A Likert scale with five possible answers—Totally disagree, Disagree, Do not agree nor disagree, Agree, Totally agree—was used in a total of 37 questions to obtain more precise answers and facilitate the analysis.

Table 3 summarizes teachers' opinions about teaching during ERT, the most important items to analyze, as well as whom teachers agreed and disagreed with the most. Therefore, the mean was analyzed:

- "Students were more distracted in online classes" ($M = 3.55 \approx 4$), meaning that teachers, on average, agreed that students were more distracted in online classes;
- "Students were less engaged in the class ($M = 3.5 \approx 4$), meaning that teachers, on average, agreed that students were less engaged in online classes than in face-to-face classes;
- "Students were affected negatively as well as their learning and evaluation process ($M = 3.48 \approx 4$), meaning that teachers, on average, agreed that students' learning process was negatively impacted during ERT;
- "During the ERT, technology was hard to use" ($M = 2.38 \approx 2$), meaning that teachers, on average, disagreed that there was a lack of knowledge regarding technology.
- "Lecturing online is the same as face-to-face lectures" ($M = 2.28 \approx 2$), meaning that teachers, on average, disagreed that remote teaching is the same as face-to-face teaching during lectures;
- "Oral Discussion" ($M = 3.69 \approx 4$) was the most used assessment technique according to teachers' perceptions;
- "Educational Games" ($M = 2.29 \approx 2$) was the less used assessment technique according to teachers' perceptions.

Table 3. Mean, standard deviation, and number of responses by item.

	Question	Mean (μ)	Stand. Dev. (σ)	N
Based on your experience during the ERT, did you encounter any of the problems described?	Students did not cooperate (I1)	2.6 ≈ 3	0.989	103
	Students were more distracted (I2)	3.55 ≈ 4	0.883	
	Students were less applied (I3)	3.5 ≈ 4	0.884	
	Technological means were insufficient or uncooperative (I4)	3.02 ≈ 3	1.18	
	Technological means were hard to use (I5)	2.38 ≈ 2	0.971;	
	The training of teachers was insufficient in the technological field (I6)	2.92 ≈ 3	1.073	
	It is the same, teaching in person or remotely (I7)	2.28 ≈ 2	0.901	
How often did you use the instrument described for online assessment?	Did the ERT negatively impact the learning process and student assessment? (I8)	3.48 ≈ 4	0.989	101
	Did the evaluation tests allow the teacher to know if the subject was well taught? (I9)	2.93 ≈ 3	0.855	
	Did the assessments during the ERT allow the teacher to know if the student acquired the expected knowledge corresponding to that period under assessment? (I10)	2.99 ≈ 3	0.922	
	Were the assessment techniques that you implemented sufficient to identify what must be worked on with the students? (I11)	3.29 ≈ 3	0.898	
	Should a student who fails to achieve satisfactory results be retained even during ERT? (I12)	3.1 ≈ 3	1.035	
	Based on your experience, during the ERT, did the retention rate increase? (I13)	2.23 ≈ 2	.754	
	Quizzes (I14)	2.87 ≈ 3	1.242	99
	Presentation (e.g., Power Point, Prezi, etc.) (I15)	3.05 ≈ 3	1.137	
	Text Processor (e.g., Word) (I16)	2.78 ≈ 3	1.174	
	Oral Discussion (e.g., Teams, Zoom) (I17)	3.69 ≈ 4	1.075	
	Simulation of Dialogue between students (e.g., Teams, Zoom) (I18)	2.89 ≈ 3	1.347	
	Didactic Games (e.g., drag-and-drop activities, others) (I19)	2.29 ≈ 2	1.206	
	Work review and online peer review (e.g., Teams, Zoom) (I20)	2.7 ≈ 3	1.216	
	Traditional test (several questions with timeout) (I21)	2.68 ≈ 3	1.268	

In a few of the free answers, teachers made it clear that the most positive part of ERT was that they could use more new and different technology to connect to students as well as to evaluate them. Students were also forced to use new technology and present different types of work, which was also important. Teachers have also claimed that they gained experience and skills in the digital area and that ERT opened a new door towards remote teaching for teachers and students in case of need. Teachers also stated that the most negative part of remote teaching was that they could not manage the teaching as they wished not the students' learning process, because they simply could not control them as they used to in face-to-face classes. This means that, during ERT, it was more difficult to ascertain if students were indeed watching the classes—allegedly, some students just leave the session open and when questioned they use the lack of internet as an answer, and the evaluation is as hard for teachers to control due to not being able to check whether the students are cheating or not.

