







Article

LearningRlab: Educational R Package for Statistics in Computer Science Engineering

Juan J. Cuadrado-Gallego ^{1,2} , Josefa Gómez ¹ , Abdelhamid Tayebi ^{1,*} , Luis Usero ¹ , Carlos J. Hellín ¹ 
and Adrián Valledor ¹ 

¹ Computer Sciences Department, University of Alcalá, 28801 Alcalá de Henares, Spain

² Department of Computer Science and Software Engineering, Concordia University, Montreal, QC H3G 2J1, Canada

* Correspondence: hamid.tayebi@uah.es

Abstract: This paper describes and evaluates the educational interest of LearningRlab, an educational R package developed for teaching statistics in computer science engineering. The package was developed by final degree project students to be used as an educational environment for statistics students who evaluated the package and provided feedback for future versions. Such a process increases the motivation of both groups of students. This paper presents how the use of the R packages conceived and developed for engineering education can improve the learning process in the computer science engineering bachelor's degree. Two different evaluations, one performed by a group of statistics students, and the other performed by final degree project students, were used to evaluate the impact on the learning process of the first version of the package to develop the second version of the package, which corrects and enhances the first version. The evaluation results show a positive effect on the learning process in both subjects. The analysis of the learning outcomes reflected in the grades of the experimental and control groups demonstrates that LearningRlab can be used as a teaching aid for statistics and final degree project subjects of the computer science engineering degree. The average laboratory grade of the students who used the package (5.76) was noticeably higher than the average laboratory grade of students who did not use it (1.84).

Keywords: computer science engineering; statistics; descriptive statistics; probability; final degree project; R packages



Citation: Cuadrado-Gallego, J.J.; Gómez, J.; Tayebi, A.; Usero, L.; Hellín, C.J.; Valledor, A. LearningRlab: Educational R Package for Statistics in Computer Science Engineering. *Sustainability* **2023**, *15*, 8246. <https://doi.org/10.3390/su15108246>

Academic Editors: Julio Ruiz-Palmero, Moussa Boumadan, Roberto Soto-Varela and Melchor Gómez-García

Received: 13 March 2023

Revised: 16 April 2023

Accepted: 20 April 2023

Published: 18 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

This paper presents the educational R package LearningRlab, which aims at improving teaching and learning processes in computer science engineering (CSE). This paper describes how the learning process of two CSE bachelor degrees, statistics for computer science and final degree project students, can be improved by developing and using an R educational package. The package, called LearningRlab [1], was developed by three bachelor students attending a large public university in Spain in two different academic courses. Two of them developed the first version, and the other developed the second version. The package aim to improve the teaching and learning process of another subject within the CSE degree, statistics, whose students not only use it to improve their learning but also to participate in the correction and improvement of the newer versions of the package. Establishing this strong connection between both contributors, the first being at the beginning of their studies and the other at the end, benefits from the students being from different generations, separated by four years.

Consequently, this paper describes the two main educational applications for which the package was conceived. Using this software package, coded in R, statistics students can obtain, in their laboratory practical teaching sessions, a deeper knowledge of the concepts acquired in their theoretical teaching sessions while learning the practical application. In

this way, the package allows students to strengthen their understanding of concepts related to descriptive statistics and probability fundamentals.

R is an environment conceived and developed for statistics; consequently, R implements all the statistical techniques from the R core, the standard library, and packages in repositories, such as in the CRAN (comprehensive R archive network).

R was first intended to be used in statistical studies; consequently, most packages were developed to solve statistical problems. However, R evolved, and many efforts were made to develop R packages that could help in multiple disciplines. Some examples include the works of [2–4], as well as the R packages for environmental scientists, engineers, and regulators described in [5]. Among the several works found in the literature that deal with the use of R to solve statistical problems [6], there are some that explore statistics from different perspectives. As an example, some R packages make statistical analysis easier and more intuitive and show students how to solve various statistical problems [7,8].

On the other hand, although both descriptive statistics and probability fundamentals are profusely implemented in R, there are few educational packages about these topics. That is, although packages can be used to calculate the results of different techniques [9–13], they fail to include theoretical descriptions of the functions and how they operate to explain the practical application of theoretical concepts learned in the classroom. Even if they have a brief description, they were not primarily designed and developed for educational purposes. For instance, according to R core documentation [14], it just computes the arithmetic mean, with the function `mean()` of the dataset indicated as an input parameter. Recent work has proposed the use of e-learning platforms [15,16] and web-based apps [17] to teach statistics.

For the second application in education, by contributing to the development of the R package, the final degree project students can obtain a level of proficiency in their knowledge of R, beneficial to their successful entry into their professional career. Simultaneously, they are highly motivated to develop the project because the results will be used by future students.

The research questions of this work are as follows:

- Can an educational R package designed to teach statistics in computer science engineering be developed by final degree project students?
- Can this educational R package improve the learning of students?

The rest of the paper is organized as follows: Section 2 presents the characteristics of the subjects involved in the development of the package, as well as the statistical theoretical knowledge implemented. Section 2 also describes the main features of LearningRlab and how the package can be used in statistics to reinforce concepts of descriptive statistics and probability. The results of the validation process are discussed in Section 3. Future work and concluding remarks are presented in Section 4. Appendix A includes the main characteristics and intended use of all the functions implemented into the first version of LearningRlab. Finally, the new functions implemented in the second version of LearningRlab are included in the Appendix B.

2. Materials and Methods

2.1. Fundamentals of Statistics and the Final Degree Project

The two subjects of the CSE degree, for which LearningRlab was planned and developed to improve the teaching and learning, are statistics for computer science and the final degree project. Statistics is a subject taught in the first semester and the final degree project is the final assessment of the degree. Due to this, it can be said that this work connects the first and last students of the degree in an educational improvement process. In this section, both subjects are discussed in detail.

The CSE degree program at the research university is eight semesters long and spans four academic years. The current study plans are structured into three formative blocks: basic, mandatory, and optional.

- Basic subjects are general subjects within branching knowledge, not necessarily specific to the degree.
- Mandatory subjects constitute specific content of the degree, corresponding to the specific competences of the degree.
- Optional subjects are chosen by students from a list of optional subjects of the corresponding degree.

The first year of the CSE degree includes eight subjects, including statistics. The fourth year, consisting of the seventh and eighth semesters, includes seven subjects, with the final degree project being the last one.

