UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE,
ECONOMICHE E SOCIALI

CORSO DI LAUREA IN ECONOMIA E MANAGEMENT

CLASSE L-33 – SCIENZE ECONOMICHE

DATA in SCHOOL DIGITALIZATION: IMPROVING STUDENTS' PRESENT AND TOMORROW'S EDUCATION

Relatore:

Prof. Giancarlo Manzi

Tesi di laurea di:

Luisa Ferrari

Matr. 903320

Anno Accademico 2019 – 2020

Index

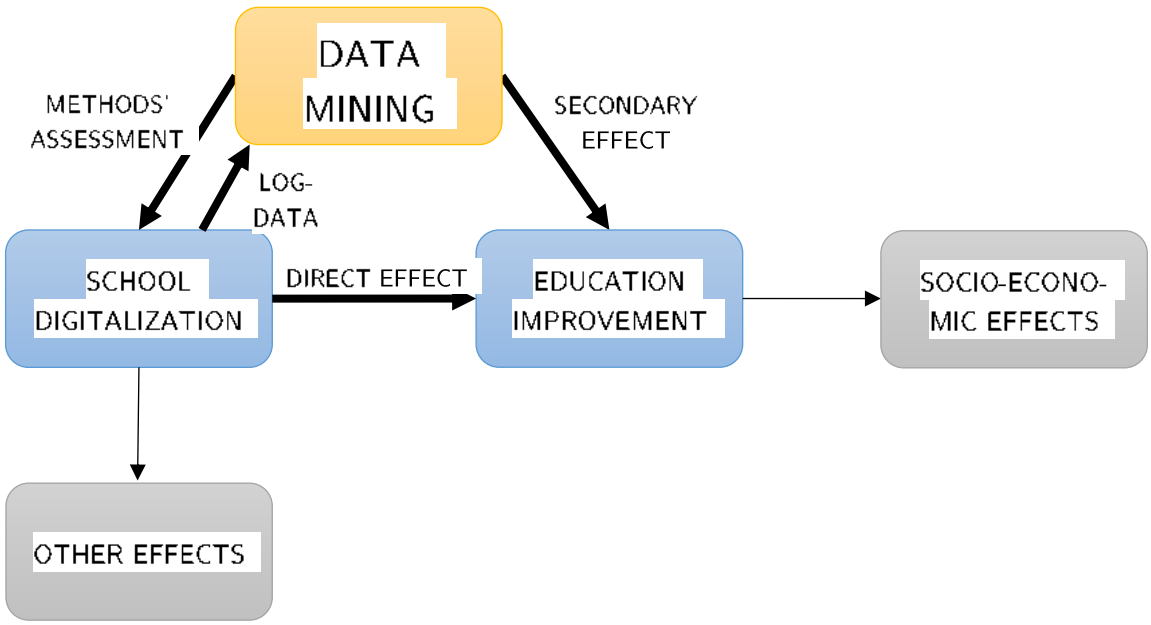
1 Introduction

The introduction of technological devices in learning provides the educational systems around the world with opportunities and challenges. The direction that digitalization should follow in this sensitive field does not appear clearly, but more or less courageous attempts keep being made all over the world. The topic attracts many criticisms, regarding every aspects of the issue. That is because the radical change that underlies digitalization does not only consist in a shift towards technological devices from paper-based and frontal lectures content: in fact, this process has the potential to undermine traditional teaching and learning methods, with obvious consequences on future generations' knowledge and skills.

The basic assumption on which this paper relies is that this transformation process is already taking place, although at different rates around the world, and it is likely to take root. This assumption is based on the evidence that this dissemination phenomenon has already taken place in a number of other areas of society, such as business. The purpose of this paper is to investigate the value and the relationship between two potential effects of school digitalization, both related to data analysis.

The main effect, which is easily conceivable, is the one on students' performance, which is perceived as the primary advantage that can be achieved through this whole process. If the effect proved positive, it would cause in turn socio-economic benefits both in the short and, more importantly, in the long term. If evidence supported the hypothesis of a positive effect, policy makers should be compelled to take into account these potential, and fairly noteworthy, advantages. However, as said before, digital learning is a rapidly growing sector in which a vast number of approaches coexist. Whether the effect is positive, negative, or uninfluential, it is highly likely to depend on the type of methods employed. Although it can be argued that other criteria should be employed as well in the evaluation of a learning method or approach, data analysis can provide decision-makers with useful evidence by estimating its impact on performance.

Figure 1.1 School Digitalization's Effects



This is where data firstly appears in the mechanism revolving around school digitalization: it can be used to test hypotheses and provide evidence to support decisions.

The first part of the paper consists in the analysis of data coming from an experimental study, conducted specifically for this research project, with the aim of assessing the impact of a fairly simple digital application on students' performance. This is used as an example in order to show how data can provide additional and useful information on the value of these services.

> **Research Question 1:** ***Does school digitalization have an effect on students' performance?***

Such analyses are already employed in educational research in order to evaluate the best methods among the traditional ones. However, when technological devices are employed, there is a game-changing difference: most of the data needed for research is automatically collected by online platforms or programs, thereby making it readily available. This incredibly reduces the cost of collecting and accessing data. Moreover, the collected information does not represent just a

randomly chosen sample, but in theory all of the observations are available and new data is continuously collected at every user access: datasets become extremely larger and dynamic. Thereby, the overall process of assessing a method's effectiveness becomes much less time-consuming, having almost immediate feedback, and the costs of conducting rigorous research drop dramatically. Moreover, innovation in digital learning will provide more and more accurate and appropriate information about students, which will make the data more precise and representative.

The second part of the paper focuses on this particular aspect: the potential of the large amount of data collected by digital devices in education. This is usually an underestimated, if not utterly neglected, benefit of school digitalization. By providing additional, and otherwise unknowable information, data mining can unleash a number of socio-economic benefits at every level of the education structure, from individual students to the district-level, to the entire system.

Starting from an economic point of view, the purpose of this paper is to show how data mining, only possible thanks to school digitalization, can be conscientiously employed in order to respond to the specific needs of an education system. The potential of educational data mining and its consequences should be therefore considered by policymakers when assessing the cost-effectiveness of a school digitalization policy.

Research Question 2*: How can data mining help education at its different levels?*

In summary, school digitalization causes two potential effects related to data: it facilitates the assessment of learning methods providing indication for future changes; it collects large amount of data that may suggests policy or method shifts, regardless of the technological aspect. Both these effects aim in the same direction that is the education system improvement, which in the long term is going to affect the socio-economic stability and productivity of a country.

1.1 Log-Data in Education

The subject of this paper is to prove the potential of log-data whose collection is an inevitable consequence of school digitalization.

As said before, the exploitation of log-data may have two major effects in the mechanism. The main distinct features of log-data regard the temporal dimension.

Firstly, data are immediately collected at every user access. Therefore, the dataset is dynamic. In addition to the reduction of time and money spent collecting data, this allows for the usage of new techniques that can learn from the data and adjust according to their evolution in time. These machine learning methods are already employed in the business and social network fields with extraordinary results, and can be employed in education as well, especially at the student level. User-based tutoring strategies are just an example of how machine learning can be employed in order to help students.

Secondly, log-data usually collect information about time during each session. For instance, the time spent on a problem before answering can be measured. These variables, although easy to measure for a single student, would be hardly available at a low cost and at a large scale otherwise. The precise and again, evolving, results that log-data offer can be used to create evaluation metrics that are not commonly employed, such as reading or problem-solving speed, readiness, and their evolution in time.

In conclusion, log-data not only will accelerate the process of evaluating learning methods but will also allow for the development of new, more sophisticated ones, that could take into account original ways of assessing performance and improvement.

1.2 Data as a Check

The first part focuses on analysing the results of a randomized controlled experiment, designed with the aim of assessing the effect of online tutoring strategies. In this first application, data will be used in order to test a hypothesis and, therefore, corroborate the theory supporting the usefulness of digital learning methods. The study was conducted in sixth-grade classrooms in the USA.

In this part, a deductive approach is employed since the starting point is a hypothesis which is tested based on evidence. This application represents the immediate advantage that data can provide in school digitalization.

A traditional method in educational research such as multilevel modelling is employed so that the model will estimate the coefficients while considering effects both at the student-level and the class-level.

However, the collected data, being log-data, differ from traditional datasets in that it contains a number of variables that would be hardly measurable otherwise, such as response time to each question. These variables allow us to include original metrics in the multilevel analysis, in order to create a more accurate model and control for more external effects. The analysis will lead to an estimate of the effect of online tutoring strategies, its confidence interval, and its statistical significance.

1.3 Data as a Tool

In addition to the direct effect that digitalization may have on the educational system, there is a more subtle benefit that can come from the exploitation of the data, especially of the log-data collected over time.

As businesses first envisioned the upcoming world of the information technology and now are increasingly recognizing the value of data analysis, the education system is likely to come across this huge opportunity in the near future.

Digitalization is not a cost-free process, therefore, wasting some of the value it contains or delaying its exploitation seems nonsense. However, data analysis is also an expensive and time-consuming activity, whose results and benefits may often only be seen in the long-term. Therefore, it is fundamental to assess the priorities of the education system at each of its levels in order to carry out analyses that specifically address these needs.

Additionally, data mining techniques should be adjusted to the context of education, whose features as well as data differ from the ones typical, for instance, of the business environment, where statistics is intensively employed.

Educational data mining is a rising, interdisciplinary field of inquiry, whose primary aim is to develop and refine statistical and machine learning techniques for the exploration of educational data. Classic data mining techniques have been

adapted to this type of data and its common hierarchical structure, in order to gain further insights about different aspects of education. These techniques have proved useful in a wide range of applications regarding the fields of teaching methods, learning approaches, and course management.

Gaining further knowledge about the functioning of the education system will help identify its fragilities and shortcomings and improve its quality. Instructors and policy makers will be provided with much more information in the form of easy-to-understand tools such as summary metrics and visual content, thus reducing the uncertainty they face in their decisions.

In the second part of the paper, a popular method in Educational Data Mining will be applied to a case study. The approach will be different from the traditional one employed in the first application. As in business data mining, the analysis will start from the identification of a specific need of the client (which could be a school, a district, or a teacher) and it will continue with the choice of a strategy or method (e.g. unsupervised data mining techniques) that allows to retrieve the information of interest from the available data. The insights will then be used to describe accurately the situation to the client and to develop recommendations for an appropriate solution.

2 Educational Data Mining

Educational data mining is defined by the most prominent researchers in the filed as:

> ***"the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational settings and using those methods to better understand students and the settings which they learn in".*** [1]

Specifically, educational data mining differs from traditional data mining because it tries to develop methods that work especially for and exploit the multiple levels of the hierarchy which often characterizes educational data (e.g. problem, assignment/test, student, class, school, district levels)

Data mining is the field of inquiry in which methods and techniques are employed in order to extract meaningful and hidden information from data that could potentially affect decision-making. Its application in several fields has made it clear that it has the potential to deeply transform the way research inquiry is conducted. Applying data mining to the educational setting was not only encouraged by the excellent results of the methodology in other fields, but also necessary given that students data, thanks to technological devices and digitalization, grow larger and larger, exceeding the human ability to extract useful information from them.

Educational data mining is a multidisciplinary field, which uses theories from the learning sciences, education philosophy, and psychometrics literature, and methods from the data mining, statistics, and machine learning fields. Most of the researchers use a theory-oriented approach that guides their choices during analyses: this is a distinctive feature of educational data mining which diverges from the classic data-driven approach employed in other fields.

2.1 History

The field of educational data mining is relatively new: the first journal entirely dedicated to the topic, *Journal of Educational Data Mining* [2], began publications in 2009, preceded by few conferences in the early 2000s. Other venues have

been established and publish regularly research in this area: *Journal of Learning Analytics; International Conference on Educational Data Mining; Conference on Learning Analytics and Knowledge; International Conference on Artificial Intelligence in Education;* and others. [3]

Two different societies, IEDMS [4] and SoLAR [5] have been established around 2010: they both promote the use of analytics in education and serve the purpose of bringing together technical, statistical, psychological, and pedagogical skills to improve learning and education.

The main reasons for the growth of educational data mining are definitely related to the increasing quantity of data available for research and the development of computational and analytical tools to analyse it.

The amount of data stored has been growing exponentially since the employment of digital devices and online platforms in schools and other educational contexts.

The main sources of data for EDM include traditional ones, such as surveys and questionnaires, but the field focuses on the exploitation of the so-called log-data, which differs from offline educational sources in a variety of aspects. Sources of educational log-data can be categorized as follows:

- **E-Learning Platforms and Learning Management Systems.** Students are provided with content, instructions, and communication. Reporting is available for teachers.
- **Intelligent Tutoring Systems and Adaptive Educational Hypermedia Systems.** The learning process is customized on the basis of each student's profile.

However, the first analyses on log-data used to spend an incredible amount of time just transforming the files into a readable format. Nowadays, this is not an issue any more since standardized formats have been developed in order to log and store effectively students' data.

Moreover, educational data has become more accessible to researchers thanks to the creation of public online repositories such as the PSLC DataShop [6] and the National Center for Education Statistics [7]. The main advantage of the uploading of log-data in public repositories is that datasets can be used to answer research questions even if they are fairly different from the ones the original study was set up for: on contrast to traditional surveys, log-data has generally a standardized

format, which generally includes a wide variety of all of the important features that can be collected through e-learning.

Conversely, theories and models that prove feasible on a dataset can be transferred to new datasets in order to be validated and generalized. It is common procedure to replicate the same data mining analysis in more datasets, differing in terms of educational context or learning systems.

Finally, the large amount of data which is collected in fairly similar learning contexts definitely helps understand how contextual factors influence the learning process and students' performance.

On the other hand, statistical and data mining tools, which are becoming increasingly easier to use, have been accessible for researchers, making it easy to store and analyse data. Thus, more and more learning sciences researchers have been driven towards the inclusion of data mining techniques into their studies.

2.2 Educational Data Mining and Learning Analytics

Two research communities, often overlapping, have been growing in the area of data mining in education, whose perspectives are different but complementary. They share similar values and goals, but they have distinctive traits with respect to methodological and ideological approaches. [8]

- **Educational Data Mining (EDM).** It is usually more interested in automated methods and theoretical, technical approaches. Its main applications aim at automated adaptation, such as user-based educational software that learns from student's performance and adapts the content to individual needs. The emphasis in modelling is on breaking down the phenomenon of interest into basic components and analysing the relationships and the interactions among them. Techniques that are widely popular within the EDM community include prediction, clustering, and Bayesian models.
- **Learning Analytics (LA).** The ultimate goal is to inform and empower teachers and learners (leveraging human judgement). Researchers in this community are also more interested in understanding systems with a holistic approach. The most preferred methods are social network analysis, sentiment analysis, concept analysis, and sensemaking models.

However, the term *educational data mining* is commonly used to refer to the interdisciplinary research field in which data science techniques are employed in an educational setting.

2.3 Main Methods and Techniques

EDM employs different data mining techniques in order to achieve its purposes. Literature has categorized the most common EDM methods as described in [3]. The first three categories of methods are classic techniques employed in data mining, while the last two are typical approaches adopted in EDM.

Prediction

The goal is to predict the value of a target variable, given a set or combination of other variables called predictors. The main types of prediction used in EDM are classification (the target variable is binary or categorical), regression (it is continuous), and density estimation (the target is a probability density function). Traditional EDM applications of prediction models are used to predict students' educational outcomes, such as final performance in terms of score or speed, probability of failure, or probability of school dropout.

When using prediction models, it is essential to consider the non-independence due to the hierarchical structure of educational data.

Structure Discovery

It represents an exploratory approach to the data with the aim of identifying patterns, groups, similarities either in the observations or variables. Common methods include:

- **Clustering.** The goal is to find groups made up by observations that naturally group together because of their feature similarities. It has been used both to group students and students' actions, as well as schools at a higher level.
- **Factor Analysis or Principal Component Analysis.** The goal is the same as in clustering, but the interest is on variables rather than observations. The features are turned into a new set of latent factors.
- **Social Network Analysis.** It develops a model about relationships and interactions among members of a group.

- **Domain Structure Discovery.** The goal is to "*find the structure of knowledge in an educational domain*". [3]

Relationship Mining

Its aim is to find out meaningful and strong relationships between variables. This category contains some of the most popular methods in EDM. These methodologies help finding relevant factors and it is used for feature selection and extraction. Four main types of relationship mining are commonly employed.

- **Association Rules.** It attempts to find rules such as if a set of variable values is present, then another variable will take a specific value. (i.e. {if→then} rules)
- **Correlation.** It analyses the strength and signs of linear correlations among variables.
- **Sequential Pattern Mining.** It attempts to find associations between events in time.
- **Causal Data Mining.** It attempts to find out whether an event has triggered or caused another one, often using covariances.

In order to be considered relevant, relationships have to prove both statistically significant and interesting. Statistical significance is assessed to understand whether the relationship or pattern was due to chance or not: statistical tests are usually employed, such as F-tests. On the other hand, interestingness measures the strength of the relationship, in order to reduce the set of rules to the strongest and most supported by the data: the most common interestingness measures in EDM are lift and cosine.

Discovery with Models

The outcome of a model (often a prediction model) is used to conduct a second analysis and to create a final model. The most popular way to conduct discovery with model employs two different prediction models, the latter using the predictions of the former as a new predictor. Another type of discovery with models uses predictions to search for relationships with other variables. Discovery with models may also employ knowledge engineering models (i.e. human-made models).

Distillation for Human Judgement

It is one of the most important tasks of EDM, which is to provide teachers and educators with useful, comprehensible information about their students. This is commonly done through data visualization methods such as heatmaps, scatterplots, learning curves, or learnograms, in which the number of opportunities to practice is plotted against a performance measure. This approach is most common in LA research. Distillation is used for two key purposes: classification and identification. When the purpose is classification, distillation can be seen as an exploratory phase for the development of a prediction model. On the other hand, identification is the process of displaying the retrieved information to human judgement for pattern recognition or labelling.

2.4 EDM Applications

The ultimate goal of educational data mining is not only research and knowledge enrichment: in fact, having a positive impact on learners and learning environments and process is the final aim. The tasks that EDM can mostly contribute to have been categorized into five categories [9]:

- Evaluation of student performance and learning process;
- Development of adaptive learning and intelligent tutoring systems based on learners' individual behaviour;
- Assessment of material and content in online courses;
- Development of valuable feedback systems for instructors and learners;
- Detection of unusual student behaviours.

Data mining techniques can be used to predict students' performance more accurately thanks to the account of new features and factors that would be unknowable without log-data. Before predicting specific outcomes, it is essential to develop accurate student or learner models that explain which factors (and how much) affects individual behaviours. The development of a robust student model is one of the primary tasks of EDM.

Once relevant features have been identified, other models can be developed, for example, to predict final scores or drop-out probability, or for data visualization techniques. These methods can be employed to build systems that assess students' behaviour and respond intelligently and adaptively to their own needs.

On the other hand, information can provide feedback support to teachers about the trends and composition of their pools of students. For instance, content offered in different media types can affect particular students' categories more than others: how well each category responds to different media can be useful to teachers to understand which types of content are more appropriate to categories of students. [10]

2.5 Present and Future

Educational data mining has already proved its value in a variety of applications. In particular, there are some areas where EDM has heavily impacted on the understanding of some aspects related to learning sciences. For example, a number of papers have been published regarding the issue of student disengagement detection in which automated models have been proposed that allow to detect specific students' behaviour. [11]

In general, the employment of data mining has given positive feedback in learning sciences, especially with discoveries and evidence that led to practical changes and improvements in the educational setting. In turn, new strategies and teaching approaches can be studied to find out more about new questions that may arise.