3.2. Assumptions

To conduct parametric tests, data must follow the normal distribution, and the variances must be homogeneous. We used the tests described above to verify both assumptions.

3.2.1. Normal Distribution

As the investigation followed the required criteria, there was a need to recur to normal distribution to understand if the observations were likely to fall above or below the mean in a distributed environment. Because of that, and as the first analysis, a normality test was taken.

From the 37 quantitative questions submitted to the normality test, 82% showed a *p*-value superior to 0.05, and the remaining 18% showed a *p*-value not inferior to 0.01, which means that the study follows, as expected, a normal distribution [18].

3.2.2. Levene Test

To apply a parametric hypothesis test regarding the comparison of a population mean obtained from the samples in the survey, it was necessary that the population variances, which were previously estimated, be homogeneous, or, in other words, equal [17]. Therefore, there was a need to use the Levene test for this purpose. This test is one of the most robust to calculate deviations from normality and one of the most powerful tools in testing the homogeneity of variables [17].

The hypotheses to be tested in Levene's Test are the Null Hypothesis, H0, where the variances are homogeneous, and hence equal. Therefore, they are connected. In the Alternative Hypothesis, H1, where variances are not homogeneous, they are different and have no connection with each other [17]. If $p > 0.05$, the valid hypothesis is the null hypothesis (H0). If $p < 0.05$, the valid hypothesis is the alternative hypothesis-H1 [17].

As shown in Table 4, about 91.7% of the Levene test results present a *p*-value > 0.05. Even though the remaining 8.3% presented a *p*-value < 0.05, they are also more remarkable than 0.01, which means that, although slight, there was some chance of homogeneity. As such, we conclude that most of the variables analyzed influence our answers. Hence, we could proceed to the cluster analysis.

Table 4. Levene test.

Investigation Goals	School	Levene	*p*-Value	Item
Students at the first years of elementary schools kept their grades during the remote emergency education and if there was an evaluation pattern.	Elementary	5.575	0.020	I13
		0.964	0.329	I14
		1.008	0.927	I15
		1.204	0.275	I16
		0.226	0.636	I17
		3.436	0.067	I18
		5.390	0.022	I19
		0.188	0.665	I20
		0.171	0.680	I21
	High	4.147	0.045	I13
		0.149	0.701	I14
		0.053	0.818	I15
		0.260	0.611	I16
		0.062	0.805	I17
		2.154	0.145	I18
		1.940	0.167	I19
		2.412	0.124	I20
		0.14	0.907	I21
The difficulty degree in implementing the evaluation at different levels of education was similar among teachers;	Elementary	1.086	0.300	I5
	High	0.981	0.324	
The remote emergency education caused changes in grades;	Elementary	0.109	0.742	I10
	Elementary	0.155	0.695	
There was an evolution in learning during the remote emergency education.	Elementary	0.354	0.553	I8
	High	0.894	0.347	

3.3. Cluster Analysis

Cluster analysis is usually a powerful tool when trying to do pattern recognition. When clustering, the main goal is connecting objects or variables and creating groups in a specific dataset. Cluster analysis is, basically, a way of representing similar objects in a graphic form. This similarity is calculated based on the principle of similarity, which claims that dots represent the variables under study on a graphic. The similarity is measured by the distance between those dots [19].