2.1.1. Statistics for Computer Science

Statistics is a mandatory subject in the first semester of the CSE degree worth six ECTS credits. Its importance as a basic instrumental science for data analysis is not only applicable to undergraduate CSE studies, but also in practically all the undergraduate study plans, with the exception of the humanities, both in Spain and abroad.

The statistics subject covers the fundamental concepts and methods in statistics, including both descriptive and inferential statistics. Additionally, students learn the necessary probability concepts to understand and apply statistical methods. The subject is taught both theoretically and practically, utilizing a computer environment for statistical data analysis.

The course comprises the following seven lessons:

- Lesson 1. Introduction to statistics.
- Lesson 2. Descriptive statistics: description of one variable.
- Lesson 3. Descriptive statistics: joint description of several variables.
- Lesson 4. Probability.
- Lesson 5. Random variables and probability distribution models.
- Lesson 6. Statistical inference: estimation.
- Lesson 7. Statistical inference: hypothesis contrast.

The instructional methodology for teaching all lessons includes two kinds of sessions held sequentially over two weeks:

- Theoretical lessons are conducted in the classroom without software, where the fundamental concepts are presented using presentation tools. In these sessions, practical cases and exercises are solved, complementing the theoretical content.
- Using software for the practical part of the statistics subject is indeed important for the learning process. By solving the same practical cases and exercises using software, students can see first-hand how the theoretical concepts are applied in a real-world context. Additionally, using software can help students better understand complex statistical calculations and concepts, as the software can perform these calculations quickly and accurately. Overall, the combination of theoretical and practical sessions can provide a comprehensive learning experience for students in the statistics subject.

2.1.2. Final Degree Project

The final degree project is a mandatory subject in the eighth semester of the CSE degree, worth 15 ECTS credits. The main objective of this subject is for students to apply and develop all their competencies required for their future professional work while completing an original, autonomous, and individual project. In this way, the project serves as proof of the student's knowledge and competence just before graduation and moving into the professional world. To partake in the subject, the students must have passed all basic subjects, with good marks in subjects related to the project.

It is crucial to emphasize that the final degree project report must be entirely original, meaning it cannot reproduce any previously presented works by the student or anyone else. Once the originality is validated, the work must be defended in a public dissertation in front of a panel of three supervisors.

Another important aspect of the final degree project is the role of the supervisor involved. Every final degree project must have a supervisor in charge of the student and the project. The supervisor must approve the project, the planning of the development tasks, and the final presentation to the commission. Additionally, the supervisor must periodically review the student's work and provide corrections where necessary.

The student is required to dedicate a minimum of 300 h to the final degree project, and their results are evaluated based on the following criteria:

- Project and development of the work: work developed by the student at different stages of the final degree project described in the preliminary project.
- Scientific report: scientific technological quality of the developed report, as well as writing the scientific report that respects the format and content detailed in the higher polytechnic school regulations.
- Oral presentation: presentation and defence of the work carried out and the results obtained.

2.2. Descriptive Statistics and Probability

This section presents the theoretical statistical knowledge incorporated into the LearningRlab educational package. While some concepts are well-known, it is important to clarify them to understand the intended theory being taught.

The adaptation of the statistical knowledge to computer science considers the most essential concepts that the CSE students need to acquire. The statistics subject is composed of three main parts: descriptive statistics, probability fundamentals, and statistical inference.

The LearningRlab package is designed to cover the entire statistics course. However, as detailed in the package section, it has been developed iteratively and incrementally. For this reason, each part of the course is intended to be implemented in a different version of the package. Consequently, the two versions of the package described in this paper are version 1, implementing descriptive statistics, and version 2, implementing probability. Future versions are planned to include statistical inference and completion of all the functions covered in the theory section.

A list of the main theoretical concepts of descriptive statistics and probability implemented into the two versions of the package are described in the following:

- Descriptive Statistics. The concepts included in the statistic subject program for descriptive statistics were carefully selected to ensure that all students acquire the necessary knowledge on the fundamentals of statistics, enabling them to apply these concepts in other subjects and future professional careers; particularly, software management and fundamentals of data science, which rely heavily on descriptive statistics. Lessons 1 and 2 of the course introduce the key concepts of descriptive statistics:
 - Concepts related to the number of times a studied variable is observed in the dataset. Concepts related to frequency: absolute frequency, relative frequency, absolute accumulated frequency, and relative accumulated frequency.
 - Concepts related to central tendency metrics for non-ordered data that represent the whole dataset: arithmetic and harmonic mean, and mode. Concepts related to central tendency metrics for ordered data that represent the whole dataset: median, percentile, and quartile.
 - Concepts related to dispersion parameters that allow verification of whether the central parameters fulfil their objective to represent the whole dataset. These concepts include: variance, standard deviation, average deviation, and variation coefficient.
 - Concepts related to parameters that allow students to determine the degree of association between two variables. These concepts are: covariance and Pearson's correlation coefficient.
- Probability. The concepts of probability that are included in the statistic subject program were selected taking into account the knowledge of probability needed to understand the main concepts of statistical inference. In addition, the relations of these

concepts with other subjects taught in the CSE degree were considered, between all the fundamentals of data science, whereby inferential statistical knowledge is profusely used. The concepts introduced in the probability part of the course, lessons 4 and 5, are:

- Concepts related to the fundamentals of probability, including conditional and Bayesian probability. These concepts are: probability definition, probability laws, conditional probability and Bayesian probability.
- Concepts related to stochastic variables and probability distributions. These concepts are: stochastic variable, mode, Laplace probability distribution, probability distribution binomial, Poisson's probability distribution, probability distribution normal, Student's t probability distribution, chi squared probability distribution and Fischer's probability distribution.

2.3. R Educational Package LearningRlab

LearningRlab [1] is an R package specifically designed and developed for statistics teaching in CSE studies, making it an R educational package. It is a live R package that currently has two versions, both developed by CSE final degree project students. It has been designed to not only provide solutions to basic functions required for the statistics subject, but also serve as a tool for teaching and learning the theoretical basis of these functions and how they operate. Consequently, the functions are implemented in a theoretical, step-by-step manner, as recommended by [18].

In this section, the main features of the two features of LearningRlab are presented, starting with an introduction to the R environment. Furthermore, an example of how each type of implemented function (main, explained and interactive) operates is included to better illustrate the purpose of the R educational package.