One of the main consequences of EDM has also caused to raise awareness about the impact of social factors, contexts, and learning environments on the learning process. This is likely to influence the way e-learning, as perhaps traditional learning, is structured and conducted.

The field of educational data mining continues to expand, with more and more research questions rising and searching for data-supported answer. The range of settings and levels in the educational systems is widening as researchers trying to apply data mining techniques to unexplored aspects of education. The development in recent years also suggests that the field is likely to grow fast in the years to come. The publication trends show that research in the field has been constantly growing since 2010 [12]. Educational data mining is likely to thrive in the following years and to unleash its full potential. Given the positive results achieved so far, it is certainly a powerful tool in the hands of researchers and learning scientists, as well as a supportive partner for teachers and instructors. It has the power to support and improve both theory and practice in the educational field.

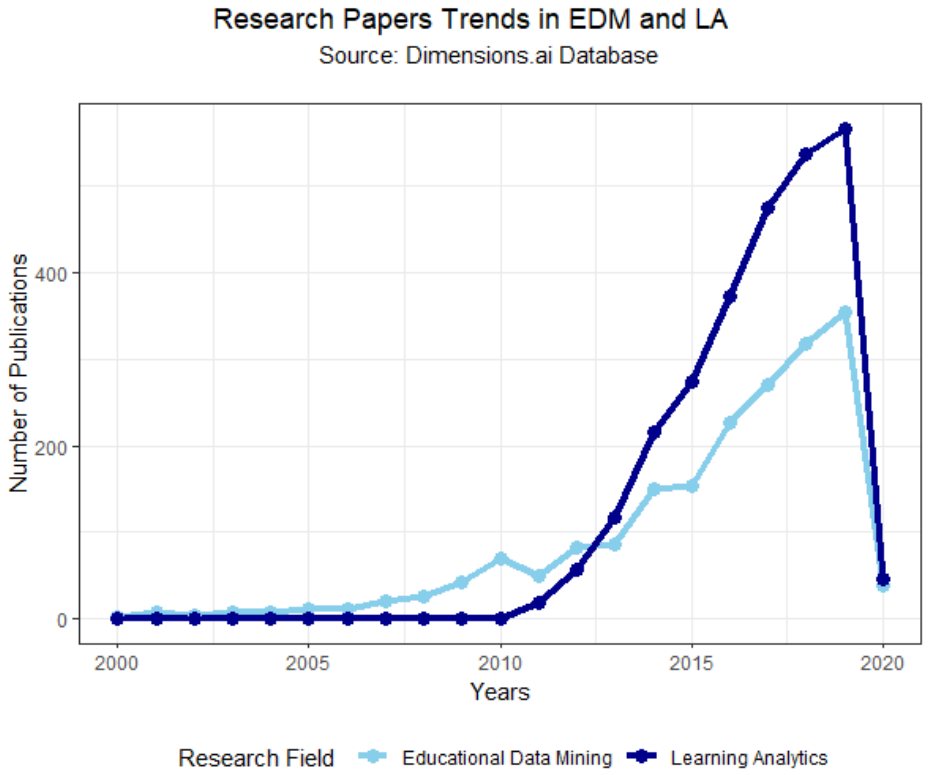


Figure 2.1 Literature Trends of EDM and LA

With respect to the future, the state of the art in data science is the employment of deep learning techniques that have achieved incredible results in a variety of applications, such as image recognition and natural language processing. [13]

Machine and deep learning techniques have already been employed in educational data mining, for instance, to perform handwriting recognition for e-learning systems [14]. Other applications that have been outlined regard adaptive learning systems based on individual students' behaviour. Although these algorithms produce accurate models, particularly in the prediction field, interpretability is definitely not one of their strength points: it is difficult to interpret the results in terms of the impact of each features involved, as well as to extract cause-effect relationships, which then prevents from having a "story" that easily explains the model and can be understood even by those who are not familiar with data mining (e.g. teachers). Therefore, an issue raises since EDM is particularly sensitive to interpretability, given its main purposes and stakeholders. However, as EDM is making its progress, methods for deep learning interpretability are being developed as well, so that hopefully these powerful tools will be successfully employed in the educational tasks and applications that most would benefit from it, which include text mining and behaviour prediction.

3 Impact of School Digitalization: An Experimental Study

3.1 ASSISTments by WPI

The purpose of this experiment is to assess whether the use of a particular digital implementation during homework has an impact on students' learning.

Specifically, the experiment was conducted in collaboration with ASSISTments, which is an online learning platform, developed in 2003 by Neil and Christina Heffernan, and now supported by the ASSISTments Foundation of Worcester Polytechnic Institute. [15]

ASSISTments is an online platform which is completely free of charge for students and teachers. It offers homework content that teachers can assign to their classes. Although it is mainly used for math homework assignments, the platform now contains many problem sets related to other subjects. Teachers are also allowed to create, develop, and assign their own content. It was initially used only by K-12 teachers, but now many colleges and universities use it to assign content to their students. ASSISTments assists students during their homework, with immediate correctness feedback and tutoring strategies, while assessing them and providing teachers with real time data about their class' performance.

On the other hand, ASSISTments' ultimate mission is to improve education through scientific research, by providing researchers and scientists with a useful tool to create randomized controlled trials to run on the platform. Studies are created to be minimally disruptive for students and teachers' experience and not to compromise students' learning time. 18 studies were published on randomized controlled experiments run on the platform.

Teachers can voluntarily participate in the experiment or just assign the problem set developed for the study as a regular assignment: students' experience does not differ from normal content assignments.

ASSISTments also has a large database containing the results for each problem set: this offers a great amount of data useful for analysis in the field of learning sciences.

In 2014, on average 4,000 students used the platform each weekday: about half of them used it for homework, while the other half during schooltime. [15]

Last year, more than 10 million problems were solved using ASSISTments, which is currently used in 46 states of the US and in 14 countries. In 2019, more than 2,500 teachers used ASSISTments content. [16]

3.2 Randomized Controlled Experiment

The following study was developed with the purpose of investigating whether simple tutoring strategies during homework could help students' learning process and enhance their future performance. In order to accomplish this goal, a randomized controlled experiment was designed for the purpose. The digitalization method that is assessed in this study is online tutoring strategies that are available on ASSISTments platform, as well as in other online learning tools. These strategies include hint requests and scaffolding strategies, which provide the breakdown of the problem into steps through which the student is led towards the solution. ASSISTments is widely popular for those features, together with the Skill Builder option. In this study, the purpose is to assess whether students learn faster when they have access to these types of help tools during their math homework.

A randomized controlled trial is an experiment whose aim is to test the effect or impact of a specific condition, by controlling for other factors, hence reducing possible sources of bias. In order to achieve this aim, subjects are randomly assigned to two (or more) groups, which then undertake different treatments: by randomly allocating subjects with different characteristics, it reduces selection bias and confounding. In general, one of the groups does not receive any treatment, and therefore, it is called control group, while the other is called treatment group. In order to test the efficacy of the treatment, the results of the two groups are statistically compared. Randomized controlled trials (RCT) are popular in the medical and in the social sciences.

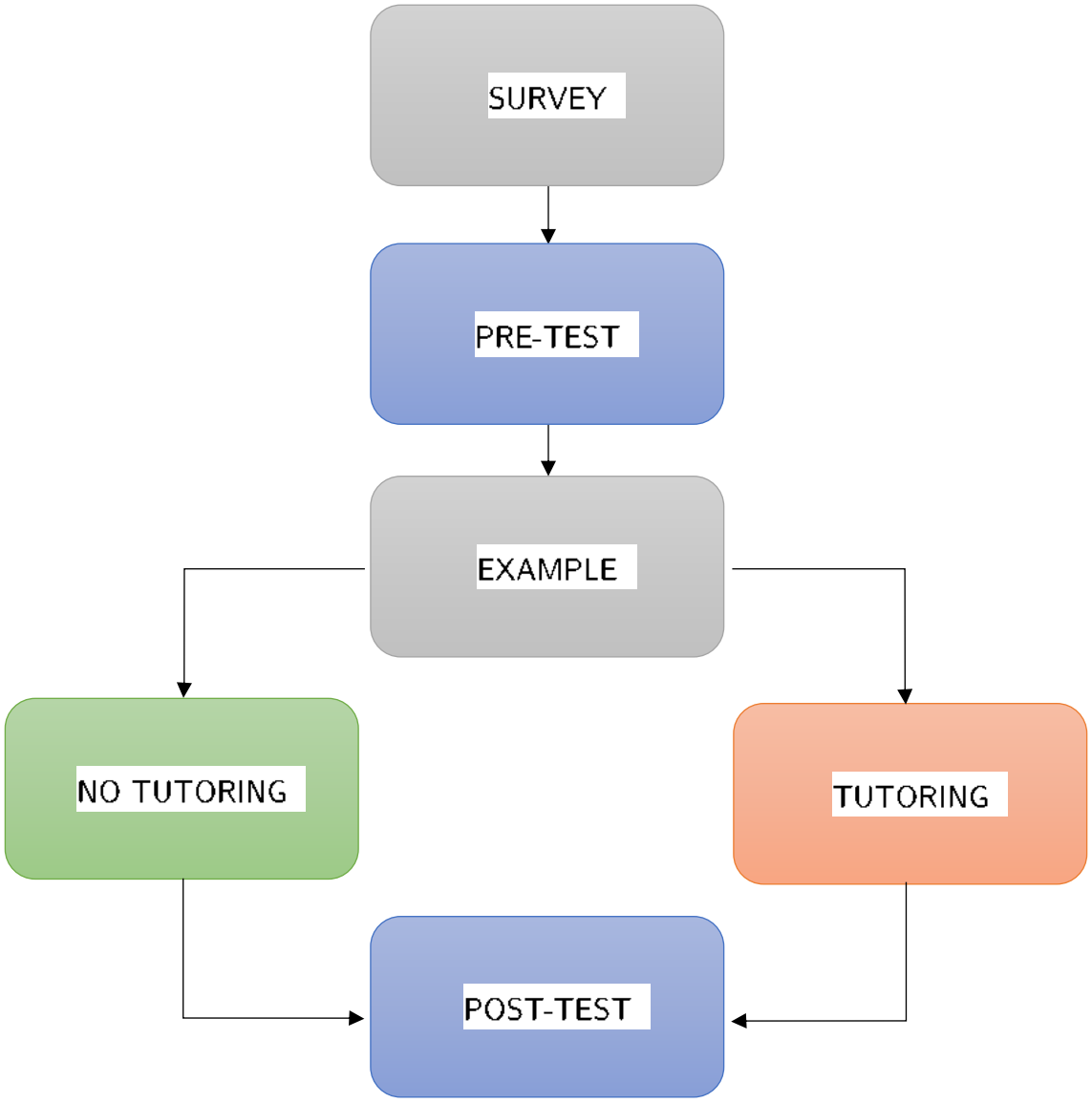
ASSISTments allows researchers to design their own experiments by writing problem sets, divided into sections. Usually, RCTs in the social sciences are composed of three main sections:

- **Pre-Test.** Subjects' initial features are recorded and used to weight the final results with the prior situation.

- **Condition.** Treatment group undergoes the condition, while the control group experiences business-as-usual.
- **Post-Test.** Subjects' final features are recorded: the results of the two groups are compared to assess the impact of the condition.

In order to create the experiment, a setting had to be chosen: the target population of choice consists in sixth-grade students, while the topic of interest regards the ability of writing expressions with variables, which is a component of the standard math curriculum in America. A problem set was built using ASSISTments pre-existing certified content. The problem set is accessible at [17]. The problem set is made up by 5 sections and has the following structure.

Figure 3.1 Flowchart of the Experiment



- **Survey.** Students are asked off-topic questions to gather information about the setting where the experiment is taking place and their general attitude towards ASSISTments. Questions are multiple choice.

    1. *Whose device is this?*
    2. *Have you got any similar device at home?*
    3. *Have your grades improved since you started using ASSISTments?*
    4. *Do you think ASSISTments or other digital systems help you learn faster than traditional methods?*

- **Pre-Test.** In the first section containing math content, students' prior knowledge is assessed through questions about a much easier mathematical topic, related to the number system. Students do not get any correctness feedback on these questions; therefore, they do not know how they performed.

- **Example.** Before the actual experiment, every student has the chance to revise or learn about the strategy that is needed to solve the type of problems that they are about to encounter.

- **Condition.** Students are randomly assigned to one of the following sections and are not able to access the other part, neither to choose which one to complete. The questions of the two sections are identical and are related to the ability of writing expressions with variables. The difference is in the type of feedback and the accessibility of tutoring strategies.

    1. **Control.** Students do not have access to tutoring strategy. However, they have immediate correctness feedback. They need to answer correctly to each of the problem in order to complete the section. They have access to the correct answer, but whenever they access it, their answer is marked as wrong.

    2. **Treatment.** From the beginning of each problem, students have access to tutoring strategies (i.e. hints or scaffolding tutoring). If they access help, their answer is marked as wrong. If they make a first wrong attempt, tutoring strategies are automatically turned on. Students then need to follow the help strategy until the correct answer is achieved.

- **Post-Test.** In the final section, students do not receive any type of correctness feedback and they cannot access any type of tutoring strategy. The questions are still related to the topic of the condition part. This section is needed to evaluate students' improvement after homework completion.

Sixth-grade math teachers were asked whether they were willing to take part in this research projects. The experiment could be conducted either at home or a school, but always as an individual activity. Within the established deadline, 6 teachers in many different states of the US had agreed to participate in the study.

3.3 Dataset Description and Preparation

The data about the problem set was collected by the ASSISTments Team of WPI and it was composed of two different datasets: student-level and problem-level data. All of the following analyses have been implemented using R, R Studio, and open source R packages [18] [19].

Of all of the 116 students who started the problem set, only 78 of them completed at least one problem in the post-test section. The subjects with incomplete information or missing values were dropped. Assigning median or mean values to these observation may be useful for other types of data mining analysis, but since the purpose of this experiment is to test the effectiveness of a methodology, estimating (and guessing) variables' values will bias the results. Unfortunately, the class sizes after cleaning is highly variable and unbalanced. The number of students in the *treatment group* is 35 (44,87 %), while the students who did not have access to tutoring strategies are 43.

The student-level dataset contains a row for each student and offers information about the three different sections: pre-test, condition, and post-test. In addition to problem set-related variables, measures of students' prior history are available, regarding practice level, percentage of correctness, probability of help requests, etc. Moreover, there are survey answers.

With respect to the class-level, the datasets only contain the identification numbers of teachers and classes. More class variables have been computed aggregating (mean values) student-level variables, such as average prior measures and modal survey answers.

Binding all these variables, a new dataset was created which includes two hierarchical levels: student-level and class-level.

Table 3.1 - Classes Composition

Class ID	Size	Control	Treatment
19788	14	10	4
27207	25	12	13
38527	21	10	11
67653	9	5	4
108312	8	5	3
113278	1	1	0
Total	**78**	**43**	**35**

Student Level Variables

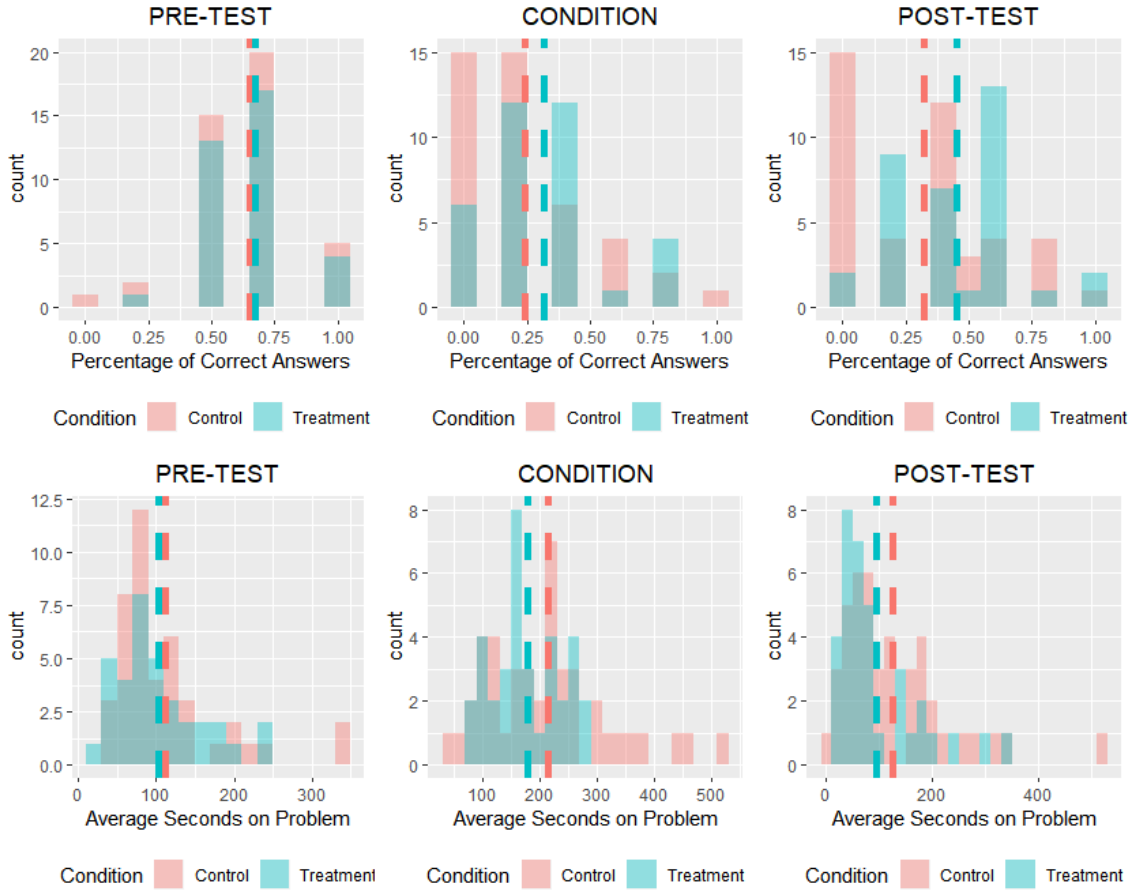
For each of the sections, the percentage of correct answers is available (only first actions), as well as the average time spent on a problem (seconds). For the treatment group, there are also the number of requested hints: however, in this problem set, if students answer wrongly or access help at first attempts, then a scaffolding strategy system opens and leads students to the solution. Therefore, these additional variables are not of interest in this example. ASSISTments also provides information about students' prior knowledge, number of problems completed, average speed, average number of attempts, and average hint count.

During the problem set, students' average performance went from quite high in the pre-test section, to particularly low in the condition part, ending with a general improvement in the final part: although this is true on average, the variability of the correctness percentage variable steadily increases during the problem set. Both before and after the experiment, the treatment group performs better in terms of percentage of correct answers. This slightly unbalanced division between treatment and control groups may represent a problem in the following analyses, especially given the restricted number of observations. The positive gap between

the two groups may be an effect of a pre-existing knowledge gap, instead of a consequence of the experiment. Therefore, students' prior knowledge must undoubtedly be considered. However, from the condition to the post-test section, the gap widens, which is a good signal for our hypothesis.

The average time spent on a single problem reaches a peak in the condition section: surprisingly, the control group spends more time than the treatment group solving the problem, despite the "Show Answer" button. Again, treatment group's students seem to be faster, which could be another indicator of their higher preparation. Interestingly, "treatment" students are faster than the "control" students in the post-test section.

Figure 3.2 Evolution of Performance and Response Time by Condition



Students' pre-existing metrics are aligned with the pre-test results. However, the pre-test section only yields a discrete distribution of students' knowledge, while the prior cumulative average performance on ASSISTments is a much more precise metric of student's preparation. This variable's probability distribution is close to a Gaussian one. All of the students but one had used ASSISTments

before, although their practice level is highly variable. The majority of the students is not used to requesting hints.