"The possible methods differ in how groups are defined in the algorithm used to create the groups. In general, group definition is based on within-group measures (e.g., high similarity between observations) or alternatively on between-group measures (e.g., maximum distance between objects), while clustering algorithms are based on different ways to define proximity, either similarities or dissimilarities" [19] (para. 1). When the researcher chooses these types of algorithms, there is a large set of options. Each method has a different shape and "different characteristics in terms of shape, dimension and density, and each different cluster analysis approach is more oriented towards detecting a particular type of cluster rather than others they work better when objects form round, dense clusters, rather than having elongated, overlapping distributions." [19] (para. 1).

Using this multivariate analysis technique, it will be possible to aggregate teachers into homogeneous groups, considering the frequency of use of specific assessment elements, and identify the Portuguese assessment standards used during the ERT phase. For this study, supported by SPSS Statistics (v. 25; IBM SPSS, Chicago, IL, USA), we used Ward's method. After defining the metric (the Euclidean metrics), we calculated the distance matrix and the corresponding similarity. Firstly, SPSS identified the two more similar variables. Secondly, these variables were linked in a cluster and checked the new similarity. "The lines which depart from each object are connected according to the degree of similarity at which the linkage between objects or clusters happens, so that it is possible to visualize in a fast way which level of similarity intercourses among the samples" [19] (para. 1).

The way of interpreting this cluster is by looking at the dendrogram and checking whether any pair of lines join. The lines, or variables, with the lowest distance join in the first place, as shown in the dendrogram (Figure 1).

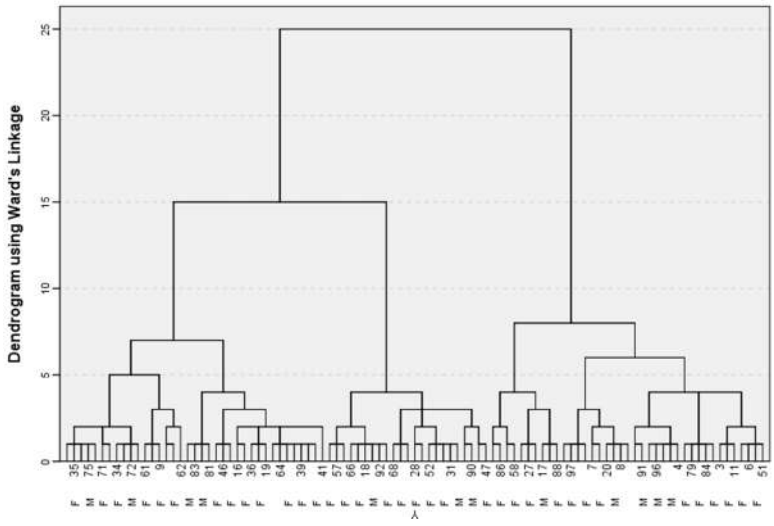


Figure 1. Dendrogram using Ward's Linkage (hierarchical).

A dendrogram is a diagram produced by a clustering algorithm during a cluster analysis that represents a tree—in this specific case, a hierarchical clustering [19]. This type of diagram usually represents the arrangement of the clusters produced corresponding to each analysis. It is also a branching strategy that reflects the relationships and differences

within a group of entities or variables. A dendrogram is a network structure constituted by a root node that splits into several other nodes connected by branches. The closer the clusters in the diagram, the more they are related. That is, they influence each other [19].

The horizontal axis (X) represents the distance between each cluster after using Ward's method, the method chosen due to its potential, while the vertical axis (Y) represents the entities regarding the used evaluation techniques used [19].

Analyzing the graph in a bottom-up approach (Table 5), it is notable that the two groups of clusters are becoming bigger and more heterogeneous along the axis, meaning that there is more variation within the cluster. That is, if we choose two responses from the same group, they will be more similar than if we choose one from each cluster. This analysis can also be checked in Table 5, which shows that the results are precise and the variables in cluster two are far more similar than in cluster one, since most of the variables' distance is superior in cluster one than in cluster two [19].

Table 5. Final cluster centers.