2.3.1. R Project

The first page of the R project [19] describes R as “a language and environment for statistical computing and graphics”. The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accumulation of very specific and inflexible tools, as is often the case with other data analysis software. New functionalities are implemented through new packages published by the community. These packages are found in different repositories, with the main one being CRAN. This is the primary characteristic that makes R so useful and interesting. All published packages are documented. R packages are installed in libraries, which are directories on the file system that contain a subdirectory for each installed package. R comes with a single library, R-4.0.3/library, which is the value of the R object ‘.Library’ containing the default and standard packages. However, users can create other libraries and make use of the packages installed in them (or not) in an R session or in all R sessions.

2.3.2. First Version of LearningRlab

On the one hand, the statistics subject for computer science has three main parts: descriptive statistics, probability fundamentals and statistical inference. On the other hand, from the beginning, the development of the LearningRlab educational package was designed as an iterative incremental project, in which more functionalities could be included in each iteration. The result of combining both perspectives was the decision to only implement the first part of the statistics subject, descriptive statistics, in the first version of the package.

Consequently, the first version of LearningRlab included thirty-six functions dealing with the main parameters used in descriptive statistics. These thirty-six functions come from three different educational approaches to the twelve main statistics parameters. Each approach was defined as a set of twelve functions organized under three definitions: main function, explained function and interactive function. The main characteristics, intended use and examples of these are presented in the Appendix A.

Next, an example of each type of function implemented into LearningRlab (main, explained, and interactive functions) is included. Specifically, the selected function for the example below is the binomial probability distribution. The function has three input parameters: 'n', 'p' and 'x', where 'n' is the number of trials, 'p' is the probability of success, and 'x' is the value of the variable.

- Main function

In this case, the user must indicate the value of each parameter and execute the function.

```
>n = 3
>p = 0.7
>x = 2
>binomial(n, x, p)
[1] 0.441
```

- Explained function

The user must also indicate the value of each parameter and execute the function. However, in this case, the function not only provides the result, but also a detailed explanation of the steps followed to obtain the result.

```
>n = 3
>p = 0.7
>x = 2
>explain.binomial(n, x, p)
```

```
[1] BINOMIAL DISTRIBUTION
```

```
Binomial distribution with parameters n
and p is the discrete probability
distribution of the number of successes
in a sequence of n independent experiments,
each asking a yes or no question,
and each with its own
Boolean - valued outcome: success
(with probability p) or failure
(with probability q = 1 - p)
```

```
Formula -> ((factorial (n) / (factorial (x)
* factorial (n-x))) * (p ^ x) *
(1 - p)^(n - x))\n")
```

```
Use Example
```

```
First of all, we need to know the n,
the number of trials
```

```
In this case n=3
```

```
Second, we need to know the p, probability
of success
```

```
In this case p=0.7
```

```
Finally, we need to know the x, random
variable
```

```
In this case x=2
```

```
Formula applied -> (factorial (3)
/ (factorial (2) * factorial (3-2))) *
(0.7 ^ 2) * (1-0.7)^(3-2) = 0.441
```

Now try by your own!

Use `interactive.binomial` function
to practice.

- Interactive function

Interactive functions do not receive parameters. The user must execute the function and follow the instructions shown. The result will vary according to the values indicated by the user in each parameter.

```
>interactive.binomial()
```

```
[1]  $\frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$ 
```

Insert n, the number of trials.

```
>3
```

Insert p, the probability of success.

```
>0.7
```

Insert x, the binomial random variable.

```
>2
```

OK! Next move!

Please, insert the result of the calculus for your data (if the result has decimal part, round to the 3rd)

```
>0.823
```

Ups, that might not be correct...

```
>0.761
```

Hint 1 → Psst!... Look at the formula on the plot panel at your side

```
>0.445
```

Hint 2 → Check that you are entering your result correctly. It's easy to be wrong.

```
>0.441
```

Well done!

2.3.3. Second Version of LearningRlab

Following two different evaluations of the first version, described later, the second version of the package was developed as part of a CSE final degree project [1]. This new version addresses errors identified during the evaluations and includes improvements to clarify the operation of some functions. In addition, eleven new functions have been developed to help users understand the second part of the course, which focuses on the concepts of probability covered in the theoretical lessons. Following the same methodology as in the first version of the package, three subfunctions (main, explained and interactive) have been added for each new function. The new functions can be found in the Appendix B.

Furthermore, LearningRlab was not only developed to teach descriptive statistics and probability fundamentals but also to enhance the knowledge and skills of students engaged in the final degree project, focusing on R and R packages. One such skill is to obtain approval for LearningRlab to be published in the official R repository, CRAN. To achieve approval, not only must the package be developed, but the CRAN package webpage must also include various information, products, and links. First, there are nine options that provide information about the package:

1. Version: It gives the most recent version of the package. Here, it is version 2.3.
2. Depends: Dependencies. Packages that must be uploaded for the package to work. It is uploaded on [magick](#) and [crayon](#).
3. Suggest: Suggested packages to improve the package. These include: [knitr](#) and [rmarkdown](#).

4. Published: Date of the last publication, 17 June 2022
5. Author: Authors names.
6. Maintainer: Maintainer of the package.
7. License: Package license type, Unlimited.
8. NeedsCompilation: Indicates if the package needs compilation. In this case, Yes.
9. CRAN Checks: Provides the proofs of the package (see, LearningRlab results).

Following are the links to the package documentation:

1. Reference manual: Provides the operating reference manual of the package, [LearningRlab.pdf](#) [20].
2. Vignettes: Small documents that illustrate and explain the operation of the package, [Introduction to LearningRlab](#) [21].

Following are the links to the package downloads:

1. Package Source: Ffiles that have the source code of the package, [LearningRlab_2.3.tar.gz](#)
2. Windows binaries: Installation files of the package in Windows. These include: [r-devel: LearningRlab_2.3.zip](#), [r-release: LearningRlab_2.3.zip](#), [r-oldrel: LearningRlab_2.3.zip](#)
3. macOS binaries: Installation files of the package on Mac. These include: [r-release \(arm64\): LearningRlab_2.3.tgz](#), [r-release \(x86_64\): LearningRlab_2.3.tgz](#), [r-oldrel: LearningRlab_2.3.tgz](#)
4. Old sources: Older versions of the package, obtained at: [LearningRlab archive](#).

Another item on the page is available to link to the package page.