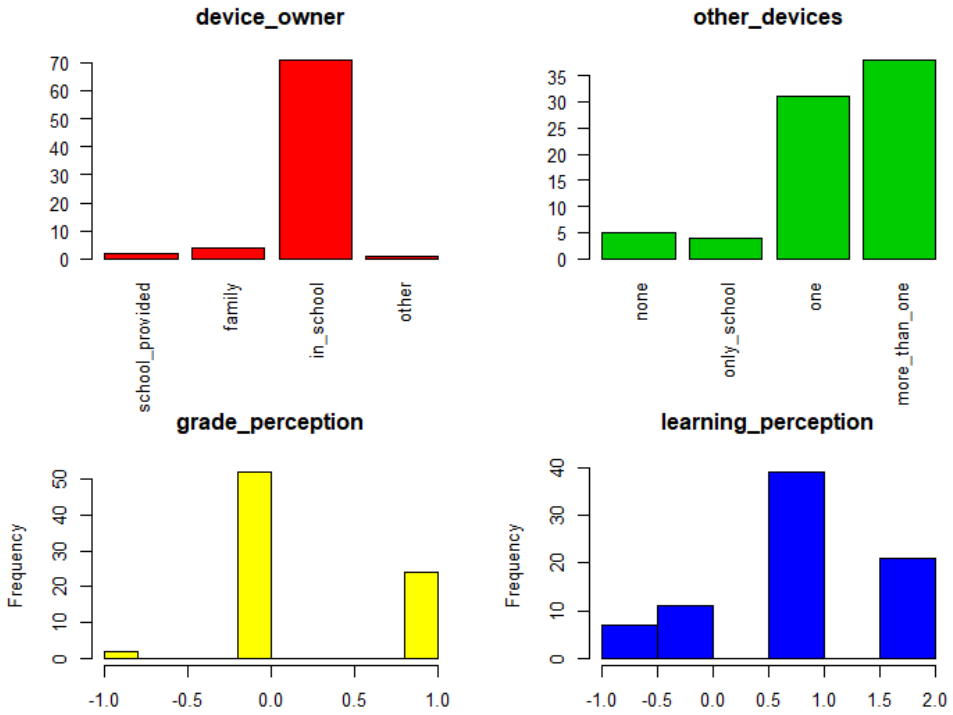
Table 3.2 - Student Level Variables

Variables	Mean	Median	S.D.	Treatment	Control
pretest_percent_correct	0.660	0.750	0.197	0.671	0.651
pretest_avg_seconds	107.691	87.409	63.558	102.356	112.034
condition_percent_correct	0.274	0.200	0.248	0.314	0.242
condition_avg_seconds	197.597	183.213	92.166	178.210	213.377
posttest_percent_correct	0.377	0.400	0.274	0.449	0.319
posttest_avg_seconds	111.737	72.985	93.689	94.586	125.697
prior_percent_correct	0.613	0.614	0.155	0.640	0.591
prior_avg_seconds	100.183	99.572	41.705	100.565	99.872
prior_problems_completed	93.192	54.000	81.671	109.343	80.047
prior_avg_hint_count	0.117	0.000	0.212	0.113	0.12

The survey answers also provide some major information about the setting in which the experiment was conducted and some students' features that were not directly accessible. Almost the totality of the students completed the problem set in the school environment, but individually. Only four of them stated that they were using a personal/family device. Given this extreme concentration, it is chosen to use this variable at the class level.

The vast majority of the pupils has access to similar technological devices at home (i.e. PC, tablets, smartphones). Technological acquaintance can be interpreted in two different ways: first, it can be used to roughly estimate students' economic situation; secondly, their familiarity or the lack of it may affect their experience on online platforms such as ASSISTments, giving them advantage or disadvantage over their classmates. Therefore, it looks like a fundamental factor to control for students' heterogeneity (student level).

Figure 3.3 Survey Questions



Another phenomenon that was assessed through the survey question is students' perception of ASSISTments as a learning tool. When questioned about the actual improvement in terms of grade, two thirds of them could not answer. Nevertheless, the remaining students gave a positive feedback, except two pupils who affirmed that their grades had been worsening.

On the other hand, answers about the comparison between traditional paper-based methods and technological learning tools have been much more variable: the majority says that digitalization helps them learn faster, but some others say that they prefer traditional methods. Grade and learning perception are used to describe individual students.

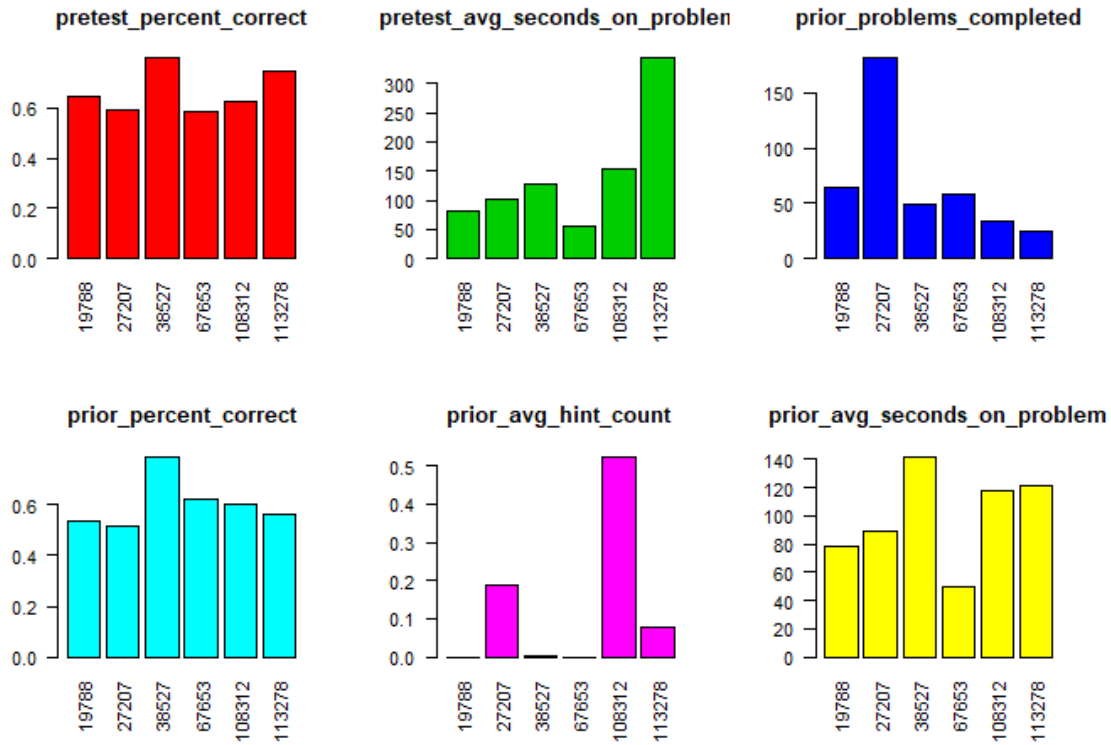
Class Level

The experiment setting in all of the classes is the school environment, therefore, this variable is constant for all of the students and can be discarded.

As a consequence, there are three components that describe classes: average prior level (aggregation of prior and pre-test variables); average technological familiarity/economic status; teacher or class dummy variables (which represent all of the other factors that affects the response variable at the class level). Classes differs

in terms of prior average scores, as well as completion time and practice level. Therefore, it is important to consider those discrepancies among students.

Figure 3.4 Class Average Variables



3.4 Exploratory Analysis

The hypothesis that needs to be tested is whether tutoring strategies used by the treatment group's students have a positive impact on their learning process and final performance. In order to assess the impact of the condition on these phenomena, different response variables are chosen:

a) Students' progress can be assessed by looking at the post-test score and comparing it with the condition score. The goal is to estimate whether the condition has a significant effect on improvement.

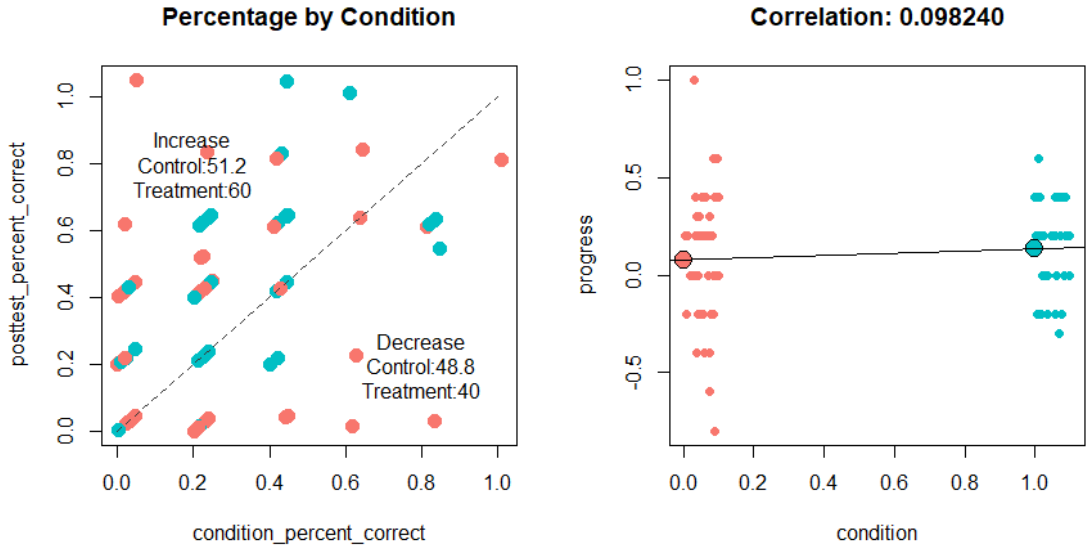
b) Students' final performance and knowledge level can be estimated using post-test section score. Given a constant level of prior knowledge, students in the tutoring condition should outperform the rest of the students.

In these first analyses, the hierarchical structure of the dataset will be ignored, and only student-level features will be considered.

Progress

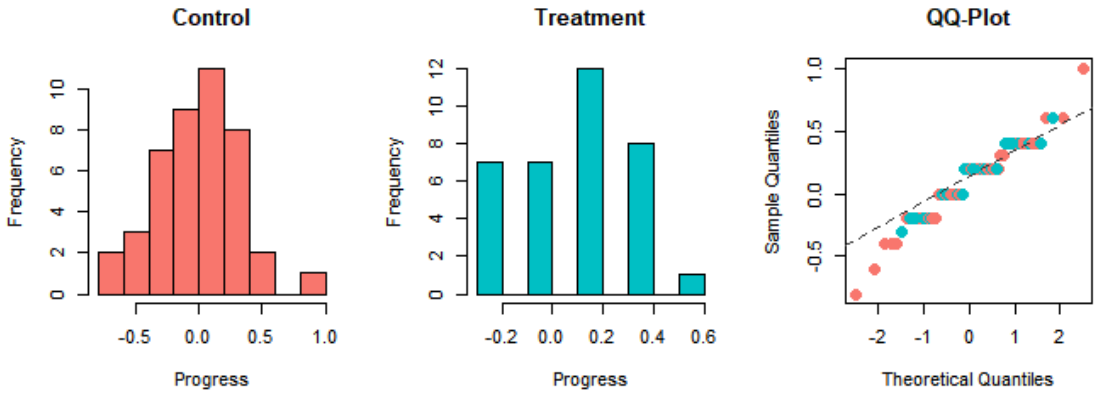
The percentage of students whose score improves is higher in the treatment group than in the control group, in which, however, the majority of the students improved despite the lack of tutoring. The difference is computed so as to have a measure for each student, which will be used as the response variable in the following analyses.

Figure 3.5 Scatterplots of Condition Score and Post Score, and of Progress and Condition



The new variable will be called "progress". Its range goes from -1 to 1: the worst result can only be achieved by students who have a perfect score in the condition section and no correct answer in the post-section; conversely, a progress of 1 is only achievable by students who scored 0 in the condition section. Therefore, the progress' feasible range depends linearly on the condition score. The response variable has a bell-shaped distribution, with mean 0.10 and standard deviation of 0.29: the QQ-Plot suggests that progress has a normal-like distribution.

Figure 3.6 Histograms of Progress by Condition and QQ-Plot



The average progress is higher in the treatment group than in the control group. This provides evidence for the hypothesis that tutoring strategies help students learn. However, the magnitude of this effect must be estimated so as to determine whether it is significant. The linear correlation coefficient between the condition and the progress is particularly low: such a value does not allow to draw conclusions about the relationship, whose positivity may or may not be due to chance. However, variances are significantly different between the two groups: therefore, the treatment at least causes heteroskedasticity in the progress variable (Bartlett's test [20]).

In order to assess whether the progress feature is significantly different in the two groups, an analysis of variance is performed. Inferential statistics allows to generalize conclusions retrieved from a sample to the overall population, thanks to probability confidence intervals and hypothesis testing. In the ANOVA test, two or more samples are compared in order to test whether their differences are due either to chance or to the grouping factors. The null hypothesis is that their means are equal, while the alternative is that at least a couple of means are significantly different. In this case, there is only one factor of interest that distinguishes the two groups, which is the presence of absence of tutoring strategy access. Classic ANOVA has assumptions that must be met in order to reach robust results: the response variable in this example is normally distributed but it does not respect homoskedasticity between groups. Therefore, instead of the classic method, the Welch's t test is chosen and performed [21].

In the Welch's unequal variances test, the following statistic is computed (see appendix 3.9):

$$\boldsymbol{t = \frac{\overline{X_A} - \overline{X_B}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \sim T_\nu}$$

This statistic has Student-t distribution with ν degrees of freedom under the null hypothesis ($h_0 : \overline{X_A} = \overline{X_B}$). In this case, the alternative hypothesis is that the treatment group has a higher mean value than the control group and therefore, the test is one-tailed ($h_1 : \overline{X_A} < \overline{X_B}$ where A is the control group). The Welch's statistic

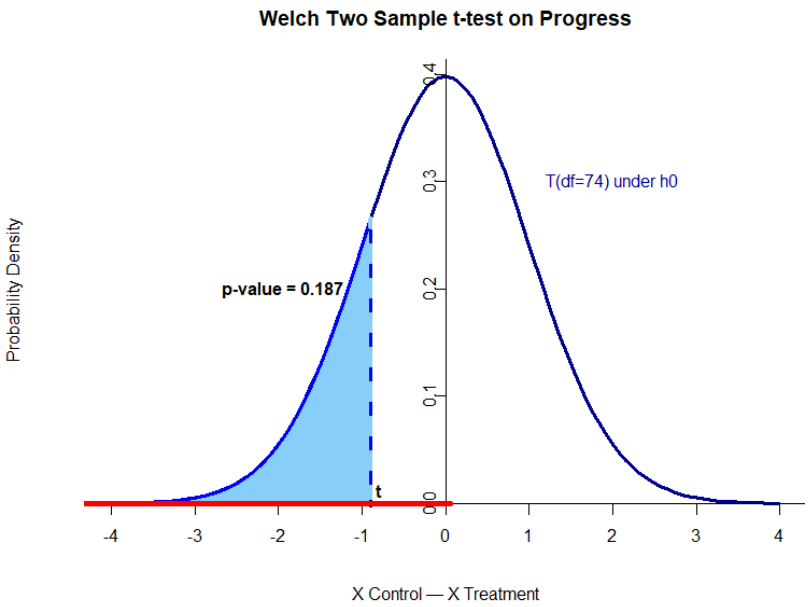
for the progress variable using the condition as the only factor yields the following results:

Table 3.3

Two Sample t-test

$\bar{X}_{CONTROL}$	0.0767
$\bar{X}_{TREATMENT}$	0.1343
ν	74.121
t	-0.8931
C.I (0.95)	[-∞; 0.050]
p-value	0.1874

Figure 3.7 Student-T Distribution with 74 degrees of freedom



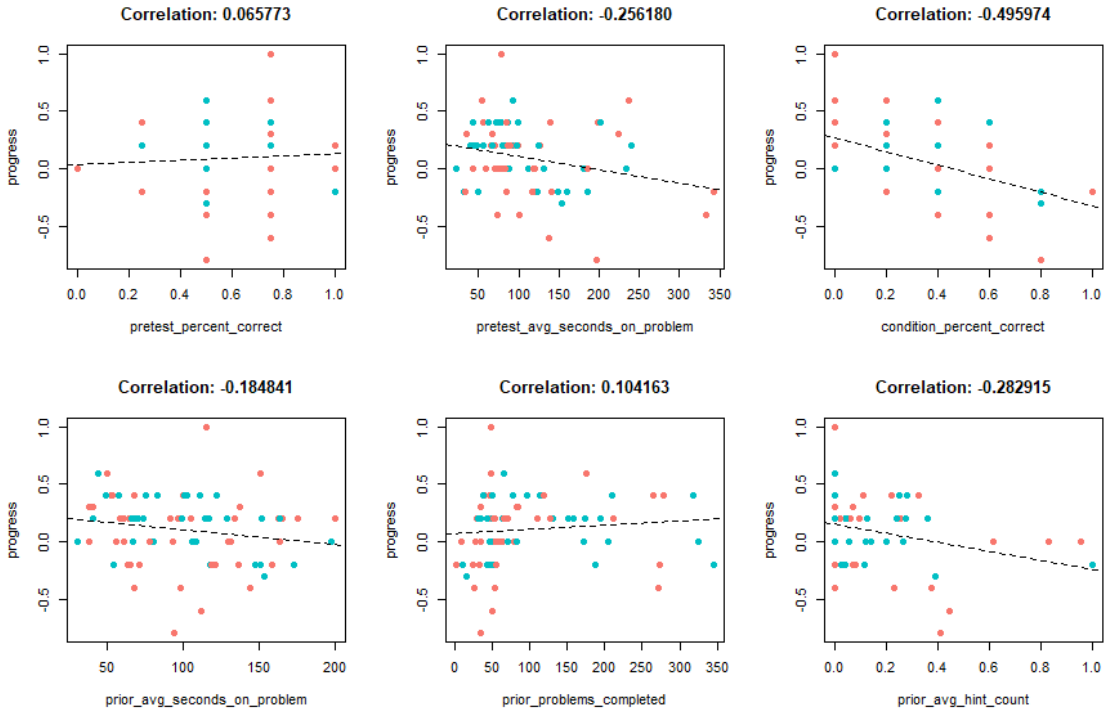
The t-test accepts the null hypothesis for which the average progress values in the two groups are not statistically different: although in this sample the alternative hypothesis holds, this cannot be generalized to the whole population.

This test only considers one factor and ignores other features that might affect the progress value for each student. It is relevant to identify the other factors that are correlated to the response variable.

At the student-level, the correlation coefficients are all very low. There is almost no linear correlation between progress and prior correctness percentage or pre-test section's score. Surprisingly, there is a much higher correlation with response time variables: the negative correlation suggests that those students who answer faster also seem to learn faster. The practice level on ASSISTments is slightly positively correlated to the outcome. In conclusion, the frequency of help requests shows a negative relationship with students' improvement: the more students are used to requesting help, the lower the probability of a high level of progress.

With respect to the survey answers' features, progress does not seem to be highly correlated to none of them. An ANOVA on each of the survey questions do not show significant differences in terms of progress.