	Cluster 1	**Cluster 2**
Quiz	3.27	2.48
Online Presentation	3.31	2.80
Text Processor	3.04	2.52
Oral Discussion	4.24	3.14
Dialogue Simulation	3.88	1.92
Educational Games	2.82	1.78
Work and Peer Review	3.49	1.92
Traditional Test	2.65	2.70

The final cluster centers, shown in Table 5, are calculated based on the mean for each variable on each final clusters [20]. These results reproduce the characteristics of the typical case for each cluster:

- Teachers in cluster 1 tend to agree more with the use of the online evaluation tools described in Table 5. Overall, there is a high tendency and preference for Oral Discussion and Dialogue simulation and a clear disbelief in Traditional tests.
- Teachers in cluster 2 tend to disagree more with the use of the online evaluation tools described in Table 5. Overall, there is a greater preference for Oral Discussion and a greater disbelief in Educational Games, Dialogue Simulation, and Work and Peer Review.

3.4. Results

To validate this study and the previously stated hypothesis and determine whether the results were indeed statistically significant, we needed to compare the differences between the means and compare the *p*-value according to its significance level (Maroco, 2007). Two one-way ANOVAs was performed on the data, one for elementary school teachers and another for high school teachers. This means that the *p*-value needs to be analyzed according to each variable's significance level and that we need to assess the null or the alternative hypothesis: for *p*-values above 0.05, we should accept the null hypothesis; for values under 0.05, we should reject the null hypothesis. A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference [21] and the sum of squares quantifies each item variation. Dividing the sum of squares by the degrees of freedom, it is possible to compare the proportions and determine if there is a significant difference.

Regarding the ANOVA made to the data collected from elementary school teachers (as shown in Table 5), since a large percentage of the various *p*-values is greater than 0.05, we reject the alternative hypothesis, which represents the inequality of means for any level of significance. Thus, the ANOVA allowed us to conclude that, except for educational

games and the traditional test, the means of the various groups are all similar for any level of significance. This result means that, regarding the usage of each technique, there are no significant differences between answers and opinions about the evaluative techniques according to their category (elementary and high school teachers).

It was also revealed that the difficulty of implementation in each technique differed significantly between categories compared to the usage, meaning that personal data such as disciplinary area, age, and years of experience impacted these results.

The number of observations in each group is equal, so the ANOVA is robust to the violation of the assumption of the equality of variances [17]. Concerning the ANOVA test to both clusters (shown in Tables 6 and 7), we can assume that all variables, without exception, are statistically significant.

Table 6. First One-way ANOVA for Cluster 1.

		Sum of Squares	Medium Square	Z	Sig.
Quizzes	Usage	3.72	3.72	2.44	0.12
	Easy to Implement	0.21	0.21	0.32	0.57
Online Presentation	Usage	0.15	0.15	0.11	0.74
	Easy to Implement	0.17	0.17	0.22	0.64
Text Processor	Usage	0.46	0.46	0.33	0.57
	Easy to Implement	2.16	2.16	3.24	0.08
Oral Discussion	Usage	0.03	0.03	0.03	0.87
	Easy to Implement	0.06	0.06	0.07	0.79
Dialogue Simulation	Usage	0.23	0.23	0.12	0.73
	Easy to Implement	0.07	0.07	0.08	0.78
Educational Games	Usage	4.39	4.39	3.08	0.08
	Easy to Implement	3.89	3.89	4.00	0.05
Peer and Work Review	Usage	0.92	0.92	0.62	0.43
	Easy to Implement	1.45	1.45	1.76	0.19
Traditional Test	Usage	5.54	5.54	3.53	0.06
	Easy to Implement	1.28	1.28	1.50	0.23

Table 7. First One-way ANOVA for Cluster 2.