1. Please use the canonical form: Furthermore, the following links to the homepage from others: <https://CRAN.R-project.org/package=LearningRlab>

2.4. Teaching and Learning Process of Statistics Using LearningRlab

This section showcases how the LearningRlab package is utilized in the statistics subject to enhance the students' understanding of descriptive statistics and probability fundamentals. The teaching and learning process for all lessons comprises two distinct, yet entirely interconnected, approaches. These approaches are covered over consecutive weeks, with four hours dedicated to each: two hours in the classroom followed by two hours in the laboratory. The sessions are staggered such that in the first week the classroom covers the theoretical foundations, while the laboratory addresses the previous lesson. In the first week, there is no laboratory lesson. The following week, the theoretical foundations are continued and the laboratory begins exploring the application of concepts using software, which continues into the next week. The laboratory lesson serves two main purposes: demonstrating how statistical techniques are applied using software, as they would in industry, and enhancing the understanding and knowledge of theoretical concepts in greater depth.

Various software packages were chosen in the laboratory to cover the same subject matter in the lessons. Almost all of these packages are regular packages, meaning they are not specifically designed or developed to teach the parameters they implement, but rather to solve statistical problems. In most cases, although it would be preferable if they were developed for educational purposes, they are sufficiently useful for teaching. However, for some content that is more challenging for students to understand, a dedicated package designed from the ground up for educational purposes would be beneficial. This is the case with LearningRlab.

After several years of teaching the statistics subject, the lessons or parts of lessons that pose the greatest difficulty for students to understand and learn have been largely identified based on their assessments and questions asked during lessons or tutorials. Our findings align with those of other researchers from previous studies [22]. Some of the descriptive statistics and probability topics were identified as more challenging to understand, and no particularly effective packages for teaching these topics were found. Consequently, it was decided to develop a new package designed for educational purposes, following the same sequence as the lessons being taught. This led to the creation of LearningRlab.

The application of the new package, LearningRlab, to the teaching of descriptive statistics and probability was as follows:

1. At the beginning of the course, the purpose of using R packages is explained to the students, ensuring they understand that the goal is to not only utilize the packages to solve exercises with software but also reinforce their comprehension of the theoretical knowledge taught. After they become familiar with the regular packages, they are introduced to the LearningRlab educational package.
2. To enhance the learning experience with the LearningRlab educational package and the descriptive statistics and probability subjects, students must first independently study the package. The fundamental concepts of descriptive statistics and probability have already been introduced in two previous theory lessons. Subsequently, students must use LearningRlab to solve the same problems they encountered during the theory lessons. It is crucial that they learn to use the package independently.
3. After students have learned and applied the package independently, the instructor guides them through the package in a second laboratory lesson to address any questions or concerns they may have regarding its use.

This approach facilitates the iterative process introduced in the previous section. The current course students evaluate the difficulties they encountered while using and learning the package, such as how the package helped them gain a deeper understanding of the theoretical concepts, any deficiencies they discovered in the package, or potential improvements they believe could be incorporated. Students are asked to provide this information in a report to be submitted to the instructor.

The outcomes of using the package are assessed in two ways, one for each subject:

- For statistics, the evaluation of the package is based on a qualitative assessment performed by the students and the results they achieve in various subject assessments throughout the course.
- For the final degree project, the evaluation of the package by the current students helps to define and identify the requirements for the new version of the package developed by the next cohort of students undertaking their final degree project.

Figure 1 illustrates the chronological progression of the implementation process as well as the relationships between both subjects: statistics and the final degree project.

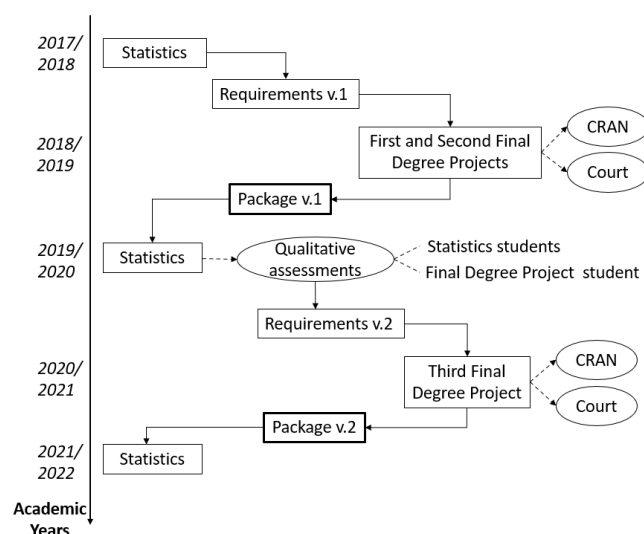


Figure 1. Overview of the implementation process of the R educational package, LearningRlab.

Methods

This section outlines the process followed to validate the impact of the R educational package, LearningRlab. A study was conducted to measure the impact of LearningRlab on

the teaching and learning process in the statistics subject. An evaluation of the influence the package has had on improving the knowledge and skills of final degree project students was assessed through their professional development, as both students are currently employed after completing their studies.

As mentioned earlier, there are two versions of LearningRlab. The requirements for the first version were derived from the teaching process of the statistics subject (see Figure 1). It was developed during the academic year 2018/2019 by two final degree project students, due to the complexity of implementing an R package from scratch. The requirements of the first version focused on descriptive statistics as it is the initial module of the course. The package was included in the CRAN before the students completed their final degree projects, as obtaining approval from the R foundation for inclusion in the CRAN was a requirement for passing the subject. Later, an experimental group of twenty-three statistics students used the package during the first semester of the 2019/2020 academic year; specifically in laboratory sessions to reinforce their understanding of descriptive statistics. They were asked to evaluate the package from an educational point of view, evaluating not only if the package correctly performs its functions, but also if the functions correctly explain the theoretical foundations of the statistical parameters implemented. They prepared and submitted a report (qualitative assessment) in which they analysed the usefulness of the package from an educational perspective. As a result of these evaluations, new requirements were identified to enhance the package in a subsequent final degree project.

On the other hand, the control group, consisting of thirty-one students, did not use LearningRlab in any laboratory sessions.

The twenty-three students (twenty males and three females) who participated in the study were asked to submit a report in which they had to use LearningRlab to solve some exercises that had been previously solved in the theory lessons. The students worked in groups of two or three to analyse the impact of the package on their learning process. Once the groups were formed and the package was analysed, the reports were submitted to be evaluated by the corresponding supervisor.

The following five questions were asked to each group of students:

1. Do you think the package is educational and easy to use?
2. Does the use of the package allow the consolidation of theoretical concepts?
3. Are the graphical simulation and results useful?
4. Is the documentation (vignettes) clear and useful?
5. How could the package be improved in future versions?