Figure 3.8 Scatterplots of Progress and Student-Level Variable with Linear Regression Lines



Final Score

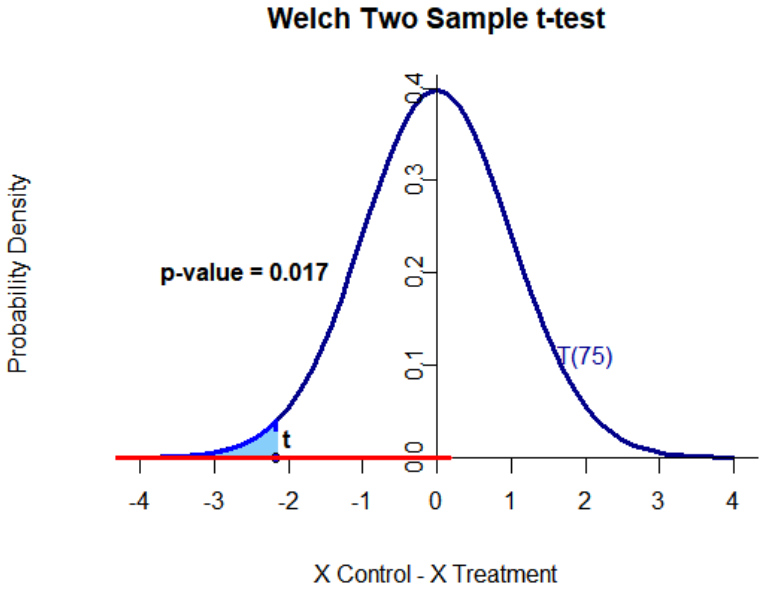
As the histograms show, the final score (post-test section correctness percentage) is on average higher for the treatment group. The Welch test finds a significant difference between the average final scores of the two groups (p-value: 0.017): however, this might be due to the fact that the treatment group has a slightly higher prior knowledge level.

Table 3.4

Two Sample t-test

$\bar{X}_{CONTROL}$	0.3186
$\bar{X}_{TREATMENT}$	0.4486
ν	75.871
t	-2.1644
C.I (0.95)	[-∞; -0.03]
p-value	0.0168

Figure 3.9 Student-T Distribution with 75.8 degrees of freedom



In fact, the final score is highly correlated to all of the previous correctness percentage metrics. As for progress, the correlation with speed variables is also low: this could be explained by the fact that the experiment was not time limited. Final performance is also negatively correlated to the probability of help requests.

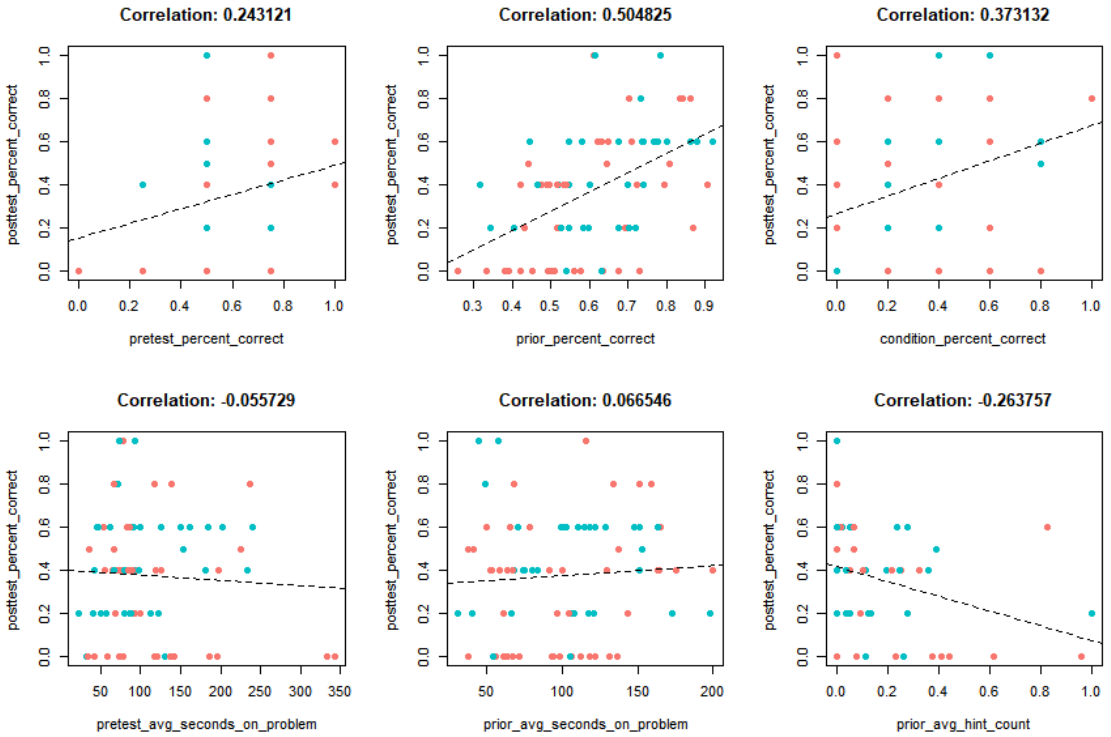


Figure 3.10 Scatterplots of Final Score and Student-Level Variable with Regression Lines

With respect to the survey questions, the ANOVAs indicates that there are no significant differences in terms of final score among students who answers differently to the survey questions. Even the technological familiarity variable yields a p-value of almost 50 %. Therefore, the survey questions do not appear to be significant explanatory variables for the post-test performance.

Interestingly, the dummy variable that shows the highest correlation coefficient is the one indicating whether or not student completed the assignment at home (which is actually a class-level variable), whose value is -0.24 the negativity of the coefficient indicates that students who completed that study at school have on average higher scores: although students were asked to complete the assignment individually, collaboration between students may explain this phenomenon.

3.5 Linear Regression Models

By design, the progress feature is highly correlated with the variables that have been used to create it. It may be more interesting to attempt to explain the final post-test score on the basis of the condition and the control/treatment section score, so as to estimate the impact of the initial score on the final one and whether or not the condition affects the magnitude of this effect. By considering the condition and the prior knowledge level, the pre-existing gap between treatment and control groups is levelled out.

In order to explain the distribution of the post-test section score, a simple linear regression model is used. Firstly, only the condition factor will be included (i.e. same results of a classic ANOVA); then, other features are used as predictors as well. Since variables are expressed in different metrics, standardization is applied on all of the variables. After the standardization, the intercept will be the final score of an average student (0 in all of the variables).

$$\boldsymbol{Final\ Score_i = \beta_0 + \beta_1 \cdot Condition_i + \ \ldots \beta_j \cdot Predictor_{ij} \ \ldots + \epsilon_i}$$

In Model 1, since the condition is a binary variable, β_0 is the mean score of the control group, while $\beta_0 + \beta_1$ is the average of the treatment group. In this case, the test is conducted on the significance of β_1 (h_0: $\beta_1 = 0$): the coefficient is 0.13, which is the difference between the two groups, and the p-value of the t test is 0.0366. With a confidence level of 95 %, the condition has a significant effect on the final score.

More features are added so as to include all of the relevant effects: since the purpose is not prediction but hypothesis testing, the effects that are assumed to be relevant are included, such as students' prior performance level, condition score, average response time, practice level, and probability of requesting help.

Table 3.5 - Linear Regression Models on Final Score

Predictors (X)	1	2	3	4	5	Best
Intercept	0.319 ***	0.226 ***	0.123	-0.176	-0.402 **	-0.365 **
Condition (C)	0.130 **	0.102 *	0.122 **	0.076	0.748 **	0.710 **
Condition_Percent_Correct		0.383 ***			-0.108	
Pretest_Percent_Correct			0.3203 **			
Prior_Percent_Correct				0.925 ***	1.115 ***	1.025 ***
Pretest_Avg_Seconds			-0.0001			
Prior_Avg_Seconds				-0.001	7.2e-05	
Prior_Avg_Hint _Count				-0.095		
Prior_Problems_Completed				0.0003	0.001 **	0.001 **
C · Condition_Percent_Correct					0.613 **	0.505 ***
C · Prior_Percent_Correct					-0.732 *	-0.641
C · Prior_Avg_Seconds					-0.002 *	-0.002 **
C · Prior_Problems_Completed					-0.002**	-0.002 **
RSE	0.2683	0.2528	0.264	0.2361	0.2226	0.2202
Adjusted R^2	0.0438	0.1512	0.07435	0.2599	0.3418	0.3564
F value (significance)	4.53 **	7.86 ***	3.06 **	6.41***	5.44 ***	7.09 ***
From 0.1 to 0.05: *; from 0.5 to 0.01: **; beyond 0.01 ***						

Except for model 4, the condition coefficient is always positive and significant at least at a 90 % confidence level. The variable that best describes the variability of the final score is prior percentage of correct answers: when it is included, the coefficient of the condition is not significant. However, there are other factors that should be taken into account. The last model includes a variety of features and also allows for their effects to change based on the condition: all of the interaction terms are significant. It explains only 34 % of the final score variability.

The prior correctness percentage is positive and have a strong effect on final performance. However, the interaction term suggests that the influence of prior knowledge level is stronger when students do not have access to help; then tutoring strategies are available, the effect is less than half of the original one. On the other hand, condition score shows a non-significant negative coefficient, which becomes positive for the treatment group: students in the treatment group can see explanations and hints so that their understanding of the skills, and therefore, their score is likely to increase. Practice level positively influences the final score for the control group, while the effect is cancelled for the treatment group. In conclusion, the speed feature shows a negative coefficient for the treatment group: it seems that fast students are more likely to benefit from tutoring strategies than pupils with longer response times.

Best subset selection allows to find the best models out of all the possible combinations of a set of features. With a maximum number of 10 features, the best model in terms of Mallow's Cp criterion contains the same variables as model 5 except for the condition score and prior average response time. The adjusted R squared for this model is 35.6 %. The coefficient of the condition is 0.71 and its p-value is 0.016. The model's overall significance is assessed through an F-test whose p-value is 2.152e-06. The signs of the coefficients are the same of Model 5.

With respect to the survey questions, the only features that are slightly correlated to the final score are about the environment where the experiment took place. However, school environment yields a non-significant coefficient: yet, the condition coefficient is significant.

3.6 Multilevel Analysis

The statistical concepts and formulas of the following sections refer to [22].

Table 3.6 - Classes' Means

Class	Treatment	Control
19788	0.5	0.28
27207	0.338	0.308
38527	0.545	0.56
67653	0.65	0.2
108312	0.233	0.12

In the linear models, it was assumed that subjects were all independent from each other. However, this assumption is incorrect because students are nested into classes. There is one class that contains a single student: this subject is excluded in the following analyses. In four out of five classes, the average final score is higher for the treatment group than for the control one; however, the impact as well as the average final score may depend on differences at the class-level.

Since the data have a nesting structure, a hierarchical model best suits the need to consider external factors that may vary the effect at the group/class level. In such a structure, there are two populations and two samples: classes in the dataset are just a sample of all of the sixth-grade classes in the US; students are a sample of all of the students in the classes that make up the first sample.

As a consequence, micro-level features' variability is affected by two random effects due to the sampling structure of the experiment. However, if the class level does not significantly influence the micro-level features, the hierarchical structure of the data can be neglected. In order to assess whether students in the same class have a higher correlation than students from two different classes, a random effects ANOVA model (i.e. null model) is computed.

$$\boldsymbol{Final\ Score_{ij} = \overline{Final\ Score} + U_j + R_{ij}}$$

U_j is not a parameter but an additional error term that changes for each class. The average final score added to U_j gives the average response value for each of the groups. Its variance is the so-called between variance (τ^2) and it is computed

as the variance from the overall mean of the group averages. When the group sizes are not equal, the between variance must be adjusted [see appendix 3.9].

R_{ij} is the student-level error term that has mean 0 and it is assumed to have a normal distribution. Its variance is the within variance (σ^2), which is the weighted mean of each group's variance. The total variance is just the sum of between and within sum of squares, divided by the total number of subjects minus 1.

The intraclass correlation coefficient simply measures the correlation between two subjects from the same group: it is computed as the ratio between the between and total variances. If this coefficient is significantly large, then the hierarchical structure should not be neglected because it influences the distribution of the response variable. The intraclass correlation coefficient is [see appendix 3.9]:

$$\boldsymbol{ICC = \frac{\tau^2}{\tau^2 + \sigma^2} \simeq \widehat{ICC} = \frac{\widetilde{B}}{\widetilde{B} + \widehat{W}}}$$

In this case study, its value is 0.186, which is relatively low but not unusual in educational research. In order to prove the significance of the group variance, the classic F-test of ANOVA is performed, whose statistic distributes as a Fisher-Snedecor under the null hypothesis.

$$\boldsymbol{F = \frac{\widehat{B} \cdot \widetilde{n}}{\widehat{W}} \sim F_{K-1;\, N-K}}$$

The F-value is 4.365 and significant at a 99 % confidence level. This test proves that the multilevel structure of the dataset should not be neglected because group membership significantly affects the final score.

In the null model, there are only the overall average response and the two levels' error terms. The class-level random effect is also called random intercept because the intercept of the model becomes group-dependent. The fixed intercept should be interpreted as the expected value of the final score for a random student in a random class: the reason why it is not identical to the overall raw mean is because it is weighted for the various classes. The parameters are estimated using the restricted maximum likelihood estimation method.

Table 3.7 - Empty Model

Fixed Part	**Estimate**	**p-value**
Intercept	-0.0656	0.788
Random Part	**Name**	**Var.**
Class-Level	Intercept (τ^2)	**0.194**
Student-Level	Residuals (σ^2)	**0.850**

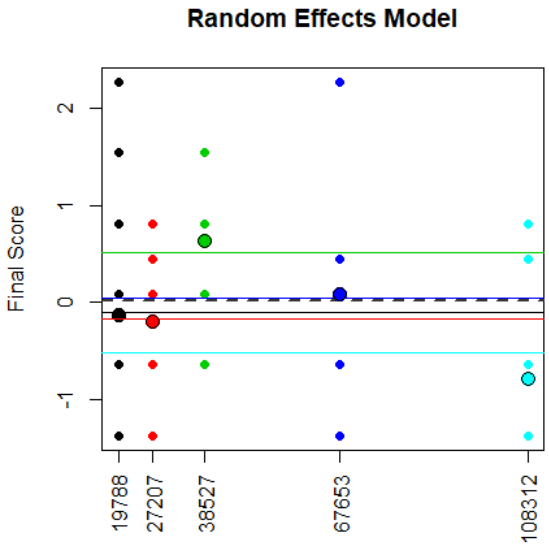


Figure 3.11 Class Intercept and Class Mean Final Scores

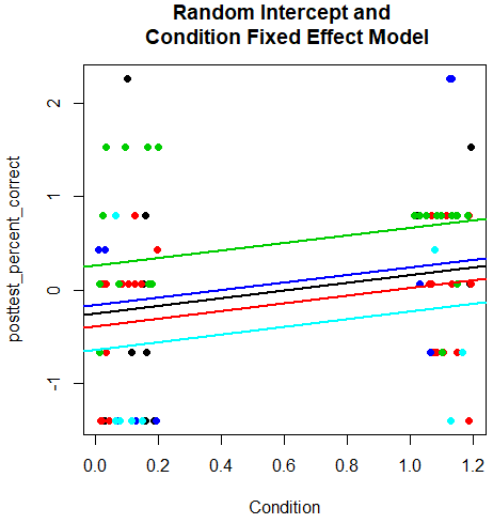
As it can be seen the intercept is not significant: because of the previous standardization, the response variable has mean 0. The class-level variance is about a fifth of the total variance. The covariance between two students in the same class is about 20 %. τ^2 and σ^2 are two terms of unexplained variability that respectively acts at the group-level and at the subject-level. These values can be reduced by the introduction of explanatory variables at both levels. Since the effect of interest regards the condition, a fixed effect for this variable is initially added to the model.

$$\boldsymbol{Final\ Score_{ij} = \beta_{0j} + \beta_1 \cdot Condition_{ij} + R_{ij}}$$
$$\boldsymbol{= \gamma_{00} + U_{0j} + \gamma_{10} \cdot Condition_{ij} + R_{ij}}$$

Table 3.8 Random Intercept Model

Fixed Part	**Estimate**	**p-value**
Intercept	-0.2405	0.3538
Condition	0.4047	0.0584 *
Random Part	**Name**	**Var.**
Class-Level	Intercept (τ^2)	0.1771
Student-Level	Residuals (σ^2)	0.8227

Figure 3.12 Class Random Intercepts



The two-levels variances have both decreased but remain high. The coefficient of the condition is positive. To test the significance of fixed terms (h_0: $\beta=0$), a t-test is performed: the t-value has a Student-t distribution. The condition coefficient is significant at a 94 % confidence level. Since this is a fixed effect, the impact of the condition is the same on each group, whose intercepts, however, are free to move. There are three schools at the center of the distribution and two schools which have very high and very low intercept coefficients.

In order to compare models with different numbers of predictors, a homogeneous measure of goodness must be employed. The most common measure is the deviance, which can be estimated for every model that uses the maximum likelihood method. It is computed as minus twice the natural logarithm of the likelihood. It measures the lack of fit between the model and the data: the model with the lowest deviance represents the best fit.

$$\boldsymbol{D_M = -2 \log (\mathcal{L})}$$

A deviance test can be performed to prove the significance of the random intercept in the model. In order to implement the test, the deviance of the model (RI) must be compared to the same model with no random intercept (i.e. OLS). The larger model will always have a smaller deviance: hence, the null hypothesis is that the difference between the two deviances, which is always positive, is however not significantly different from 0, meaning that the larger model does not represent an improvement. On the other hand, if the difference is sufficiently large, the larger one should be considered a more feasible model. The difference of the deviance has a chi-squared distribution with a number of degrees of freedom equal to the number of added parameters.

$$\boldsymbol{D_0 - D_1 \sim \chi^2_{p_1 - p_0}}$$

In our case, there is only one additional parameter (τ^2) and the deviance difference is equal to 6.234, for which the p-value is 0.012. Therefore, the random intercept is significant in this model at a 95 % confidence level.

The impact of the condition may be stronger for classes with a low level of average final score, while it could be weaker for classes in which the overall mean is higher. In order to let the slope of the condition vary, a random effect for this variable must be included. By including a random slope in the model, also the variance of the response variable within a group is dependent on the explanatory variable.

The hypothesis is that the slope will be steeper for high-intercept classes than for low-intercept ones, which can be translated into a negative correlation coefficient between the two random effects.

$$\begin{aligned} \boldsymbol{Final\ Score_{ij}} &= \boldsymbol{\beta_{0j} + \beta_{1j} \cdot Condition_{ij} + R_{ij}} \\ &= \boldsymbol{\gamma_{00} + U_{0j} + (\gamma_{10} + U_{1j}) \cdot Condition_{ij} + R_{ij}} \end{aligned}$$

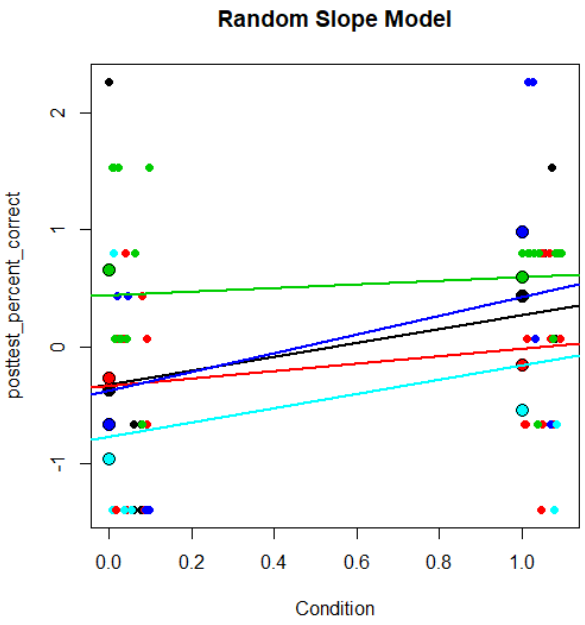
Table 3.9 - Random Slope Model

Fixed Part	**Estimate**	**Std. Error**	**T-Value**	**P-Value**
Intercept	-0.273	0.271	-1.006	0.373
Condition	0.493	0.276	1.784	0.175
Random Part	**Name**	**Variance**	**Std. Dev.**	**Correlation**
Class-Level	Intercept ($\boldsymbol{\tau_0^2}$)	0.2647	0.515	-0.64 ($\boldsymbol{\tau_{01}}$)
	Condition ($\boldsymbol{\tau_1^2}$)	0.1519	0.390	
Student-Level	Residuals ($\boldsymbol{\sigma^2}$)	0.7936	0.891	

The condition coefficient is still positive and remains positive for each of the groups. However, the fixed effect of the condition is not significant. The condition slope is $\gamma_{10} + U_{1j}$, which is a random variable with mean γ_{10} and standard deviation τ_1. In a 95 % confidence interval, the slope goes from -0.271 to 1.257: this means that the effect may also be null. The correlation between the intercept and the condition random effects is negative: this means that classes with a higher

performance for an average student have a lower within-class effect of the condition.