		Sum of Squares	Medium Square	Z	Sig.
Quizzes	Usage	2.32	2.32	1.51	0.22
	Easy to Implement	0.21	0.21	0.32	0.57
Online Presentation	Usage	0.03	0.03	0.02	0.88
	Easy to Implement	0.10	0.10	0.13	0.72
Text Processor	Usage	1.18	1.18	0.85	0.36
	Easy to Implement	0.52	0.52	0.76	0.39
Oral Discussion	Usage	0.19	0.19	0.16	0.69
	Easy to Implement	0.00	0.00	0.00	0.98
Dialogue Simulation	Usage	0.07	0.07	0.04	0.84
	Easy to Implement	0.04	0.04	0.04	0.83
Educational Games	Usage	12.38	12.38	9.23	0.00
	Easy to Implement	6.40	6.40	6.90	0.01
Peer and Work Review	Usage	3.15	3.15	2.16	0.15
	Easy to Implement	2.25	2.25	2.76	0.10
Traditional Test	Usage	7.62	7.62	4.93	0.03
	Easy to Implement	1.74	1.74	2.05	0.16

The means analyses are probably similar. Moreover, the results of the ANOVA test for cluster 1 (Table 6) are very similar to the results of an ANOVA for basic school, and the results of the ANOVA for cluster 2 (Table 7) are very similar to the results of an

ANOVA for high school. This leads to the belief that cluster 1 represents teachers from elementary schools and cluster 2 represents teachers from high schools.

The Spearman Correlation, also known as The Spearman rank-order correlation coefficient, was used to measure the degree of association between two variables. As shown in Table 7, most of the data collected are quite associated with the teacher's data, such as age, years of experience, and disciplinary area.

From Spearman's correlations between items about the use of online evaluation tools (Table 6), we verify that:

- Quizzes and online presentations are, clearly, more used in elementary schools than in high schools, although they are quite used at both levels;
- Text Processor, oral discussion, and dialogue simulations are more used in high schools than elementary schools;
- Work and Peer Review is much used in high schools but less used in elementary schools;
- The lesser-used tools by teachers from both types of schools are Educational Games and Traditional tests;
- All techniques seemed easy to implement, except for Educational Games, and Work and Peer Review. The variable that had less significance, and hence less impact in the implementation of techniques was the teachers' age. All other variables have a great deal of importance for the results.

3.5. Assessment Patterns

From the survey, we gathered a set of techniques to assess that vary in difficulty to implement and suitability, according to the teachers' perceptions (Table 5).

The most adequate and less difficult assessment techniques to implement were Oral Discussion, Dialogue Simulation, and Online Presentation. These techniques rejected the null hypothesis (p-value > 0.05), meaning that these variables are statistically significant. Teachers classified oral Discussion as the technique that resulted in better results and was easier to implement. Results with average grades that were medium-hard to implement showed that these were the predominant tools: Traditional Tests, Work and Peer Review, Text Processors, and Quizzes.

According to Table 5, the most challenging evaluation technique that was also inadequate and resulted in bad grades was Educational Games. This technique resulted in a p-value < 0.05, which rejects the null hypothesis and shows that this variable is not statistically significant. Hence, this technique did not have an impact on this study.

According to Tables 5 and 6, all techniques rejected the null hypothesis except for the Educational Games.

According to Table 6, there was an evaluation pattern both in elementary and high schools, showing that the three most significant evaluation techniques were, in order, Dialogue Simulation (p-value > 0.05), Work and Peer Review (p-value > 0.05), and Quizzes (p-value > 0.05). All techniques rejected the null hypothesis except for the Educational Games. So, it can be stated that there were several evaluation patterns used, the most used ones being Dialogue Simulation, Work and Peer Review, and Quizzes.

4. Discussion

The primary motivation to start this study was to contribute to the discussion about how the relationship between technology and education can be enhanced, especially in the current context. The truth is that the COVID-19 pandemic accelerated the increase of online scholar activities that were already under way (e.g., [22]). The ERT had ever happened. There has never been a demand to move from face-to-face to 100% online teaching. Even though this brought many social-economic problems, it opened an opportunity for studies such as this one to find the evaluation patterns during the ERT.

This study aimed to analyze how the online assessment in mandatory education, elementary and high education, was implemented during the ERT. Some of the specific goals looked to answer simple questions such as whether grades increased or decreased during ERT, the best online evaluation technique, and the difficulty of implementing a specific technique. A larger sample would certainly have allowed for a more consistent analysis. However, we came to a solid conclusion, mostly due to the significant results.