After evaluating the reports, the supervisor observed that some students identified errors in three main functions: percentile, quartile, and variance. Other modifications and improvements were also highlighted by the students. The supervisor also noticed that, in general, LearningRlab was well received among all the students. As a result, in the planned new version of the package, all these corrections were implemented, taking into account the evaluations of the statistics students. These modifications complement the planned new functions needed to reinforce the concepts related to probability, the second module of the subject.

The supervisor suggested the idea of improving an educational R package as a final degree project to the students of the last course during the second semester of the 2020/2021 academic year, with one student accepting. Before developing the new version of the package, this student considered the evaluation results of the statistics students. He made his own assessment of the first version of LearningRlab, reaching the same conclusions as the statistics students. He also considered comments from the final degree subject supervisor. The obtained package, the second version of LearningRlab, was also included in the CRAN in September 2021.

During the 2021/2022 academic year, the second version of LearningRlab was used in the statistics subject to reinforce the basic concepts of descriptive statistics and probability fundamentals. The students were asked to evaluate the second version of the package,

with prospective new students expected to implement the third version of the package, considering the evaluations of the statistics students and including new functions related to statistical inference, the third and final module of the subject. In this way, the iterative process will continue over two years when the statistics students evaluate the third version of LearningRlab. That third version would contain all the functions required by the statistics subject, so the validation process could be completed in two years, as long as positive assessments are obtained from the statistics students' reports.

3. Results and Discussion

This section presents the results obtained from the statistics students' qualitative evaluations during the 2019/2020 academic year. Additionally, the impact of using LearningRlab on the students' grades will be discussed.

From the qualitative assessment, comments were collected from the twenty-three students who worked in groups and delivered a report on their analysis of the LearningRlab package. Several comments were positive, in which students described how the package helped them gain a deeper understanding of descriptive statistics. For instance, most students highlighted the fact that the package explained in detail all the steps required to carry out the calculations: 'The functions included in the package are explained step by step' or 'The package explains in great detail the steps to follow in each function to obtain the results'. In addition, two groups of students remarked on its educational value and considered LearningRlab a good tool for students who study statistics for the first time: 'This package helps students to understand in a much simpler way how descriptive statistics works, in such a way that it is understandable for users who are not so experts in the field' and 'The advantage offered by this package compared to others is that it can also help first year students who might have difficulties with the subject'.

However, there were also other participants who found the package less useful. They remarked that the results provided by the package did not agree with the results obtained in the theory lessons: 'We have detected that the package calculates the variance, percentiles and quartiles in a different way than how we do it in class' or 'The package presents a failure according to the data that we enter in it'. After a thorough analysis of the functions implemented by the first version of the package, some bugs were found by the students in the variance, quartile and percentile functions. Furthermore, the students requested the inclusion of specific functions needed to solve the probability exercises. It is worthwhile to point out that the first version of LearningRlab was focused on descriptive statistics only. On the other hand, several comments highlighted the need for visual or graphical modifications to help users better use the package.

Therefore, the results of the qualitative evaluation of the first version of the package carried out by the statistics students can be classified as follows: correction of errors, the addition of new functions (probability functions) and graphical improvements.

3.1. Correction of Errors

The modifications made in each of the subfunctions to correct the errors found by the students are detailed next.

- **Correction in the variance function**
Based on the evaluation conducted by the students, it was observed that it is easy to make mistakes in functions such as variance, as the results may differ based on the type of data used. Population variance is calculated from population data, while sample variance is calculated from sample data. The comments received from the students led to the conclusion that the most appropriate use of variance for this package is one applied to samples, which can be used by other functions. To correct this error, it was necessary to modify the denominator of the formula by replacing 'n' with 'n-1'.
- **Correction in the quartile function**
The students pointed out that the results provided by the quartile function were sometimes incorrect and significantly different from what was expected. Upon further

analysis of the function's code, it was discovered that the way it was programmed was inappropriate. Specifically, the function calculated the quartiles based on the size of the vector, rather than its content. To fix this issue, the function was modified using the resources already present in the code, specifically the median function.

- **Correction in the percentile function**
The students also noticed that the results provided by this function were not always correct. The reason behind this was related to the errors identified in the quartile function. That is, the calculation of the percentiles was focused on the number of elements and not on the value of each element of the dataset.

3.2. Addition of New Functions

As mentioned before, the first version of LearningRlab was focused on descriptive statistics only, the first part of the statistics subject. Several students suggested the addition of new functions when answering the fifth question regarding future improvements to the package. After analysing all answers given by the students in their qualitative assessments, the supervisor realized that it would be beneficial to add more functions to the package so students could learn probability concepts while solving probability exercises.

Thus, to supplement the improvement of the package, a series of new probability functions were developed and included in the second version of the package to make it more robust and useful for the users. These include eleven widely used statistical functions that can facilitate their use and learning process in the second module of the subject related to probability fundamentals:

- Harmonic mean
- Covariance
- Pearson's correlation coefficient
- Coefficient of variation
- Laplace's law
- Binomial distribution
- Poisson distribution
- Normal distribution
- Student distribution
- Chi-square distribution
- F-Fisher distribution

It is worth noting that each of these functions has been subdivided into three subfunctions: the main subfunction, the explained subfunction, and the interactive subfunction. Consequently, a total of thirty-three functions were included in the second version of the package to cater to the improvements requested by the students.

3.3. Graphical Improvements

Regarding the answers to the third question 'Are the graphical simulation and results useful?', most students highlighted that some modifications could be introduced to facilitate the use of the package. Some identified that simple indications such as 'Please, insert your dataset' could be added in the interactive functions to minimize the possible insertions of incorrect data. Other students remarked that including an image of the implemented equations in each function would help users better understand the goal of each function.

Regarding the learning process, most students perceived that the use of the package made studying descriptive statistics easier and that their knowledge improved after using the package. They also considered that the use of the package facilitated the consolidation of the theoretical concepts, and that the user guide was useful to understand the theoretical concepts. It should be noted that LearningRlab was, in general, well appreciated among the students.