Figure 3.13 Class Regression Lines



According to the likelihood ratio test, the random slope for the condition is highly not significant. Yet, the whole model is significantly better than the OLS simplification. Given the results, the model does not need a random slope for the condition variable: it can be concluded that the condition effect is not significantly different among classes. The random intercept is instead significant, and its variance can be reduced by the introduction of class-level predictors. The irrelevance of adding a group-dependent slope to condition is also pointed out by the fact that adding other subject-level features and interaction terms with the condition completely dissolves τ_1^2 variance: the condition effect may change among students but the cause does not regard school differences but, in fact, student differences.

3.7 Hierarchical Models' Specification

As in linear models, the condition coefficient should be tested once that all of the other significant effects are included. Including student-level variables may reduce the residual variance ($\boldsymbol{\sigma^2}$), as well as the group-level variance. On the other hand, variables at the group-level can help reduce the variability of U_{0j}: adding group factors may help predict more accurately the group-dependent regression coefficient β_{0j}.

$$\boldsymbol{\beta_{0j} = \gamma_{00} + \gamma_{01} \cdot Class\,Variable_j + U_{0j}}$$

The relationship between the final score and student-level variables has already been analysed. With respect to the class-variables, the final score is positively correlated with the class average prior correctness percentage and negatively with

the class average number of help requests. Average hint count can prove more useful at the class level, indicating whether teachers use ASSISTments mainly to teach (high probability of asking for help) or to assess students (low probability of asking for help). In addition to these factors, adding the average correctness percentage in the condition section may control for the fact that students may or may not have already encountered the specific math skill being tested: it is assumed that the final score will be higher in classes in which the topic has already been covered (i.e. with an average good result in the condition section). With regard to speed variables, condition and prior speed show very similar correlation coefficients. Students' response time is assumed to affect the final score mainly at the subject-level rather than at the class one.

In conclusion, there is a slightly positive correlation with the dummy variables indicating whether or not students possess more than one other technological device at home. The aggregation of this factor at the class-level (mode value of the survey answers for each class) can be considered as an instrumental variable used to roughly assess the average socio-economic status of each class. The positive correlation supports the hypothesis that wealthier classes' students may score higher than students coming from a poorer background.

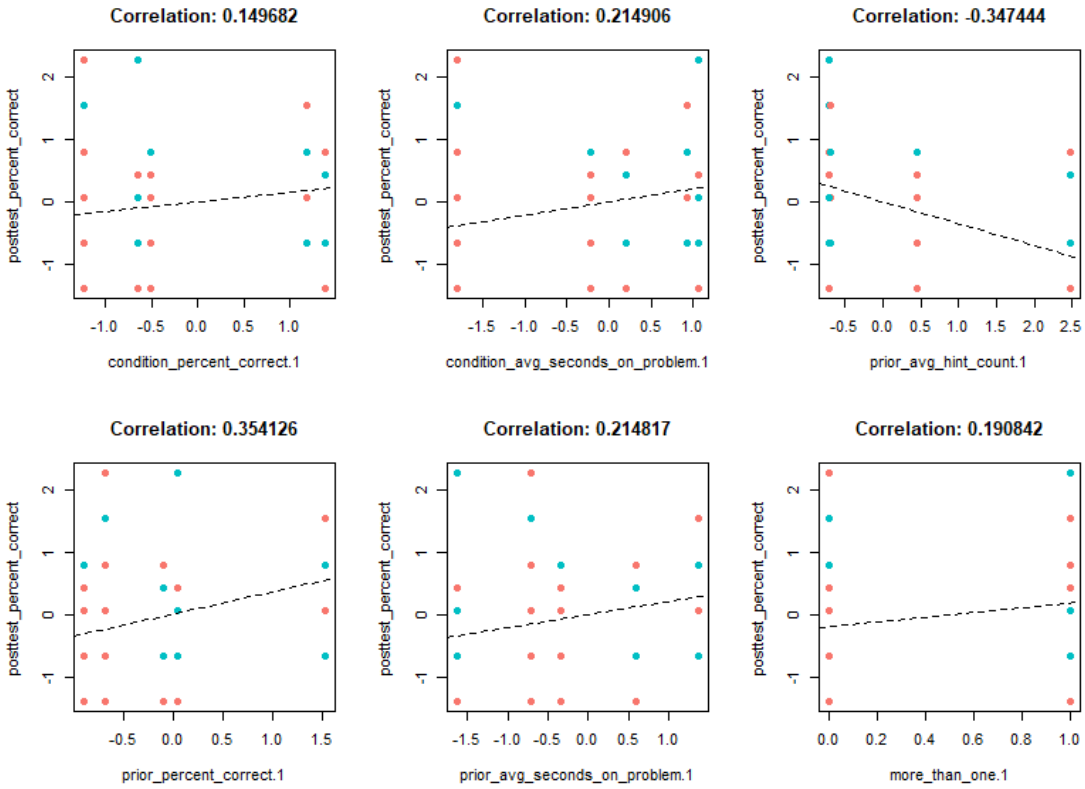


Figure 3.14 Scatterplot of Final Score and Class-Level Variable with Regression Lines

Table 3.10 - Hierarchical Linear Models

	0	1	2	3	Best
Fixed Part					
Intercept	-0.241	-0.131	0.28	0.137	0.375 *
Condition (C)	0.405 *	0.262	-0.382	0.271	0.266
Prior_Percent_Correct		0.55 ***		0.534 ***	0.37 ***
Prior_Problems_Completed		0.254		0.244	
C · Condition_Percent_Correct		0.480 ***		0.498 ***	0.471 ***
C · Prior_Percent_Correct		-0.389 *		-0.417 *	
C · Prior_Avg_Seconds		-0.372 **		-0.362 **	
C · Prior_Problems_Completed		-0.448 **		-0.456 **	
$\overline{\text{Condition\_Percent\_Correct}}$			0.765 ***	0.186	
$\overline{\text{More\_Than\_One}}$			-0.9 *	-0.451	-1.031 ***
$\overline{\text{Prior\_Avg\_Hint\_Count}}$			-0.535 ***	-0.284 **	
$\overline{\text{Prior\_Percent\_Correct}}$					0.618 ***
C · $\overline{\text{Condition\_Percent\_Correct}}$			-0.975 ***		
C · $\overline{\text{Prior\_Avg\_Hint\_Count}}$			0.183		
C · $\overline{\text{Prior\_Avg\_Seconds}}$					-0.529 ***
C · $\overline{\text{More\_Than\_One}}$			1.608 **		
Random Part					
Intercept (τ_0^2)	0.177 **	0.07	-	-	-
Residuals (σ^2)	0.823	0.613	0.689	0.606	0.617
Goodness of Fit Measures					
Deviance	211.04**	192.59	189.87	190.98	188.18
Adjusted R^2	-	-	0.231	0.394	0.383
F-Value	-	-	4.257 ***	5.936 ***	8.848 ***

In the first model, the best subset found in the linear models' analysis is implemented again with a random intercept. Starting from a group-level variance of 0.177 in the model with only the condition, Model 1 leads to a significant decrease in τ_0^2, as well as in σ^2. The likelihood ratio test states that class differences are not significantly relevant when these subject-level features are considered. The condition coefficient is still positive but not significant.

The multilevel analysis is also of interest in order to assess whether class variables and their interaction terms with the condition are significant.

In Model 2, the most relevant group-level effects are added and the random intercept variance completely vanishes. The condition coefficient is negative but not significant in Model 2, but its interaction terms are indeed significant. The condition percentage score coefficient is positive meaning that classes that have already sufficient knowledge on the skill are more likely to have students with high final scores. However, under treatment, the coefficient becomes negative: tutoring strategies seem to fill the gap between the two types of classes. The prior average hint count coefficient is significant and negative: students whose teachers are more likely to use ASSISTments as a testing tool may be more careful and less likely to request hints, hence more likely to have a higher final score; on the other hand, students who are used to accessing helps may be less careful. The effect of average hint requests is reduced under treatment. The dummy variable regarding the average familiarity with technological devices show a significant coefficient which is in contrast with the previously formulated hypothesis.

The random intercept variance is nearly 0 and, therefore, the included variables explain almost all of the between-group variability: the hierarchical structure can therefore be neglected, and normal linear regression is more appropriate. With regard to prior and condition percentage of correctness, neither of the variables show a significant between-group coefficient and, therefore, within-centering is not convenient.

In Model 3, all of the significant variables at the two levels are included. The signs of the coefficients are all equal to the previous models. The interaction terms are significant, and their signs are aligned with the hypotheses explained before.

Finally, best subset selection offers a model with 6 predictors with significant condition interaction terms and an explained variability of 38.3 %.

3.8 Results and Conclusions

Introducing and assessing the multilevel structure was necessary to secure that all of the significant levels of variability were considered. The results suggest that, although there are significant differences among classes regarding the problem set final score, these discrepancies are almost completely explained by subject-level variables. As a consequence, the hierarchical structure can be neglected in this particular case study.

As a consequence, the impact of the tutoring strategies was analysed through multiple linear regression models containing both micro and macro-level features. Micro-level interaction terms are also included. The non-significance of the random slope of the condition dummy variable suggests that its effect is not significantly different among classes, thus indicating the irrelevance of including cross-level interaction terms between the condition and class-level variables.

When the condition is considered as the only predictor for the final score (ANOVA), its coefficient is positive and significant. The positiveness of the coefficient indicates that students who had access to online tutoring strategies during homework do score higher on average in following tests than students who completed traditional homework in which tutoring strategies are not available.

The sign of the coefficient remains constant even when different predictors are included in the regression in order to control for external factors. However, in some of the models, the coefficient is not significant: although the impact is positive in this sample, its magnitude is not statistically different from 0 (i.e. its confidence interval contains 0) and, therefore, the presence and the positiveness of the effect cannot be extended to the overall population.

However, when the condition is not a significant explanatory variable, some its interaction terms always prove to have strong coefficients: the condition does not directly influence the final score, but it changes substantially the impact of other variables.

For example, tutoring strategies seem to reduce the importance of prior knowledge level on the final score: the negative coefficient of this interaction term weakens the positive correlation between prior percentage correctness and final score. Thus, tutoring strategies seem to partially fill the gap between low-achieving and high-achieving students.

Another interesting effect regards the one of the condition section score, whose coefficient is not significant for the control group; however, the predictor's coefficient becomes positive and significant if tutoring strategies are available. Surprisingly, students' performance during traditional homework does not prove to be a good predictor of their future performance; however, when homework includes tutoring, then the relationship becomes positive. This phenomenon could be explained by the fact that if a student answers wrongly when in the treatment condition, a scaffolding strategy containing explanations automatically opens, allowing students to understand their mistakes.

In general, the condition seems to weaken the influence of traditional performance predictors, such as previous score and practice level, thus partially levelling out students' individual differences in achievement.

In terms of class-level variables, three variables appear to influence the final score. Firstly, the class average score in the condition section can signal whether or not students have already encountered the particular math skill being assessed in their curriculum: contrary to the corresponding disaggregated variable, the class-level aggregated feature shows a significant positive effect on final correctness percentage.

Secondly, class average attitude towards the platform in terms of hint access can reveal whether or not the tool is mainly used for assistance or assessment: in the first case, students will be more likely to frequently access tutoring strategies (i.e. high values of class average hint count) and to answer more carelessly (lower final scores); otherwise, students should be used to evaluating carefully the problem before accessing help or giving an answer (low values of $\overline{\text{Prior\_Avg\_Hint\_Count}}$ and higher final scores).

Finally, high familiarity with technological variable seems to have a negative effect on students' performance in the problem set. In this case, evidence opposes the hypothesis that students living in a more digital home environment (possibly meaning wealthier) perform better than other students.

In conclusion, it can be said that tutoring strategies on ASSISTments have a positive impact on students' future performance. Students who completed traditional-like homework (control group) are likely to perform worse than students who had the chance of using hints and scaffolding strategies during homework.

The effect can be decomposed in mediated effects that influence other predictors, which in turn are the most relevant explanatory factors for final performance.

The model chosen as the optimal representation of the phenomenon is represented as Model 3 of Table 3.10, which explains almost 40 % of the response variability.

3.9 Appendix: Formulas

In the Welch's t-test, $\overline{X_A}$ and $\overline{X_B}$ are the sample means of the two groups; n_A and n_B are the sizes of the two samples; s_A and s_B are the sample standard deviations; ν is the number of degrees of freedom of the Student-t distribution. [21]

$$s_A = \sqrt{\frac{\sum_{i=1}^{n_A}\left(x_i - \overline{X_A}\right)^2}{n_A - 1}} \qquad \nu = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{s_A^4}{n_A^2(n_A-1)} + \frac{s_B^4}{n_B^2(n_B-1)}}$$

The formulas used for the multilevel analysis of paragraph 3.6 are taken from [22]. N is the overall number of observations; K is the number of groups; n_j is the number of observation in the j-group; $\bar{Y}_{.j}$ is the mean in the j-group; $\bar{n}$ is the average group size (N/K).

Within Variance
$$\widehat{W} = \frac{1}{N-K}\sum_{j=1}^{K}\sum_{i=1}^{n_j}\left(Y_{ij} - \bar{Y}_{.j}\right)^2$$

Between Variance
$$\widehat{B} = \frac{1}{\tilde{n}(K-1)}\sum_{j=1}^{K}\left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2 \cdot n_j$$

Adjustment for unequal sizes
$$\tilde{n} = \bar{n} - \frac{1}{N(K-1)}\sum_{j=1}^{K}(n_j - \bar{n})^2$$

Total Variance
$$Var(Y_{ij}) = \frac{N-K}{N-1}\widehat{W} + \frac{\tilde{n}(K-1)}{N-1}\widehat{B}$$

Expected Observed Within Variance
$$\varepsilon\left(\widehat{W}\right) = \sigma^2$$

Expected Observed Between Variance
$$\varepsilon\left(\widehat{B} - \frac{\widehat{W}}{\tilde{n}}\right) = \varepsilon\left(\widetilde{B}\right) = \tau^2$$

4 Analysis of Log Data: Groups Identification

At the school or class level, log-data files offer large amount of information about students that can be useful for teachers. However, analysing the data and the results for each of the students is impracticable. Therefore, some summarization and simplification of the data is needed so that they can become in fact readily understandable for teachers and administrators. This analysis implements two of the most common methods in educational data mining, which are called *clustering* and *distillation for human judgement*. [1]

Given a particular class or school, it can be useful for teachers and instructors to identify particular groups of students who have similar features at each point in time, so as to suggest different approaches towards each of these groups. Although each student is unique and has its own needs and shortages, rarely teachers have time to address them at an individual level. By dividing a classroom into homogeneous groups, teachers will be able to give different recommendations and suggestions to each of them, without losing time assessing students one at a time.

This group-based method represents the first step towards adaptive learning and teaching approaches which are likely to thrive in the years to come. This method is likely to have effects both in the short and long-term. For instance, identifying students at risk of failure in advance may prove incredibly important in preventing their actual failure at the end of the school year. Once teachers are aware of which students fall in this category, they can suggest remedial programmes, extra-practice, or whatever they think it may be useful. Not only this is likely to help the low-achieving students, but it will also improve the overall quality of the classroom. In the long run, the gap between low-achieving and high-achieving students is likely to narrow and students will be more likely to succeed in higher education.

Using log-data instead of only test or assignment scores allows to categorize students also in terms of variables such as motivation, confidence, and speed. If students limp in some of those areas, schools will be able to help them overcome their difficulties before it is too late. For instance, the school ability of identifying in time careless and demotivated students may reduce drop-out rates. In the long

term, the average educational attainment is likely to rise, as well as productivity [23].

Moreover, at higher levels of the education system, these data can be aggregated and compared so as to monitor more accurately the general composition of the student population in a given area. This can prove useful to address issues such as regional inequality and the quality of education in deprived areas.

Dividing units into homogenous groups is a widely popular application of data mining. These methods are called clustering algorithms: given a set of features, these algorithms are able to divide observations into groups or clusters based on a dissimilarity measure. [24]

However, the hierarchical structure of the educational system requires to consider the effects that can cause differences between groups at each level. As a consequence, generalization is often not possible. Nevertheless, similar patterns may emerge from different implementations which do not differ considerably.

In this application, a cluster analysis is implemented in different contexts with similar macro-level characteristics. Firstly, different models are created on the basis of a single school: the results are then interpreted, and their meaningfulness is assessed. Secondly, the same models are implemented on data from neighbour schools and from the same school but on data collected the following year: the aim is to compare the clusterizations and assess the goodness of the models.

The statistical software R has been employed for the following analyses: figures, tables, and plots have been generated using R Studio. [18] [19]

When statistical terminology and concepts are employed, they refer to the definitions and formulas given in [24].

4.1 Dataset Description & Data Preparation

For this application, a dataset provided by ASSISTments was chosen [15]. The dataset is public, and it includes 942,816 rows [25]. The dataset contains the log-actions of 1709 middle school students. Log-actions refer to two different school years and to four different schools: for the school year 2004-2005, the log-actions of all of the schools are available, while for the following year, the dataset contains only the actions of two schools. Students used the ASSISTments software twice a week for the whole school year to complete assignments on math skills. Schools

are located in the same city in central Massachusetts, while students are distilled from a diverse population, both racially and economically [26].

The dataset contains variables regarding the student, such as gender, school identification code, and final test score for the school year. The data are anonymized and do not contain personally identifiable information about the individuals. Each row represents an action made by the student while solving a specific problem: each one contains information about the type of problem, the number of attempts so far, help requests, response correctness and time.

The final score (Massachusetts Comprehensive Assessment Score) was not available for some of the students: the observations containing missing values for the final score were dropped. The variable "timeTaken" contains a high number of outliers which could damage the following analyses: the extreme of the upper whisker in the boxplot is just above 1 minute, while the maximum value is about 2 hours. Since some problems may take more time, the extreme upper whisker value appears too low, so the threshold of 5 minutes is chosen instead: all of the actions above this time are dropped.

Finally, some problems in ASSISTments do not assess any math skills since survey or research questions can be included in the problems. Therefore, the actions containing problems with "noskill" content are dropped. After the cleaning process, there are 765,092 observations and 1376 students left.