In addition to the lack of a larger sample, no gaps were identified that could alter or even discredit the analysis.

The data analysis and treatment were performed between Excel and SPSS. The analysis plan pursued was the following: Validation of data using Normal distribution, analysis of Mean and Standard Deviation to extract simple conclusions from the data collected, Levene Test to validate homogeneity, and follow through to the correlation and association of the data using Cluster Analysis, ANOVA, and Spearman's Correlation.

The analysis showed that 68% of the data collected were within one standard deviation of the mean, 95% were within two standard deviations of the mean, and 99.7% were within three standard deviations of the mean. This distribution allowed us to follow through with the study in the beginning.

Regarding the analysis itself, 37 quantitative questions were submitted to the normality test. With the normal distribution validated, the statistics such as mean and standard deviation of the several quantitative variables were analyzed.

According to the mean and standard deviation, teachers alleged the following:

- Students were more distracted and less engaged in online classes than before the ERT, as shown in I2 ($M = 3.55 \approx 4$) and I3 ($M = 3.5 \approx 4$) of the survey. These results are in line with the work of Kirschner and Mirjam [2], who argue that students are always distracted and little concentrated in online classes;
- Students were affected negatively (I8; $M = 3.48 \approx 4$), as well as their learning and evaluation process. These results can be related to the described problems in the studies of Bond et al. [3] and Kirschner and Mirjam [2], namely, the lack of preparation of students to manage their time remotely, students not always being able to conclude the assessments due to internet connection problems, and the lack of flexibility of online quizzes, among other issues that affect the online evaluation;
- According to I5, technology was not hard to use for Portuguese teachers ($M = 2.38 \approx 2$), which concurs with the study of Maroco [23], where the author analyzed the use of technologies of over 4000 teachers during the ERT;
- Lecturing online is not the same as face-to-face lecturing (I7; $M = 2.28 \approx 2$), which was also observed in Vieira and Silva [10];
- From items 14 to 21 it is possible to verify that Oral Discussion (e.g., Teams, Zoom, among others) is the most used evaluation technique in elementary and high education Oral Discussion (e.g., Teams, Zoom (I17; $M = 3.69 \approx 4$) and Educational Games (I19) is the less used evaluation technique ($M = 2.29 \approx 2$.

Based on Table 5, we verified the presence of two clusters with the followed characteristics:

Teachers in cluster 1 tend to agree more with using online evaluation tools described in Table 6. Overall, there is a big tendency and preference with Oral Discussion and Dialogue Simulation and a clear disbelief in Traditional Tests and Educational Games.

Teachers in cluster 2 tend to disagree more with using the online evaluation tools described in Table 6. Overall, there is a greater preference for Oral Discussion and a greater disbelief in Educational Games, Dialogue Simulation, and Work and Peer Review.

In general, all techniques seemed to be used and were medium–easy to implement. The less used were Educational Games, which was the variable that had less significance.

The most adequate and less difficult evaluation techniques to implement that resulted in good grades and were easy to implement were Oral Discussion, Dialogue Simulation, and Online Presentation. High school teachers classified oral Discussion as the

technique that resulted in better results and was easier to implement. The ones that resulted in average grades and were medium–hard to implement were Traditional Tests, Work and Peer Review, Text Processors, and Quizzes, both for elementary and high schools, and the most difficult evaluation technique which was not adequate and resulted in bad grades was Educational Games. As said, this technique did not have an impact in this study.

This is quite normal, since traditional tests are more commonly used in face-to-face teaching than in remote teaching. This is so because it is more difficult to implement this type of test remotely. It can be stated that there was an evaluation pattern, in general, both in elementary and high schools, with the most three significant evaluation techniques being the Dialogue Simulation, Work and Peer Review, and Quizzes.