To quantify the aforementioned improvement, the learning outcomes of two groups of students (experimental and control groups) were analysed during the 2019/2020 academic course. Tables 1 and 2 show the grades of both groups of students. A grade for the

laboratory part (LP) and a grade for the theory part (TP) for each student is included in each table. The maximum grade for each part is ten points. It can be seen that the average grade of the laboratory part for students who used the package (5.76) was noticeably greater than the average grade of the laboratory part for students who did not use it (1.84). According to Figure 2, only 30% of the students in the experimental group did not deliver their laboratory report (their grade is 0.00), whereas the percentage of students in the control group was 48%. Furthermore, it is worth noting that 43% of the students in the experimental group had the maximum grade (10), whereas no one in the control group had a grade higher than 6.50. This leads to the conclusion that the use of LearningRlab also affects the motivation of the students in the laboratory sessions. These results corroborate the conclusions obtained from the qualitative assessment carried out by the students. Regarding the theory part, the average grade of both groups was quite similar: 6.21 for the experimental group and 6.29 for the control group. The reason behind this could be related to the fact that the theory test includes questions that are not entirely covered by the package. For instance, covariance, correlation, and regression functions were not implemented in the first version of LearningRlab.

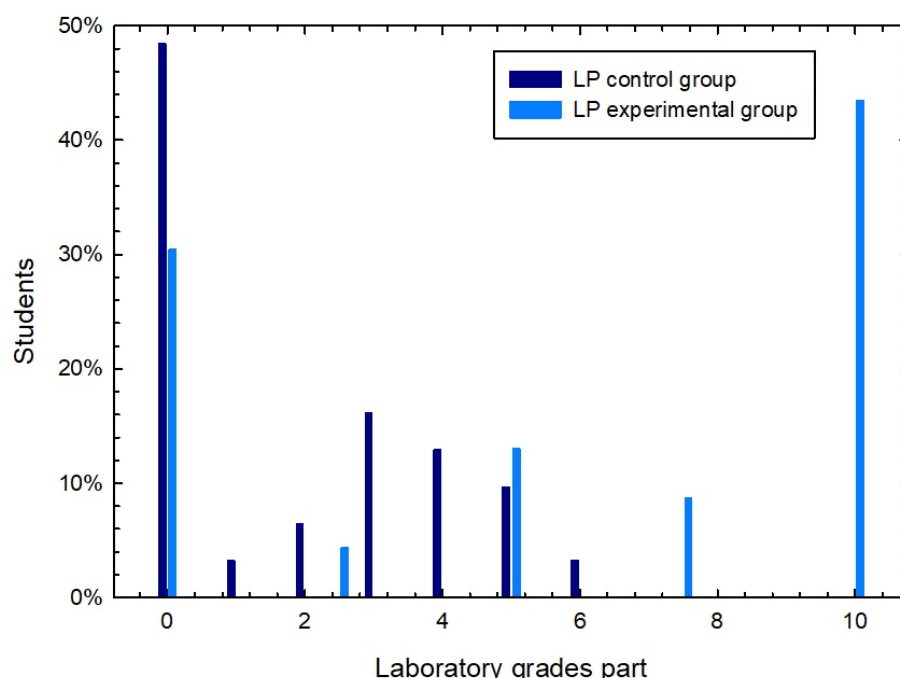


Figure 2. Comparison between the grades of the laboratory part for both groups of students.

Concerning the impact of LearningRlab's development on the final degree project students, it should be assessed qualitatively since the students completed their studies and did not participate in the paper. All of these students are currently employed in the private sector as data scientists and data analysts. As a result, it could be argued that the evaluation of the package's role in enhancing the knowledge and skills of the final degree project students is positive. Indeed, the skills and knowledge gained during their final stage of university have assisted them in securing good jobs in the fields of data management, data analysis, and data computing, which are closely related to their work as students.

Table 1. Grades of the experimental group.

-	LP	TP
S1	0.00	8.67
S2	10.00	7.00
S3	10.00	5.00
S4	7.50	7.50
S5	10.00	7.83
S6	0.00	4.50
S7	10.00	4.67
S8	5.00	4.83
S9	10.00	10
S10	0.00	3.67
S11	10.00	10
S12	5.00	4.50
S13	7.50	9.50
S14	10.00	6.50
S15	0.00	4.83
S16	10.00	5.33
S17	10.00	4.50
S18	5.00	9.50
S19	0.00	4.83
S20	0.00	3.67
S21	0.00	4.50
S22	10.00	7.83
S23	2.50	3.83
\bar{x}	5.76	6.21
σ^2	19.56	4.75

Table 2. Grades of the control group.

-	LP	TP
S1	3.00	9.67
S2	3.00	7.00
S3	3.00	8.00
S4	0.00	3.00
S5	0.00	5.67
S6	4.00	7.33
S7	4.00	6.00
S8	2.00	7.67
S9	5.00	6.67
S10	5.00	6.67

Table 2. *Cont.*

-	LP	TP
S11	4.00	5.00
S12	0.00	10.00
S13	0.00	7.33
S14	5.00	4.00
S15	0.00	5.67
S16	0.00	4.67
S17	4.00	7.00
S18	0.00	9.33
S19	0.00	3.00
S20	0.00	6.33
S21	0.00	5.67
S22	3.00	7.33
S23	0.00	5.67
S24	0.00	4.33
S25	6.00	5.00
S26	1.00	10.00
S27	3.00	5.67
S28	2.00	5.67
S29	0.00	9.00
S30	0.00	6.00
S31	0.00	1.67
\bar{x}	1.84	6.29
σ^2	4.14	4.19

Taking into consideration the research questions outlined in the introduction and the results obtained, the work has demonstrated that (1) final degree project students can develop an educational R package to teach statistics in computer science engineering, and (2) the educational R package enhances the students' learning process.

4. Conclusions

This paper highlights how an educational R package, LearningRlab, can enhance the learning process for the statistics subject in computer science and the final degree project. Developed as a final degree project by three former students, the package not only computes the main functions of descriptive statistics for a dataset but also explains the calculation steps. The package's educational value lies in providing a tool that helps students understand the content covered in theory lessons. Furthermore, since the package was developed by former students of the university, it serves to motivate current students. This work's significance lies in improving the teaching and learning process in CSE and contributing to the development of educational resources.

The paper explains how the LearningRlab package was developed in the final degree projects with its future use for the statistics subject. It also includes a description of how the package has been used in the statistics subject to reinforce concepts related to descriptive statistics and probability fundamentals. The paper also describes the main features of LearningRlab and its validation process.