Figure 4.1 Boxplot of "timeTaken" Variable

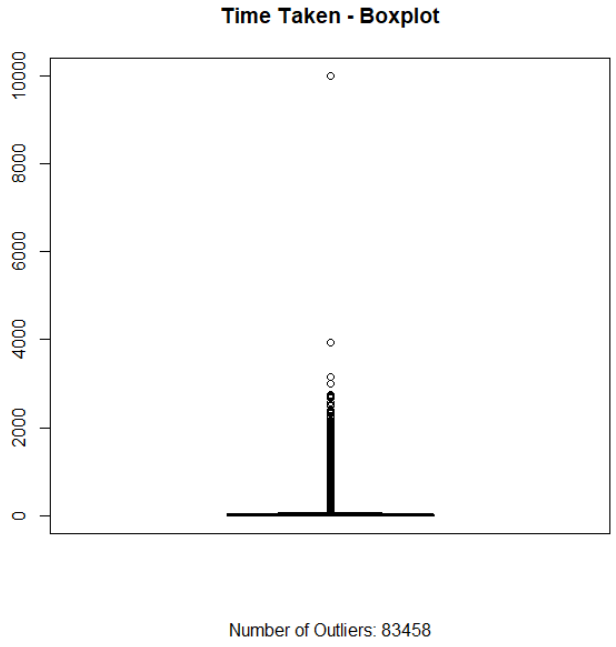


Table 4.1 - Dataset Summary

Middle School	School Year	Actions	Students
School 1	2004-2005	95,501	191
School 2	2004-2005	315,497	382
	2005-2006	105,226	295
School 3	2004-2005	20,595	125
School 4	2004-2005	191,279	238
	2005-2006	36,994	295
Total		765,092	1376

Since students did not complete the same problem sets during the school year, results for each student are likely to depend on the problems encountered. Therefore, a dataset containing details about each problem is created and the following measures are computed: total number of attempts; number of students who solved the problem; number of times the problem was encountered (i.e. number of first attempts to the problem); average number of attempts; average time of completion; average response time; percentage of correct answers; percentage of help requests; average response time in first attempts; percentage of correct answers in first attempts; percentage of help requests in first attempts. The number of unique problems containing math skills is 2094.

Then, a dataset containing a row for each problem completed by each student was created. The same measures were computed as well as the distance between the student's results and the average results for each problem: these new variables are assigned the "dev" prefix.

Since the aim of the analysis is to identify students with similar features, a "student dataset" was created. For each pupil, the mean of all of the variables was computed. In addition to these metrics, the number of different problems encountered was included as a measure of practice level.

4.2 Metrics Description and Selection

Before starting any analysis, the measures in the student dataset are explored. The dataset is divided in school years, so as to control whether there are significant differences between the two time periods. While evaluating the distributions and the correlations among features, possible differences between schools are also considered.

First, the features are divided into categories:

- Performance measures
- Time measures
- Attempt and Practice measures
- Help measures

Performance measures

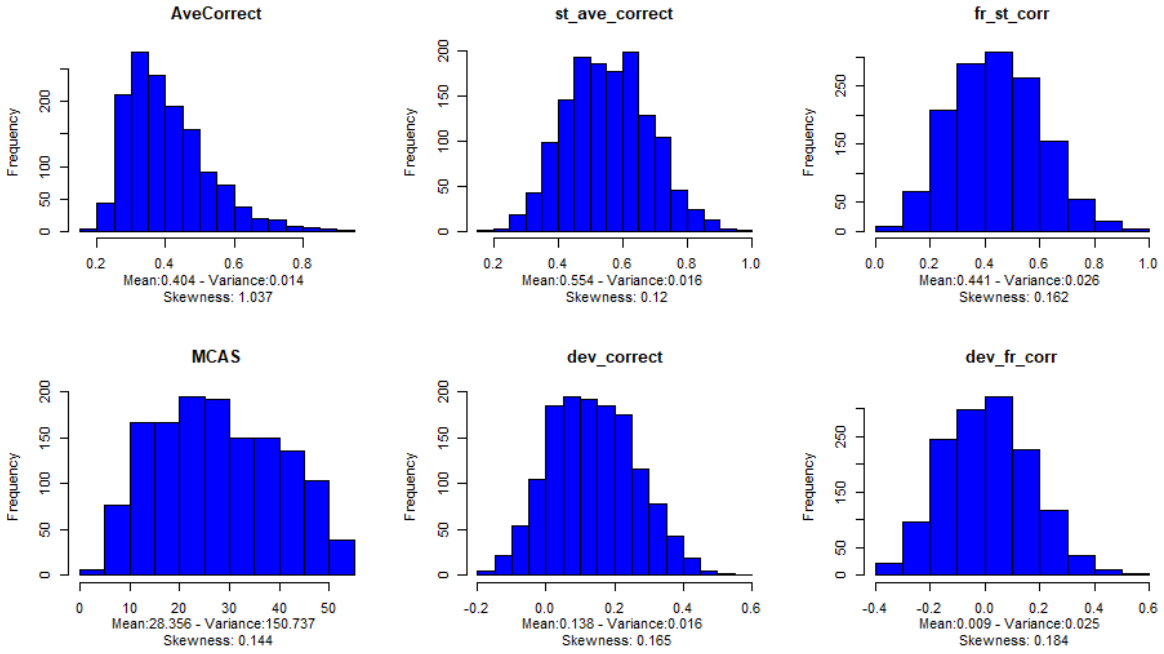
For each student, the following metrics related to score and performance are available:

- **AveCorrect**: number of actions divided by number of correct answers. It is not a reliable measure since some actions are help requests or scaffolding steps, for which the "correct answer" dummy variable collects 0. It tends to underestimate students' score. The positive skewness of its distribution supports this hypothesis.
- **MCAS**: it is the final score each student obtained in a time-limited standardized test which is taken at the end of the school year. It represents students' achievement level or test performance. It is an example of a traditional parameter used to classify students.
- **St_Ave_Correct** and **Dev_Correct**: mean of the scores gained in each problem and mean of the deviation of the scores from each problem's average. The second variable controls for the difficulty level of the problems encountered. Although the distributions are very similar, students differ both in the number and in the type of problems completed. Therefore, the latter variable seems a much less biased metric.
- **Fr_St_Corr** and **Dev_Fr_Corr**: percentage of correct answers in first attempts (mean for each problem) and its deviation from problem average.

These variables are similar to the previous ones; however, they take into account only first attempts: thus, they can be interpreted as students' knowledge before accessing any help or correctness feedback. They measure the number of times students readily gave a correct answer.

Even though there are obviously differences in the mean and variance among the different schools, the distributions look always very similar.

Figure 4.2 Histograms of Performance Variables



Looking at the scatterplot and at the correlation matrix, it appears clearly that the last four variables are highly correlated to each other. With regard to the correlation coefficients of MCAS, there is always a positive strong correlation between the performance measures and MCAS. Dev_Correct shows the highest one and this happens also when we look at single school and year datasets: this high coefficient suggests that Dev_Correct may be more useful than other metrics in predicting accurately MCAS. There are differences among the schools and the school year in terms of mean and variance values. However, the variable distributions are always very similar: the AveCorrect variable always show quite a high skewness coefficient;

Figure 4.3 Scatterplot Matrix of Performance Variables

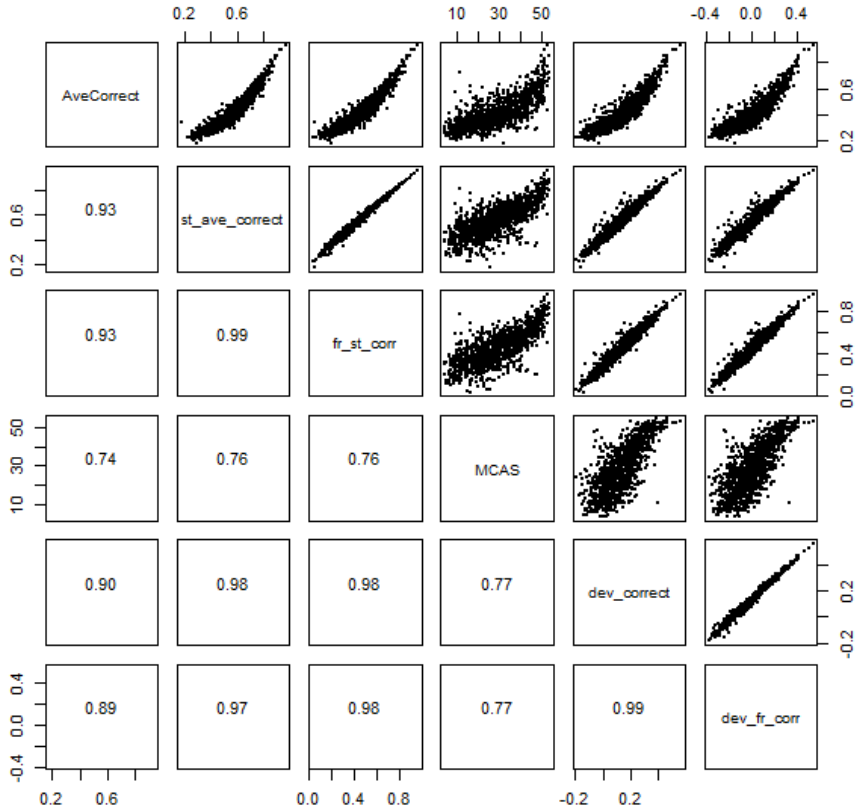
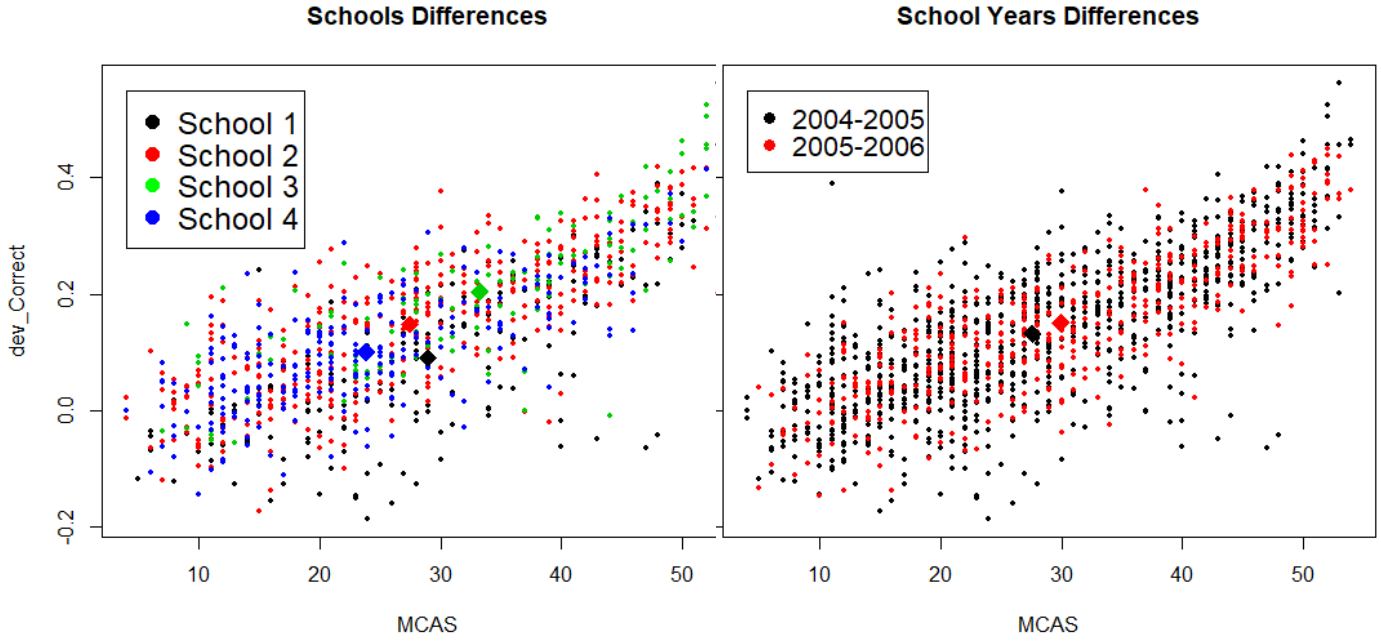


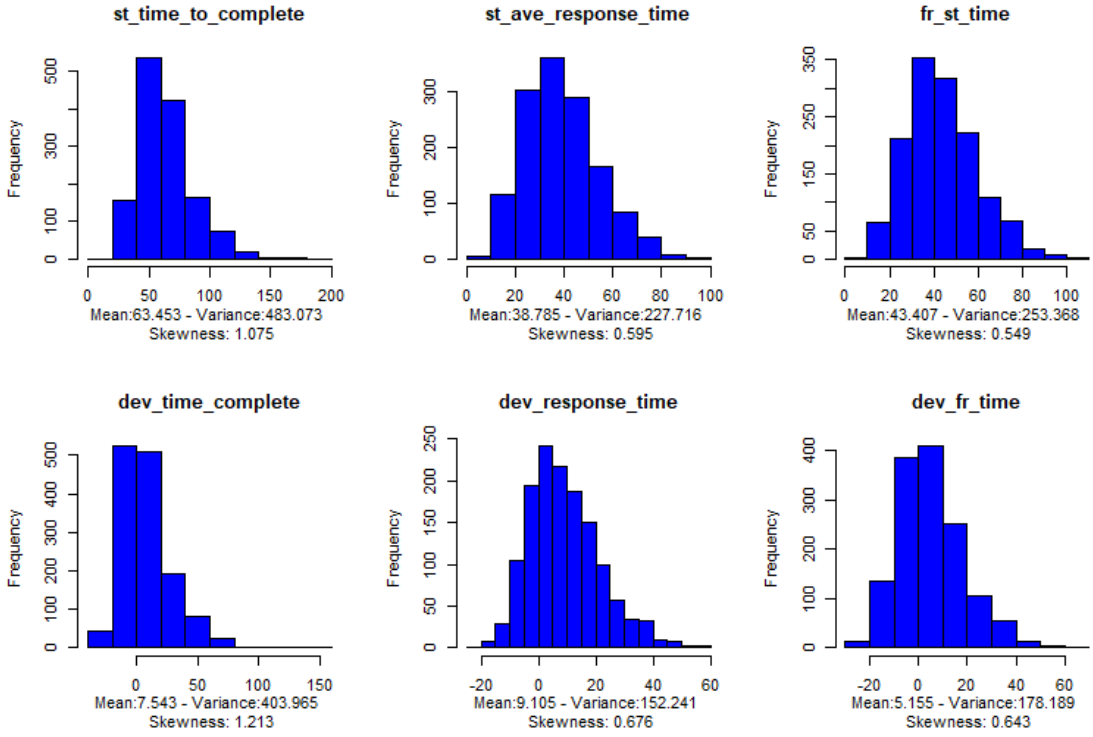
Figure 4.4 Differences in MCAS and Dev_Correct scores among Schools and School Years



Time Measures

- **St_Time_to_Complete:** it measures the average time a pupil takes to complete a single problem. It represents the time taken to solve the problem. This variable is likely to depend strongly on the difficulty level of the problem. Therefore, a better measure would be the mean of the deviations from problem average.
- **St_Ave_Response_Time:** it is the average time a student takes before choosing an action. It represents the readiness of the student and its confidence level. It is not directly related to the difficulty of the problem: after the first attempt, students may take the time to consider other solutions or just try to give a different answer. It can be interpreted as students' average decision speed.
- **Fr_St_Time:** it takes into account only first responses. It measures students' carelessness while approaching a problem. It is important to evaluate it in terms of deviations from the average since some problems might be significantly more time-consuming than others. Therefore, a short time taken before the first attempt could identify a careless student or a fairly easy problem.

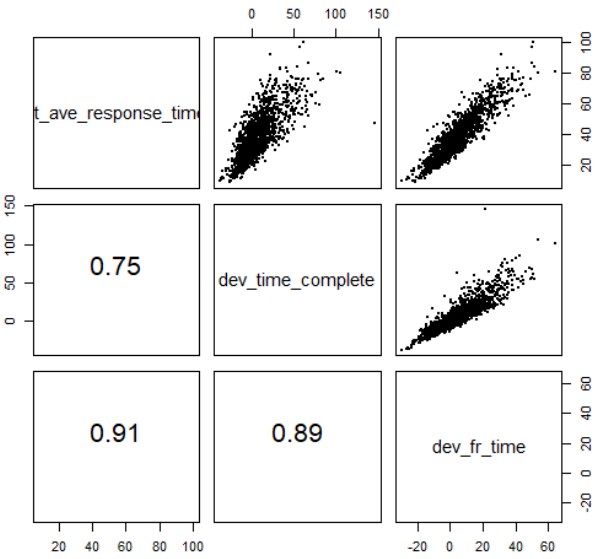
Figure 4.5 Histograms of Time Variables



Since these are time measures, a lower value means either higher speed or confidence, or higher carelessness. On the other hand, a higher value can be interpreted as a slow or more conscientious student. All of these variables have positive-skewed distributions: more than half of the students show values lower than the mean, which is a positive signal in terms of students' speed level.

The variables related to time are all positively correlated. As time of completion increases, the variance of response time becomes larger: this supports the hypothesis that some students may think the problem through (high response time) but have relatively low completion time, while others may be more spontaneous (low response time) but have the same average completion time because they use a higher number of attempts to solve the problem. This heteroskedastic relationship between these two variables may prove fundamental in the following clusterization process.

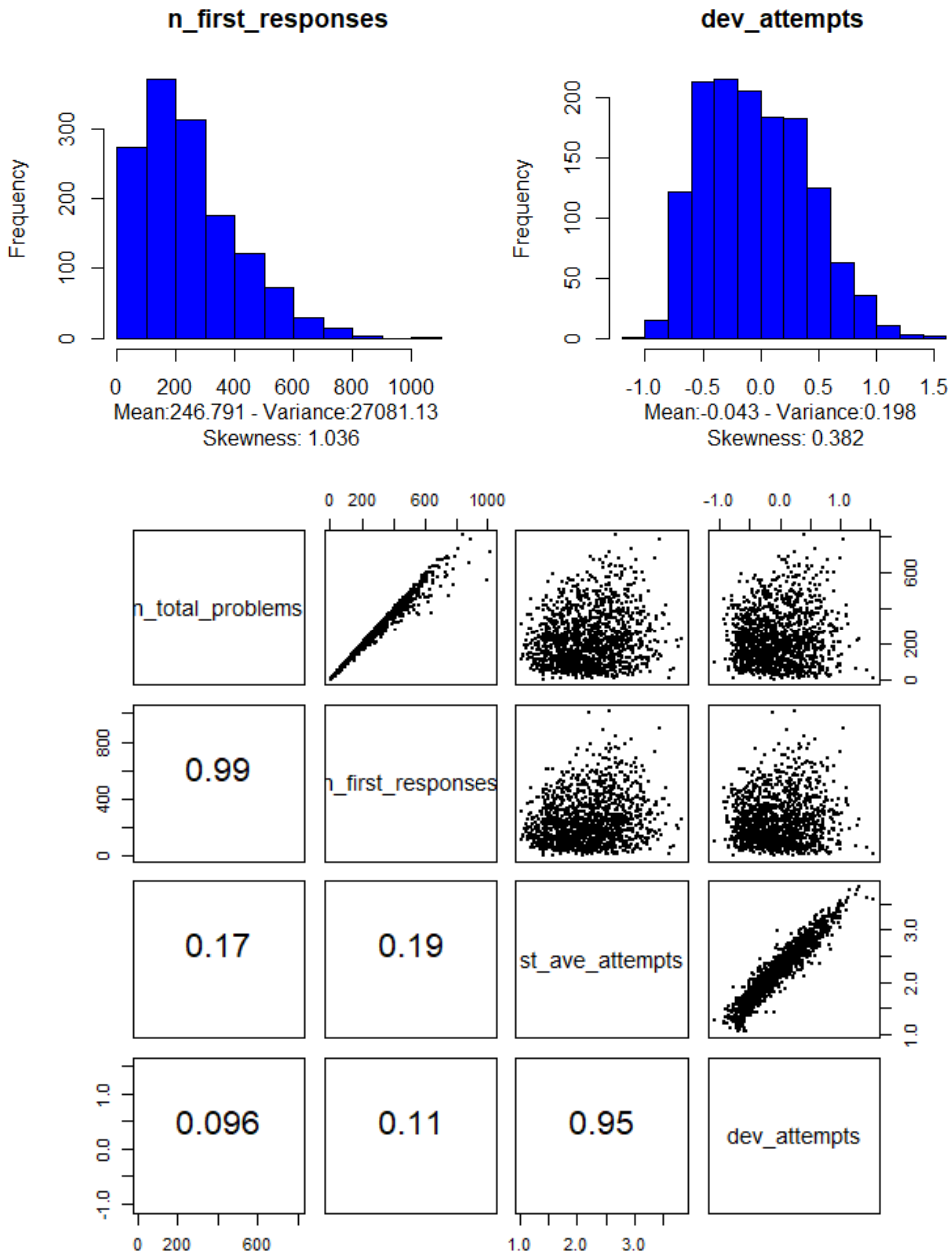
Figure 4.6 Scatter Matrix of Time Variables



Attempts and Practice Measures

- **N_Total_Problems**: number of unique problems encountered. The number of actions is not a reliable measure since it does not consider the number of attempts for each problem.
- **N_Fr_Responses**: number of solved exercises. The same problem can be solved by the same student more than once. This variable is highly correlated to the previous one (ρ=0.99). This second measure is selected because it represents more diretly the concept of students' practice level on ASSISTments.
- **St_Ave_Attempts** and **Dev_Attempts**: the average number of attempts the student takes to solve a problem. It is likely to depend on the type of problems. Therefore, the latter is selected. It depends both on students' knowledge and motivation.