This study presents the perspective of the Portuguese teachers and their patterns of evaluation during the ERT phases. We found that teachers' opinions tended to converge into two main groups: those who gave priority to oral discussion and dialogue simulation; and those who prefer oral simulations and express disbelief about educational games, dialogue simulation, peers' work and review. From the results, we concluded that teachers diversified the assessment during the ERT and used the traditional test less than before the ERT.

4.1. Limitations

No study is entirely free of error. According to Price [24], there are two most common and important groups of limitations, represented by threats to internal and external validity. Both can affect the outcome of the research. The fact that this is a study on a relatively new topic. Distance learning during ERT carries risks and means that not much research has been done on the topic so far, and there is not much information available to recapitulate. So, the information was gathered from different platforms with articles from several different authors, and it was analyzed and considered to gather more concise data. Besides, since our research was conducted during a pandemic, the number of responses was affected, which did not allow us to generalize the results to other contexts or even to all Portuguese teachers, since the content taught and school level are diverse within the sample.

4.2. Future Lines of Research

Our research leads to new possibilities of research and new questions. After two years of pandemic, it is necessary to analyze its effects on students' knowledge and the future standards of teacher evaluation, after the end of the pandemic. It is important to understand if the patterns detected in this study will be used in the coming years or if, after this phase, teachers will adopt a new assessment model that uses hybrid characteristics or completely return to a face-to-face assessment.

Finally, there might be a possibility to follow through and escalate this study using social media and a deeper analysis of the qualitative data on the future—for example, dividing the results by private and public schools, or by geographic location. It may be equally interesting to use longitudinal surveys to understand how assessment develops over time in schools, whether in ERT or outside of it.