There have been three different validations of the educational R package presented in this paper: acceptance into the CRAN repository, approval after public dissertation, and testing by a group of students. To validate the package from an educational perspective, a qualitative analysis was conducted. Most students reported having a deeper understanding of the theoretical concepts after working with LearningRlab. Additionally, the analysis of learning outcomes, focusing on the grades of experimental and control groups, demonstrates that LearningRlab can be used as a powerful teaching aid in the CSE statistics subject. Specifically, the average grade for the laboratory part for students who used the package (5.76), significantly higher than the average grade for the laboratory part for students who did not use it (1.84).

This work presents some limitations that could be addressed in future studies. Firstly, a questionnaire could have been distributed to students to analyze their satisfaction with the LearningRlab package and measure its influence on their motivation. Additionally, the package could have been evaluated by more students over consecutive years to assess the improvements implemented and measure its long-term effectiveness.

As future work, the LearningRlab package will be further improved in a new final degree project, taking into account the criticisms currently being made by statistics students during the 2021/2022 course. The third version of the package could also include functions related to statistical inference, the final module of the subject, thus allowing the package to cover all required analyses throughout the entire semester. Additionally, a broader study could be conducted with students at other universities in Spain and/or other countries to investigate the effectiveness of LearningRlab in enhancing the teaching and learning of statistics. Moreover, the impact of LearningRlab on students' motivation in laboratory sessions could be analysed to provide insights into its potential role in promoting engagement and interest in the subject.

Author Contributions: Conceptualization, J.J.C.-G.; methodology, J.J.C.-G., J.G. and A.T.; software, J.J.C.-G.; writing—original draft preparation, J.J.C.-G., J.G. and A.T.; writing—review and editing, L.U., C.J.H. and A.V.; project administration, J.J.C.-G.; funding acquisition, J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the program “Programa de Estímulo a la Excelencia para Profesorado Universitario Permanente” of Vice rectorate for Research and Knowledge Transfer of the University of Alcalá and by the Comunidad de Madrid (Spain) through project EPU-INV/2020/004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors are grateful to the students who participated in the LearningRlab package, Carlos Javier Hellín Asensio [aut, cre], Jose Manuel Gómez Cáceres [aut], Dennis Monheimius [aut], Eduardo Benito [aut].

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

This appendix includes the main characteristics and intended use of all the functions implemented in the first version of LearningRlab.

- **Main functions.** The functions included in this set receive parameters from the user and perform the desired calculation. They are similar to the descriptive statistics functions included in the core of R and the main statistical packages. For final degree project students, these functions are intended to help them start implementing the required functions in an easier way, preparing them for the complete programming of the following interactive and explained function definitions. For statistics students, these functions allow them to become familiar with their names and usefulness. The

functions, all of them descriptive statistics of one variable, that are part of this set are the following:

- [frequency_abs](#). It calculates the number of times a specific number appears in the dataset.
 - [frequency_relative](#). It calculates the number of times a specific number appears in the dataset divided by the total number of data in the set.
 - [frequency_absolute_acum](#). It calculates the absolute accumulated frequency of a certain value, calculated as the sum of the absolute frequency of the values that are lower or equal to an indicated value.
 - [frequency_relative_acum](#). It calculates the relative accumulated frequency of a certain value, calculated as the sum of the absolute frequency of the values that are lower or equal to an indicated value divided by the total number of data in the set.
 - [mean](#). It calculates the arithmetic mean of the dataset.
 - [mode](#). It calculates the mode of the dataset.
 - [median](#). It calculates the median of the dataset.
 - [percentile](#). It calculates the percentiles of the dataset.
 - [quartile](#). It calculates the three quartiles of the dataset.
 - [variance](#). It calculates the variance of a vector of data.
 - [standardDeviation](#). It calculates the standard deviation of the dataset.
 - [averageDeviation](#). It calculates the average absolute deviation of the dataset.
- Explained functions. This second set of functions developed for LearningRlab give the package a distinctive edge over other packages, making it more educational in nature. The input parameters for these functions are the same as those for the main functions. However, the key difference lies in the fact that in addition to providing the result of the function, an explanation of how the result is calculated is provided. These functions have been explicitly designed to teach users the process involved in performing statistical calculations, using examples customized to the parameters being used. This enables users to understand how statistical parameters are calculated step-by-step. Another important feature of these functions is the inclusion of visual effects such as coloured output and formula plots, making colour a useful pedagogical tool for improving knowledge understanding. The functions in this group are identical to those in the previous set of main functions and work with the same statistical parameters, with the prefix [explain](#) added to each of their names. For example, the [explain.absolute_frequency](#) details and explains the steps carried out to calculate the number of times a specific number appears in the dataset. The rest of the explanation functions are: [explain.frequency_relative](#), [explain.absolute_acum_frequency](#), [explain.relative_acum_frequency](#), [explain.mean](#), [explain.mode](#), [explain.median](#), [explain.percentile](#), [explain.quartile](#), [explain.variance](#), [explain.standardDeviation](#), and [explain.averageDeviation](#).
 - Interactive functions. This is the third group of functions included in the educational LearningRlab package and are intended to reinforce the knowledge of the statistical parameters introduced by the two previous groups of functions. The functions in this group are designed to interact with the user, using self-test questions to allow the students to check if they have understood the process that must be carried out to obtain the calculation of the requested value. A series of data is requested from the user and, based on this, the correct solution is requested. Interactive functions do not need input parameters. The function itself asks the user to enter the required parameters depending on the function. Once the user has provided the required data, the function asks the users to enter the result to check if they have understood how the function operates. Finally, the function itself will provide the correct result so that users can check if their results are correct. The functions part of this group are the same as for the previous two sets, and in subsequent work over the same statistical parameters, but with the word "interactive" before all of them. The interactive functions

are as follows: [interactive.absolute_frequency](#), [interactive.relative_frequency](#), [interactive.absolute_acum_frequency](#), [interactive.relative_acum_frequency](#), [interactive.mean](#), [interactive.mode](#), [interactive.median](#), [interactive.percentile](#), [interactive.quartile](#), [interactive.variance](#), [interactive.standardDeviation](#) and [interactive.averageDeviation](#).

Appendix B

This appendix includes the new functions implemented into the second version of LearningRlab.

For descriptive statistics of one variable:

- [harmonicMean](#). This function calculates the harmonic mean of a dataset.
- [cv](#). This function calculates the coefficient of variation of a dataset.

For descriptive statistics of two variables:

- [covariance](#). This function calculates the covariance of two datasets, each corresponding to a different variable.
- [pearson](#). This function calculates the Pearson correlation coefficient of two datasets, each corresponding to a different variable.