Figure 4.7 Histograms and Scatter Matrix of Practice Variables



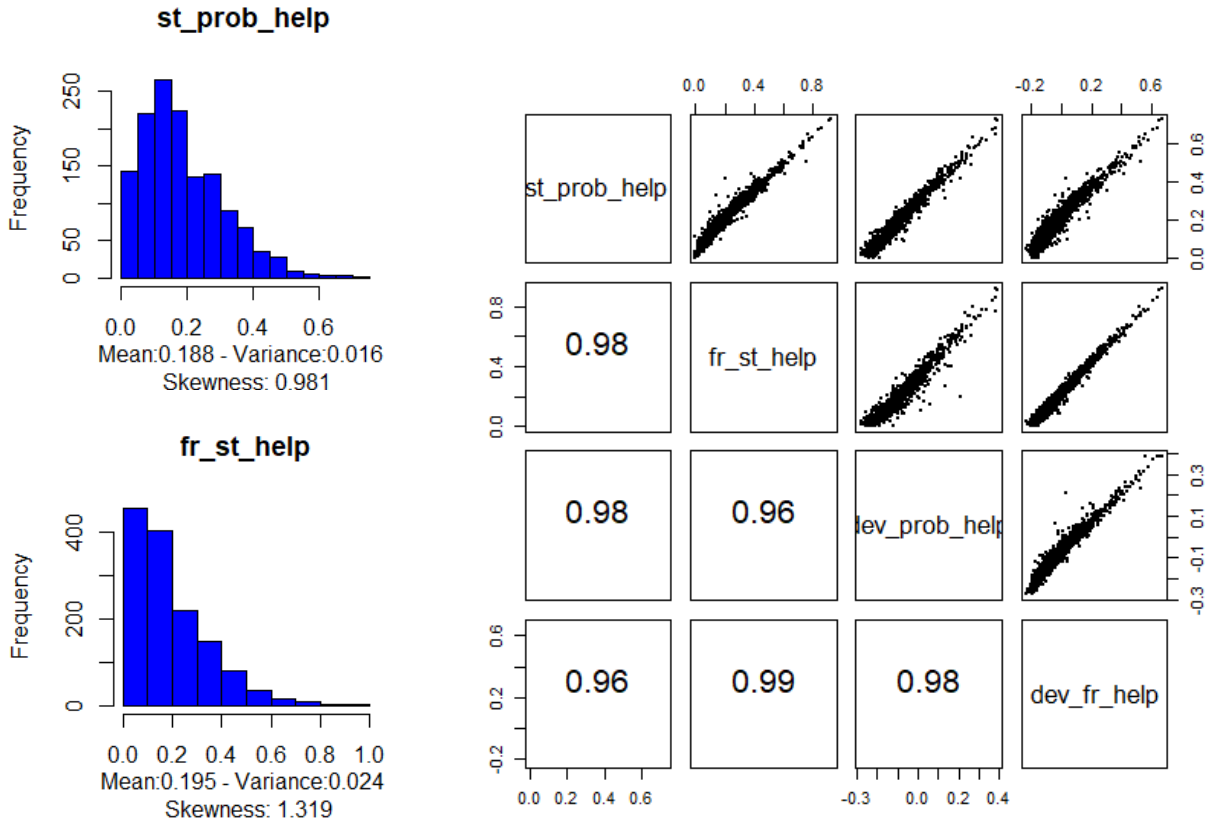
In some schools, the level of practice and the average attempt measure are negatively correlated: a possible explanation could be that the more a pupil practices, the easier it becomes for him to solve the problem. However, since it does not happen in all of the schools, this relationship is due to some factor which is not been taken into account. For instance, the average practice levels of these schools may be higher than in the other schools, meaning that students used ASSISTments intensively during the school year. On the contrary, evidence shows that these schools are characterized by a lower level of practice on ASSISTments.

This example highlights the importance of considering closely school specific features and the fact that generalization can be misleading.

Help Measures

- **St_Pr_Help:** probabity of a help request. It is computed as the number of hint or scaffolding requests on a single problem divided by the actions on that problem. Although its deviation has been computed, this variable is likely to depend on inner characteristics of the student rather than on the difficulty of the problem. Moreover, it is more useful to maintain the metric as a probability ranging from 0 to 1. This variable represents students' confidence and dedication when solving a problem. The higher this probability, the less the student relies on his own effort to solve the problem.
- **Fr_St_Help:** probability of a help request in first attempts. It also measure self-esteem and motivation. Generally, it is higher than the previous one. The deviation measure do not show significant differences: therefore, the original measures are selected.

Figure 4.8 Scatter Matrix of Help Variables



4.3 Schools' Datasets Description and Selection

Having selected all of the meaningful features, it is interesting to investigate further their mutual relationships. As it can be seen from the scatterplot, there is a number of strong relationships between features. The variables related to score appear usually negatively correlated with the number of attempts and the probability of asking for help, while the relationship between time and performance is not so neat. The response time is positively correlated with performance variables, meaning that the variable may be more accurately interpreted as a conscientiousness measure rather than speed.

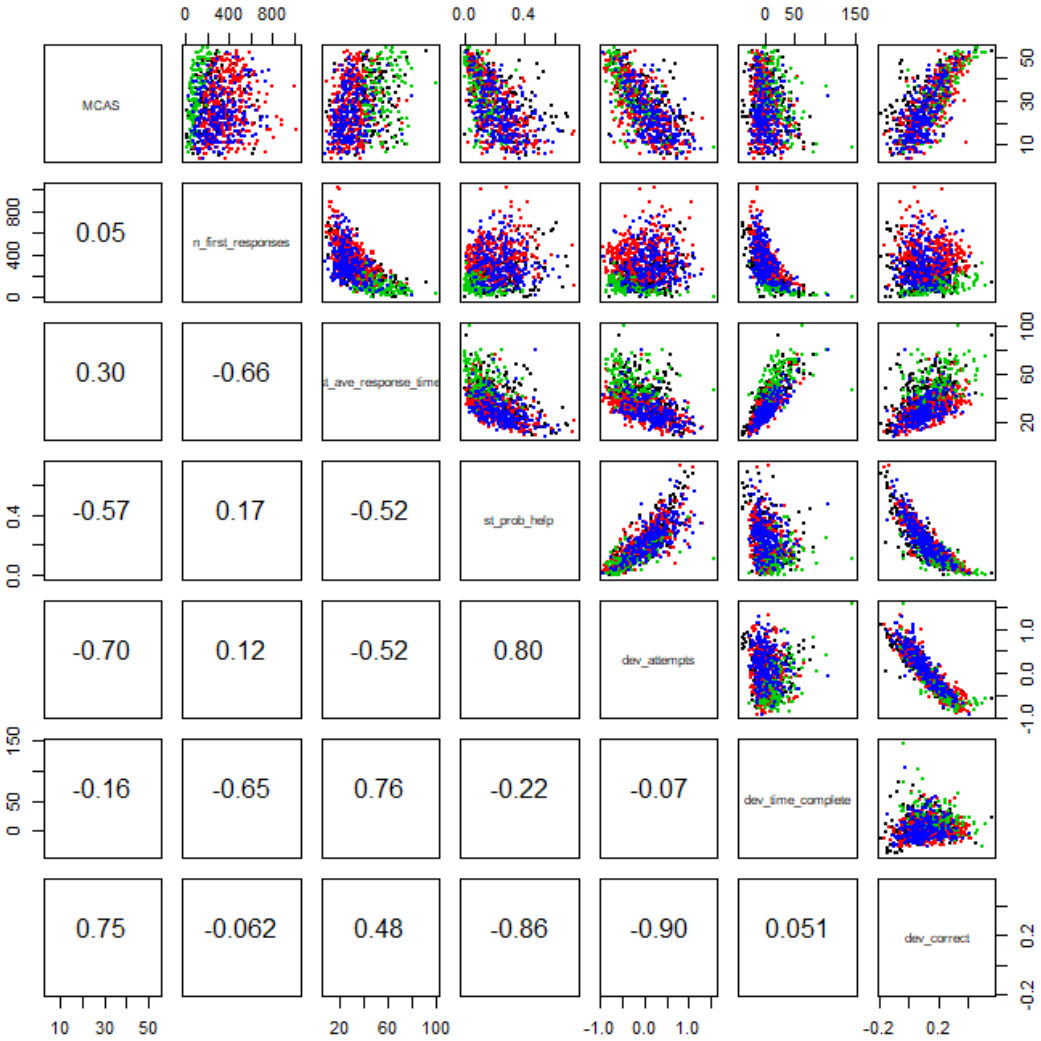


Figure 4.9 Scatter Matrix of Selected Features

However, time-related variables are also negatively correlated to the practice level, which means that the more students practice, the faster they become.

Since students' metrics from different school populations can differ for factors that this analysis does not take into account (e.g. teachers' knowledge level, school expenditures, neighbourhood median income..), it is considered important to divide students into schools and school years so as to remove these possible differences. However, since these schools are all located in the same city, it is possible that those factors are not as relevant as to cause significant differences. In order to test the hypothesis that the means are equal in all of the sample schools, an analysis of variance is conducted for each variable [27]. If the variance between groups proves to be significantly larger than the variance within groups, then it will be concluded that there are in fact significant between-group differences.

Table 4.2 - Selected Features School Averages

(School Year: 2004-2005)

Variables	**Meaning**	**1**	**2**	**3**	**4**	**ANOVA**
MCAS	Test Score	28.94	27.49	33.24	23.81	18.34 ***
N_First_Responses	Practice	219.8	363.2	98.74	315.3	121.4 ***
St_Response_Time	Carelessness (-)	41.257	31.464	54.22	28.372	155.7 ***
St_Prob_Help	Demotivation	0.2304	0.204	0.1166	0.2441	31.51 ***
Fr_St_Help	Demotivation	0.2533	0.213	0.11701	0.2497	23.34 ***
Dev_Attempts	Readiness (-)	0.05515	-0.0612	-0.2825	0.1604	31.58 ***
Dev_time_complete	Speed (-)	5.925	2.244	20.668	2.522	33.92 ***
Dev_correct	Performance	0.09014	0.148	0.2042	0.1006	30.23 ***
Dev_fr_correct	Performance	-0.04884	0.0245	0.0896	-0.0371	29.42 ***
Dev_fr_time	Carelessness (-)	4.098	1.834	15.165	0.3251	45.8 ***

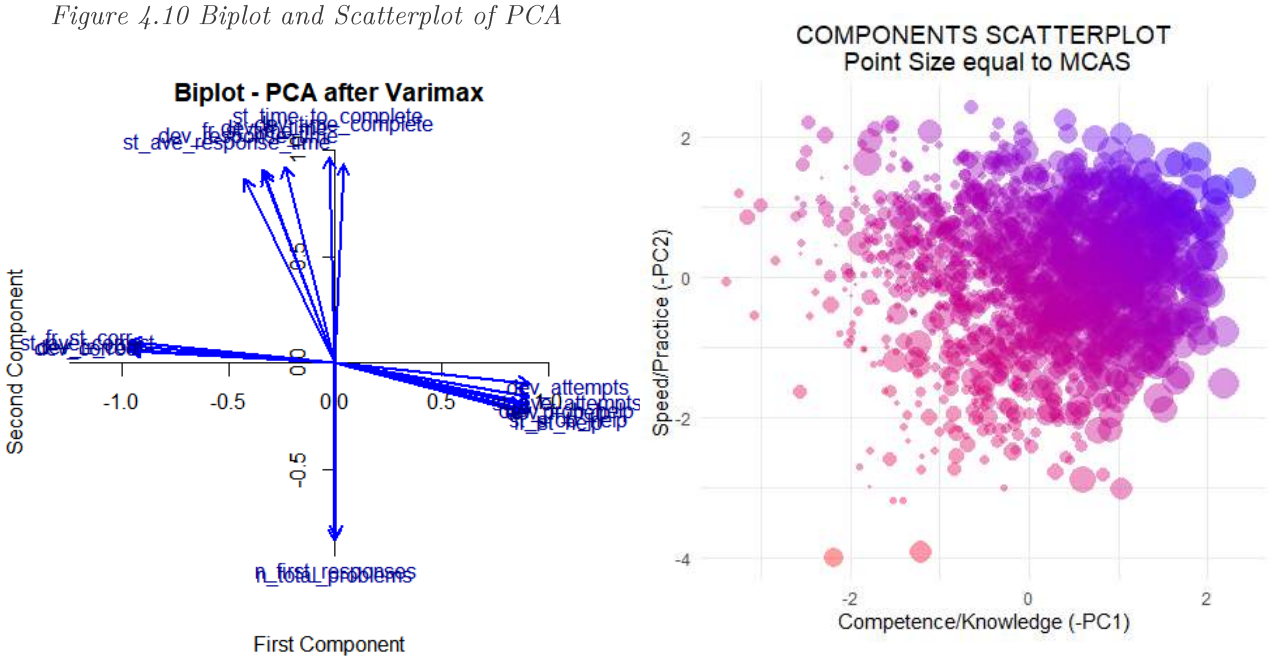
Each of the univariate ANOVA F-Test indicates that there are significant differences among schools. School 2 is chosen for the first cluster analysis because it contains the largest number of students. Moreover, its results are comparable to other schools, as well as with the same school during the following school year.

4.4 Principal Components Analysis

Lastly, since the number of variables is high and the correlation coefficients are quite strong, it is possible that dimensionality reduction techniques may yield good results in terms of interpretability variance and conservation. A principal component analysis is implemented using the whole dataset, in order to assess whether reducing the number of variables may prove convenient. All of the available features (including the ones that were not tested) but MCAS are included. Scaling is needed since variables are expressed in different units of measurement.

The results appear extremely promising given that the first two components retain almost 87 % of the original variance. However, the loadings and the biplot show that no simple interpretation is possible for the new components. A rotation of these components is performed: an orthogonal rotation, in particular the "varimax" rotation, was chosen [28]. This rotation aims at maximizing the sum of variance of squared loadings: as a consequence, the new loadings for each component will be either high or near-zero. It helps interpretability of the correlations between the original variables and the new components.

Figure 4.10 Biplot and Scatterplot of PCA



Given its loadings, the first component can be interpreted as an overall measure of students' knowledge, which takes into account both scores and students' usual approach to the assignments. The lower the score on the first component is, the better and more prepared the student will be.

On the other hand, the second component is made up by two other important aspects: time and practice. The PCA has emphasized the negative correlation running between time-related variables and practice level: as a consequence, it must be assumed in the interpretation that the more students practice, the faster they will become.

Both components' scores are multiplied by -1 so that an excellent student would be characterized by higher values both on PC1 and PC2. As PCA was performed, the new components are not correlated to each other: their distributions both have a slightly skewed bell shape. The scatterplot does not reveal any obvious group: therefore, a cluster analysis may prove a useful way to find out these subtle borders. Given the excellent results obtained by PCA, the components will be used to carry out the cluster analyses and the outcome will be compared to the one obtained using the original variables.

Regardless of the groups identification goal, which was the reason to implement PCA, this method has proved useful in itself since it summarizes 17 variables in just two practical scores for each student, whose meaning is easily comprehensible. In addition to the loadings of the first component, the high positive correlation with MCAS adds evidence to the fact that the first component is a performance measure.

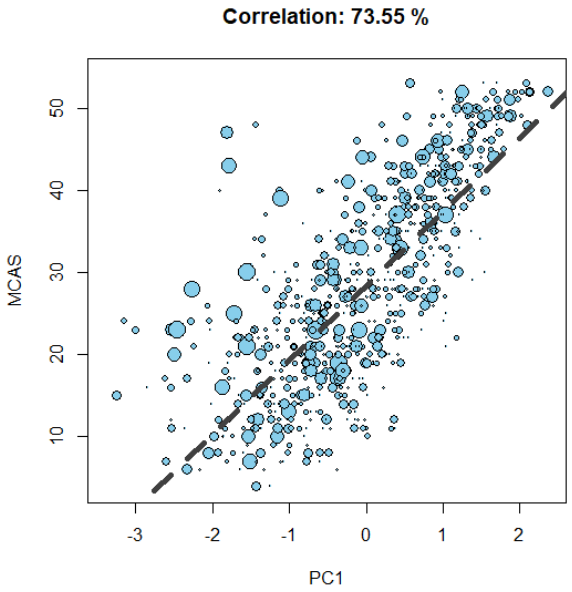


Figure 4.11 Scatter Plot of PC1 and MCAS. Point radius equal to FC2

4.5 Cluster Analysis: Goal Setting and Preparation

In order to obtain meaningful results, it is important to set the ultimate goal of the analysis. Teachers or principals may be interested in identifying categories such as at-risk, average, and excellent students. The final score is useful but not sufficient in order to describe accurately each student: other variables such as response speed and motivation may incredibly improve the ability to recognize different pupils' needs and address them. Given the information contained in the

original dataset, some variables related to these characteristics have been easily computed.

The ultimate goal which was chosen for this case study is to divide students into categories created along three dimensions: knowledge, speed, motivation. These three aspects are chosen both because of their understandability and because of measures' availability. Although other criteria are interesting as well, their conceptualization and measurement might be more complicated. The aim is to create a model that allows teachers to efficiently address the needs of each of these groups based on different scores in each of the dimensions. For instance, learning difficulties, slowness, and carelessness can be addressed respectively with revision, practice, and stimulation. The resulting clusters should be clearly divided along these three dimensions. Although all of the variables have continuous distributions, if we can think of their values as either low-level or high-level, then there are at least 8 possible combinations (2^3), so this number could be a starting point for the choice of the optimal number of clusters, which should mirror the combinations found in a given school.

It may be reasonable to think that the more information is held about the students, the better the results. However, some variables may be misinforming about students, either because they are biased or because they are highly correlated to others. This is the case of the dataset chosen for the analysis: firstly, there is a number of factors that could have biased the data and made it not comparable (e.g. cheating, later improvement, etc...); secondly, including two or more highly correlated features gives too much importance to the specific aspect they describe to the detriment of others. Since the goal was clearly set out, each investigated dimensions will be given the same importance. In view of the above arguments and of the ultimate goal, the first of the analyses should regard choosing the appropriate subset of meaningful features, on which the divisions will be made.

The final issue concerns the inclusion of MCAS: the main advantage of the implementation of this method is considered to be the ability of improving students' probability of success. Therefore, it may be more useful not to include the final score variable but rather using it to check the consistency of the results. MCAS may become the target variable, whose value may be predicted given a certain students' group. By excluding the final score, all of the variables included in the analysis are derived from the log-data file: the usefulness of the cluster analysis

will also be a test on one possible advantage of log-data exploitation (i.e. educational data mining).