Author Contributions: Conceptualization, C.R. and J.M.C.; methodology, C.R., J.M.C. and S.M.; validation, J.M.C. and S.M., investigation, C.R.; writing—original draft preparation, C.R. and J.M.C.; writing—review and editing, C.R., J.M.C. and S.M..; supervision, J.M.C. and S.M.; funding acquisition, J.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) [ISTAR Projects: UIDB/04466/2020 and UIDP/04466/2020].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available under request to the author.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Vieira, G.S. Ensino remoto de emergência: Reconhecimento e usos dos ambientes virtuais de aprendizagem pelos professores do curso de Direito/Ceres. *Anais do 39º Seminário de Atualização de Práticas Docentes* **2020**, *2*, 11–17.
2. Kirschner, P.; Mirjam, N. Emergency Online Teaching ≠ Online Learning—3-Star Learning Experiences. 2020. Available online: https://3starlearningexperiences.wordpress.com/2020/11/05/emergency-online-teaching-≠-online-learning/ (accessed on 4 January 2021).
3. Bond, M.; Bedenlier, S.; Marín, V.I.; Händel, M. Emergency remote teaching in higher education: Mapping the first global online semester. *Int. J. Educ. Technol. High. Educ.* **2021**, *18*, 48, https://doi.org/10.1186/S41239-021-00282-X/FIGURES/6.
4. Tan, D.Y.; Chen, J.-M. Bringing physical physics classroom online—Challenges of online teaching in the new normal. *Phys. Teach.* **2021**, *59*, 410–413, https://doi.org/10.1119/5.0028641.
5. Vallaster, C.; Sageder, M. Verändert Covid-19 die Akzeptanz virtueller Lehrformate in der Hochschulausbildung? Implikationen für die Hochschulentwicklung. *Zeitschrift Für Hochschulentwicklung*, **2020**, *15*, 281–301.
6. Rapanta, C.; Botturi, L.; Goodyear, P.; Guàrdia, L.; Koole, M. Online university teaching during and after the Covid-19 crisis: Refocusing teacher presence and learning activity. *Postdigital science and education* **2020**, *2*, 923–945.
7. Händel, M.; Stephan, M.; Gläser-Zikuda, M.; Kopp, B.; Bedenlier, S.; Ziegler, A. Digital readiness and its effects on higher education students' socio-emotional perceptions in the context of the COVID-19 pandemic. *J. Res.Technol. Educ.* **2020**, 1–13, https://doi.org/10.1080/15391523.2020.1846147.
8. Mertler, C.A. Patterns of response and nonresponse from teachers to traditional and web surveys. *Pract. Assess. Res. Eval.* **2002**, *8*, 22, https://doi.org/10.7275/2kdf-g675.
9. Bond, M.; Marín, V.I.; Dolch, C.; Bedenlier, S.; Zawacki-Richter, O. Digital transformation in German higher education: Student and teacher perceptions and usage of digital media. *Int. J. Educ. Technol. High. Educ.* **2018**, *15*, 48, https://doi.org/10.1186/s41239-018-0130-1.
10. Vieira, M.; Silva, C. A Educação no contexto da pandemia de COVID-19: Uma revisão sistemática de literatura. **2020**, *28*, 1013–1031, https://doi.org/10.5753/rbie.2020.28.0.1013.
11. Rosa, J.; Zaboroski, A. Ensino remoto e pandemia COVID-19: Desafios e oportunidades de alunos e professores. *Interacções* **2020**, *57*, 41–57.
12. Macdonald, S.; Headlam, N. Introductory Guide to Research Methods for Social Research. 2008. Available online: www.cles.org.uk (accessed on 15 January 2022).
13. Creswell, J. *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*, 2nd ed.; Sage Publications: Newbury Park, CA, USA, 2003.
14. Matthiensen, A. Uso do Coeficiente Alfa de Cronbach em Avaliações por Questionários. 2011. Available online: www.cpafrr.embrapa.br (accessed on 16 January 2022).
15. Maroco, J.; Garcia-Marques, T. Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? *Lab. Psicol.* **2006**, *4*, 65–90.
16. Costa, J.M.; Miranda, G.L. Desenvolvimento e validação de uma prova de avaliação das competências iniciais de programação. *RISTI-Rev. Ibérica Sist. Tecnol. Inf.* **2017**, *25*, 66–81, https://doi.org/10.17013/risti.25.66-81.
17. Maroco, J. Testes paramétricos para comparar populações a partir de amostras independentes. In *Analise Estatistica: Com Utilização do SPSS*; Sílabo: Lisboa, Portugal, 2007.
18. Boodie, K. Normal Distribution of Data: Examples, Definition & Characteristics. 2019. Available online: https://study.com/academy/lesson/normal-distribution-of-data-examples-definition-characteristics.html (accessed on 12 January 2022).
19. Tullis, T.; Albert, B. Hierarchical Cluster Analysis—An overview. In *Measyring the User Experience*, 2nd ed.; Elsevier: Amsterdam, The Netherlands, 2013. Available online: https://www.sciencedirect.com/topics/computer-science/hierarchical-cluster-analysis (accessed on 12 January 2022).
20. Wu, J. Cluster analysis and K-means clustering: An introduction. In *Advances in K-Means Clustering*; Springer, Berlin, Heidelberg, 2012; pp.1–16.
21. Duffy, T.; Gilbert, I.; Kennedy, D.; Kwong, P.W. Comparing distance education and conventional education: Observations from a comparative study of post-registration nurses. *ALT-J* **2002**, *10*, 70–82, https://doi.org/10.1080/0968776020100110.
22. Costa, J.M. Microworlds with different pedagogical approaches in introductory programming learning: Effects in programming knowledge and logical reasoning. *Informatica* **2019**, *43*, 145–147, https://doi.org/10.31449/inf.v43i1.2657.
23. Maroco, J. O que nos dizem os dados? Experiências de Ensino à Distância em Tempos de Pandemia (What Does the Data Tell Us? Distance Learning Experiences in Pandemic Times). 2020. Available online: https://somossolucao.pt/2020/08/31/o-que-nos-dizem-os-dados/ (accessed on 21 February 2022).
24. Price, J.H. Research Limitations and the Necessity of Reporting Them. 2004. Available online: https://search.proquest.com/openview/b6991f124333fca111dfbc6ef96d080c/1?pq-origsite=gscholar&cbl=44607 (accessed on 2 January 2022).