For probability:

- [laplace](#). This function applies the Laplace's law to a dataset. It has two input parameters: a vector of the data and the value considered as the favourable case.
- [binomial](#). This function applies the discrete Binomial distribution defined by 'n', the number of trials, and 'p', the probability of success, to a certain value 'x'. It has three input parameters: 'n', 'p' and 'x'.
- [poisson](#). This function applies the discrete Poisson distribution defined by 'lam', which is a positive value that represents the number of times a phenomenon is expected to occur during a given interval, and 'k', which is the number of occurrences. It has two input parameters: 'lam' and 'k'.
- [normal](#). This function applies the discrete normal distribution, characterized by a mean equal to zero and a standard deviation equal to one, to a certain value 'x'. It has one input parameter: 'x'.
- [tstudent](#). This function applies Student's t distribution defined by 'x', the sample mean, 'u', the population mean, 's', the population standard deviation, and 'n', the sample size. It has four input parameters: 'x', 'u', 's' and 'n'.
- [chisquared](#). This function applies the chi-square distribution to two datasets, which are given as two vectors.
- [fisher](#). This function applies the continuous Fisher F distribution to two datasets, which are given as two vectors.

The new explained functions are the following: [explain.harmonicMean](#), [explain.cv](#), [explain.covariance](#), [explain.pearson](#), [explain.laplace](#), [explain.binomial](#), [explain.poisson](#), [explain.normal](#), [explain.tstudent](#), [explain.chisquared](#), and [explain.fisher](#).

Furthermore, the LearningRlab version 2 interactive functions are the following: [interactive.harmonicMean](#), [interactive.cv](#), [interactive.covariance](#), [interactive.pearson](#), [interactive.laplace](#), [interactive.binomial](#), [interactive.poisson](#), [interactive.normal](#), [interactive.tstudent](#), [interactive.chisquared](#) and [interactive.fisher](#).

References

1. Gómez, J.M.; Monheimius, D.; Benito, E.; Cuadrado-Gallego, J. Introduction to LearningRlab. Available online: <https://cran.r-project.org/web/packages/LearningRlab/vignettes/learningRlab.html> (accessed on 3 April 2023).
2. Frick, H.; Kosmidis, I. trackerR: Infrastructure for Running and Cycling Data from GPS-Enabled Tracking Devices in R. *J. Stat. Softw.* **2017**, *82*, 1–29. [[CrossRef](#)]
3. Leodolter, M.; Plant, C.; Brandle, N. IncDTW: An R Package for Incremental Calculation of Dynamic Time Warping. *J. Stat. Softw.* **2021**, *99*, 1–23. [[CrossRef](#)]
4. Role, F.; Morbieu, S.; Nadif, M. CoClust: A Python Package for Co-Clustering. *J. Stat. Softw.* **2019**, *88*, 1–29. [[CrossRef](#)]

5. Millard, S. EnvStats: An R Package for Environmental Statistics. Available online: <http://www.springer.com/book/9781461484554> (accessed on 3 April 2023).
6. Tucker, M.C.; Shaw, S.T.; Son, J.Y.; Stigler, J.W. Teaching Statistics and Data Analysis with R. *J. Stat. Data Sci. Educ.* **2023**, *31*, 18–32. [CrossRef]
7. Ugarte, M.; Militino, A.; Arnholt, A. *Probability and Statistics with R*, 2nd ed.; Chapman and Hall/CRC: New York, NY, USA, 2015. [CrossRef]
8. Verzani, J. *Using R for Introductory Statistics*, 1st ed.; Chapman and Hall/CRC: New York, NY, USA, 2004. [CrossRef]
9. Dalgaard, P. Package ISwR: Introductory Statistics with R. Available online: <https://CRAN.R-project.org/package=ISwR> (accessed on 3 April 2023).
10. Signorell, A. Package DescTools: Tools for Descriptive Statistics. Available online: <https://cran.r-project.org/web/packages/DescTools/index.html> (accessed on 3 April 2023).
11. Erdely, A.; Castillo, I. cumstats: Cumulative Descriptive Statistics. Available online: <https://cran.r-project.org/web/packages/cumstats/index.html> (accessed on 3 April 2023).
12. Mostad, P. lestat: A Package for Learning Statistics. Available online: <https://cran.r-project.org/web/packages/lestat/index.html> (accessed on 3 April 2023).
13. Navarro, D. lsr: Companion to “Learning Statistics with R”. Available online: <https://cran.r-project.org/web/packages/lsr/index.html> (accessed on 3 April 2023).
14. Venables, W.; Smith, D. Documentation of the R Core. Available online: <https://cran.r-project.org/manuals.html> (accessed on 3 April 2023).
15. Burckhardt, P.; Nugent, R.; Genovese, C.R. Teaching Statistical Concepts and Modern Data Analysis With a Computing-Integrated Learning Environment. *J. Stat. Data Sci. Educ.* **2021**, *29*, S61–S73. [CrossRef]
16. Gerbing, D.W. Enhancement of the Command-Line Environment for use in the Introductory Statistics Course and Beyond. *J. Stat. Data Sci. Educ.* **2021**, *29*, 251–266. [CrossRef]
17. Yuanling, L. Web-Based Applets for Facilitating Simulations and Generating Randomized Datasets for Teaching Statistics. *J. Stat. Data Sci. Educ.* **2022**, 1–9. [CrossRef]
18. Clark, R.C.; Nguyen, F.; Sweller, J.; Baddeley, M. *Efficiency in Learning: Evidence-Based Guidelines to Manage Cognitive Load*; John Wiley & Sons: Hoboken, NJ, USA, 2006; Volume 45, pp. 46–47.
19. Bates, D. The R Project for Statistical Computing. Available online: <https://www.r-project.org/> (accessed on 3 April 2023).
20. Gómez, J.M.; Monheimius, D.; Benito, E.; Cuadrado-Gallego, J. LearningRlab Package Manual. Available online: <https://cran.r-project.org/web/packages/LearningRlab/LearningRlab.pdf> (accessed on 3 April 2023).
21. Gómez, J.M.; Monheimius, D.; Benito, E.; Cuadrado-Gallego, J.J. Package LearningRlab. Available online: <https://cran.r-project.org/package=LearningRlab> (accessed on 3 April 2023).
22. Gelman, A.; Nolan, D. *Teaching Statistics: A Bag of Tricks*; Oxford University Press: Oxford, UK, 2017. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.