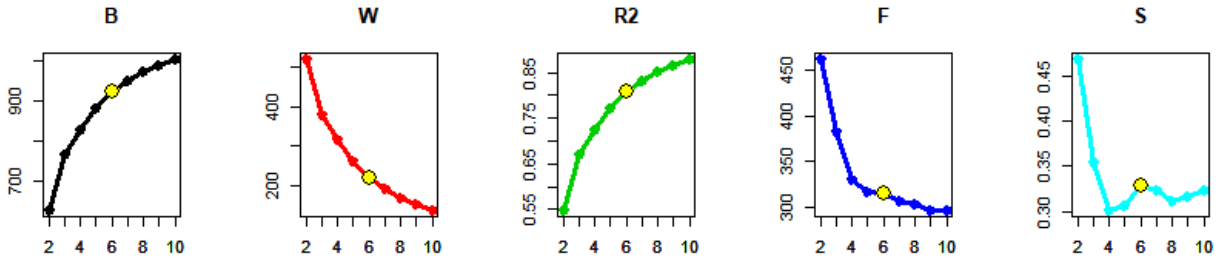
Since the variables are expressed in different units of measurement, they must be scaled before proceeding with any clustering algorithm.

4.6 Non-Hierarchical Clustering Algorithm: K-Means

In the first part, the clustering algorithm chosen is the K-means algorithm, which is a centroid method [29]. This algorithms needs a random initialization on which the results are going to depend.

Numerous combinations of the interesting variables are used and evaluated. Each subset contains a speed-related variable, a correctness variable, and a help probability variable. Some subsets also contain the average number of attempts and the average completion time. Including the number of first responses may be misleading since it probably did not depend on students' choices. Given these subsets, the ones that yield the best results in terms of low within-variance are chosen. In order to evaluate the goodness of the models and to choose the optimal number of clusters, five statistical measures are employed: between variance, within variance, R^2, F statistics or CH index [30], and silhouette index (average) [31]. Given a maximum number of clusters, each of these measures is computed and plotted against the number of groups (K). Since the goal of the analysis is to provide teachers with a simple and effective tool thanks to whom they will able to create customized approaches for each group, the optimal number of clusters must be limited: otherwise, the tool will not be very useful to teachers in spite of its improved accuracy. As a consequence, it has been decided that the number of clusters cannot be larger than 10.

Figure 4.12 Subset A's Statistical Measures for K from 2 to 10



The best subset includes the following variables: average response time; help request frequency; deviation of correctness (subset A). Although including the number of average attempts and average time to complete a problem does not yield

better results, these variables are considered important to efficiently divide students in terms of speed and motivation. In order to counterbalance the effect caused by the inclusion of these variables, the percentage of correct answers in first responses is included as well (subset B). Subset C contains the same variables of subset A, but measured only in first responses: according to the statistical measures employed, this model performs worse than the previous one and, therefore, it is discarded. Finally, the k-means algorithm is also performed on the principal components: although the performance is slightly lower, the interpretation which was given to the components will help describe the clusters.

Table 4.3 - Statistical Measures of Group Homogeneity

	K	B	W	R^2	CH	S
Subset A: st_response_time, st_prob_help, dev_correct	6	922.76	220.24	0.807	315.07	0.328
Subset B: A + dev_attempts, dev_time_complete, dev_fr_correct	6	1772.20	513.82	0.775	259.37	0.298
Subset C: st_fr_help, dev_fr_correct, dev_fr_time	7	933.28	209.72	0.817	278.13	0.330
Subset PC: Competence and Speed	6	498.52	138.33	0.783	271.02	0.335

Subset A

Starting from K equal to 2, observations are first divided into high-score and low-score students. With a higher number of clusters, divisions based on time and speed start emerging. A group of slow students is isolated from the other observations, while the faster students are progressively divided into smaller groups based on competence. As K gets closer to 10, more time-related boundaries emerge. Clusters are clearly separated along the dimensions of the variables which were employed in the algorithm: the groups have different means and are only

slightly overlapping. This is also the case for the variables that are highly correlated to the ones of subset A. However, all of the groups show high variance in MCAS: the groups overlap each other, and the interpretation becomes difficult. This could suggest that the hypothesis that MCAS accurately describes students' attitude may be wrong: the clusterization depends not only on score variables, but also time and motivation variables, which are aspects that are not considered by MCAS.

Figure 4.13 K-Means on Subset A with 6 clusters

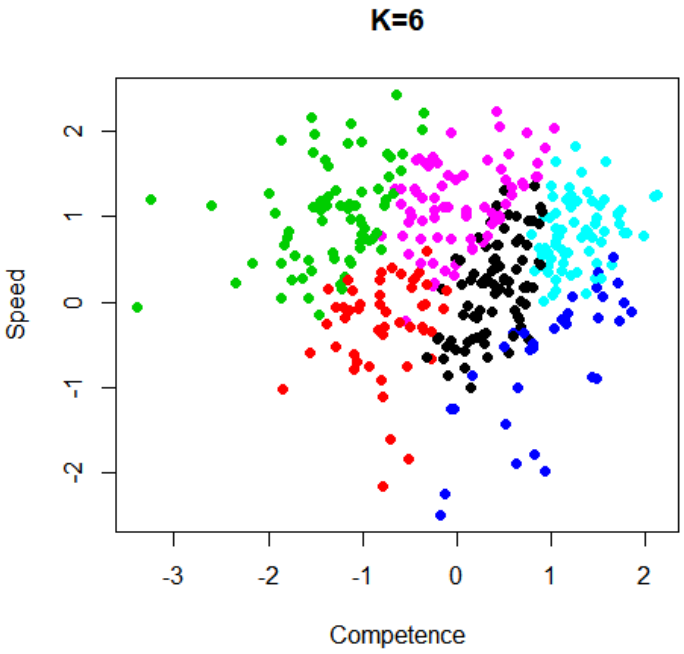


Table 4.4 – K-Means Clusters Mean Values

Subset A	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Size	83	49	73	33	70	74	382
MCAS	27.94	17.102	16.877	34.212	41.914	27.703	27.49
n_first_responses	297.157	233.000	448.219	234.000	386.543	475.338	363.2
st_response_time	35.801	35.446	19.217	53.487	34.645	23.213	31.464
st_prob_help	0.146	0.285	0.387	0.091	0.075	0.204	0.204
fr_st_help	0.137	0.306	0.446	0.084	0.065	0.204	0.213
dev_attempts	-0.227	0.236	0.514	-0.514	-0.549	0.023	-0.061
dev_time_complete	9.320	16.466	-8.216	22.175	-4.110	-7.670	2.244
dev_correct	0.183	0.061	-0.015	0.289	0.297	0.125	0.148
dev_fr_corr	0.069	-0.082	-0.180	0.200	0.210	-0.007	0.024
dev_fr_time	7.199	8.869	-9.000	20.431	1.386	-6.021	1.834

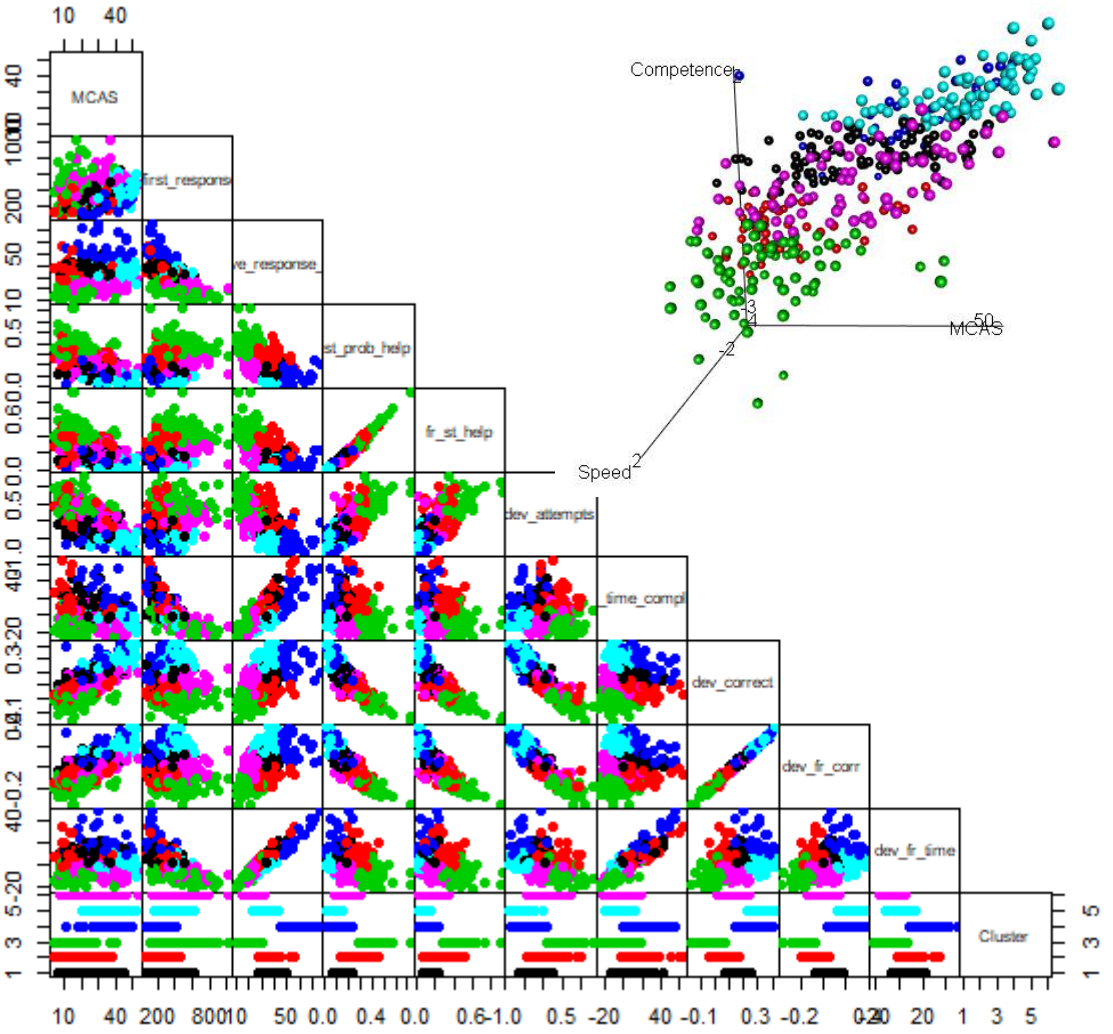


Figure 4.14 Scatter Matrix and 3D plot with clusters obtained by K-Means on Subset A

- **Cluster 1.** This is the largest cluster with an average MCAS score really close to the school average and a high variance. Its average correctness percentage lies in the middle of the distributions, slightly higher than the school averages, and its variances are the lowest among all of the groups: students' scores lie in a specific narrow region, near the middle. Students are slightly slower than the school average; they tend not to ask for help and their average number of attempts is moderate. Not much can be said about this group since it is not strongly marked by any feature. They position themselves in the middle of all of the distributions, with relatively limited variances. The average student has a score higher than 50 %.

- **Cluster 2.** This group is made up by low-achieving students, whose results and scores are slightly higher than the worst group (cluster 3). These pupils take more time on average than top students before answering a problem:

therefore, they seem to care about their assignments and their results. Their high number of attempts is another sign of their lack of knowledge.

Their probability of asking for help is pretty high as its variance is. Together with cluster 3, these are the only two groups in which the probability of asking for help increases in the steps following the first attempts. These students might get frustrated after a series of wrong attempts and, therefore, their confidence in solving the problem may waver, resulting in a help request.

In terms of MCAS score, they fail to reach good results and their average score is well below the school average. Their MCAS maximum value is lower than the one of cluster 3: this is evidence supporting the hypothesis that these students are more engaged and more motivated than others, but they lack the necessary raw material, which is knowledge and topic understanding.

Cluster 3. These students have a very low average response time and they have on average a 44 % probability of requesting help during the first attempt. This group has the worst results in terms of scores as well as number of attempts. Although the number of attempts is high, the completion time stays low since some hints or help request lead to the correct answer in few steps. In terms of score variables, this group shows the worst results: even its variance and maximum values are quite low as well. Although the variance in MCAS is the lowest among the group, it is still quite high with a maximum of 40: their true level of knowledge seems difficult to assess using assignment results.

It is possible that these students answer without care or directly ask for help, thereby showing little interest in practicing their knowledge during the assignments.

Given the strong reliance on help and the tiny amount of time spent considering the problem, this cluster contains demotivated, insecure, or careless students who appear to have little incentive to perform well during the homework. This group contains both average and low-achieving students. Although some of them do not seem at risk of failure, identifying their carelessness and demotivation may act as a warning for teachers and parents in order to prevent future issues.

- **Cluster 4.** It is the smallest cluster among the groups. It contains high-achieving students who generally score well in homework assignments: their

results are similar to the ones of the top group (cluster 5). The difference lies in speed: the average student of this group takes 20 more seconds to enter a response than the average of cluster 5. This gap can be found also in first responses. Even though the number of attempts is quite similar, the average time to complete a problem remains substantially higher. Moreover, they are the slowest group among all of the clusters. It may be that this speed difference is caused by different levels of practice and familiarity with ASSISTments platform: in fact, this group is characterized by one of the lowest practice level.

However, speed is not the only difference between the two high-achieving groups: in MCAS, the average performance is 10 points lower in this group, whose students fail to reach the top score (maximum value: 50). This might be a consequence of what the clusterization has pointed out: these students are in general slower than their fellows in group 5 and take more time reflecting on the problems. Although some of this slowness may be the effect of scarce practice, it is likely that it represents an inherited feature of these students. This hypothesis is supported by the fact that MCAS (which has a time limit) scores are lower: it may be that these students are well prepared, but they are too slow to reach top results in final tests. A recommended approach towards these students in order to smooth their performance may be extra-practice.

Another hypothesis is that these pupils lack self-esteem or confidence and, therefore, they tend to become excessively conscientious.

- **Cluster 5.** It is the best group in terms of achievement. It performs better than all of the others in percentage of correctness, correctness in first responses, and MCAS. This group is undoubtedly the best regarding the results of ASSISTments homework. This cluster is also characterized by the least average probability of asking for help. These students may not need help to answer the problems, because they are likely to be able to solve the problems on their own. The average number of attempts is also the lowest, confirming the hypothesis that these students are the best ones.

  Their good results in assignments might be due to a high level of practice on the platform: in fact, the group has the third higher number of average problems per student. However, practice level was not used in the clustering

algorithm on purpose since dividing students in terms of this variables could add bias to the results.

When it comes to the average response time, this group scores quite low in comparison to the other high-achieving group (cluster 5), but higher than other groups and the average. However, response time is not only a speed measure but also a carelessness measure, so it is important to look at other speed variables.

In fact, the average time to complete a problem is closer to the minimum value and well below the school average. This difference provides additional evidence that the average response time may be negatively correlated with carelessness. The best students in terms of confidence and scores are not the first in terms of response speed: however, since they usually answer correctly in few attempts, their speed reduces when the overall time to complete the problem is considered.

In conclusion, this group can be described as: competent, motivated, confident, relatively fast.

- **Cluster 6.** This group has the largest variance of MCAS score and the closest score to the school average: this means that belonging to this group tells us almost nothing about probability of success. In terms of speed, this group appears quite careless and demotivated. On the contrary, they tend not to ask for help neither in first nor following attempts. Their correctness percentage are slightly worse than the school means.

  It is possible that these students briefly evaluate the problem and then choose the most plausible answer without putting much effort in the decision-making process. They may not consider help useful for the completion of the problem. The fact that they score higher than cluster 3 may result from a mix of factors: they take slightly more time in the problem's evaluation, thereby increasing the chance of answering correctly; they do not use help request, which usually marks the problem as wrong; they may be more prepared given their average result in MCAS in comparison to cluster 3.

  In light of this analysis, it is likely that cluster 3 is made up of students who lack confidence and preparation: this leads them to rely on help requests whenever they do not have a clue. They often refuse to take even a first evaluation of the problem and directly dive into the help strategy. What was interpreted

as carelessness may be just a sign of profound incompetence and lack of understanding of the subject.

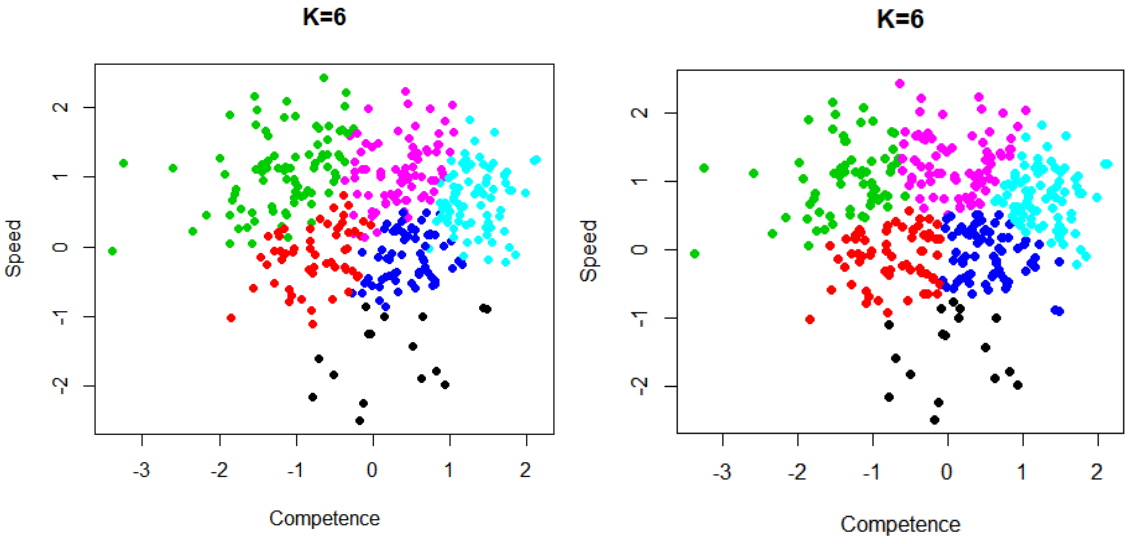
On the other hand, students of cluster 4 are more likely to get good results in the final test. They do not seem to care about help even when they cannot solve the problem. They take less time than top students to complete problems. Given these pieces of evidence, it may be argued that this is the most careless group of students in the school.

Subsets B and PC

There are some differences between the results of the two subsets. However, the pattern is similar as well as the actual boundaries as long as K is relatively small. The optimal number of clusters for both of the subsets is equal to 6. Subset B performs worse than subset A, even though the former contains more variables.

The results in subset B are quite similar to subset A's, except for the first group. In subset B, Cluster 1 contains less than 20 observations which are characterized by high response times. This group can be identified as speed outliers. These "slow" subjects are included in Cluster 4 when subset A is used. This new group has also the minimum level of practice on ASSISTments with a tight variance. Its slowness may be due both to lack of familiarity with the platform and to troubles encountered by students while completing the assignment (e.g. computer crashes, net lagging).

Figure 4.15 K-Means on Subset B (left) and Subset PC



With K equal to 6, clusterization on subset PC gives almost the same exact results as subset B: it keeps isolating the outliers while the other observations are divided along the same boundaries. In terms of goodness, subset PC performs slightly better. Using subset PC, the interpretation for clusters 2, 3, 5, 6 remains the same. As said before, Cluster 1 becomes the outlier group while Cluster 4 changes its composition: it now includes students with good scores and average speed which were part of Cluster 1 in subset A. It represents a average-achieving students' group whose performance, however, does not reach the top results of cluster 5, neither in speed nor correctness percentage. Their results in MCAS test are highly variable.

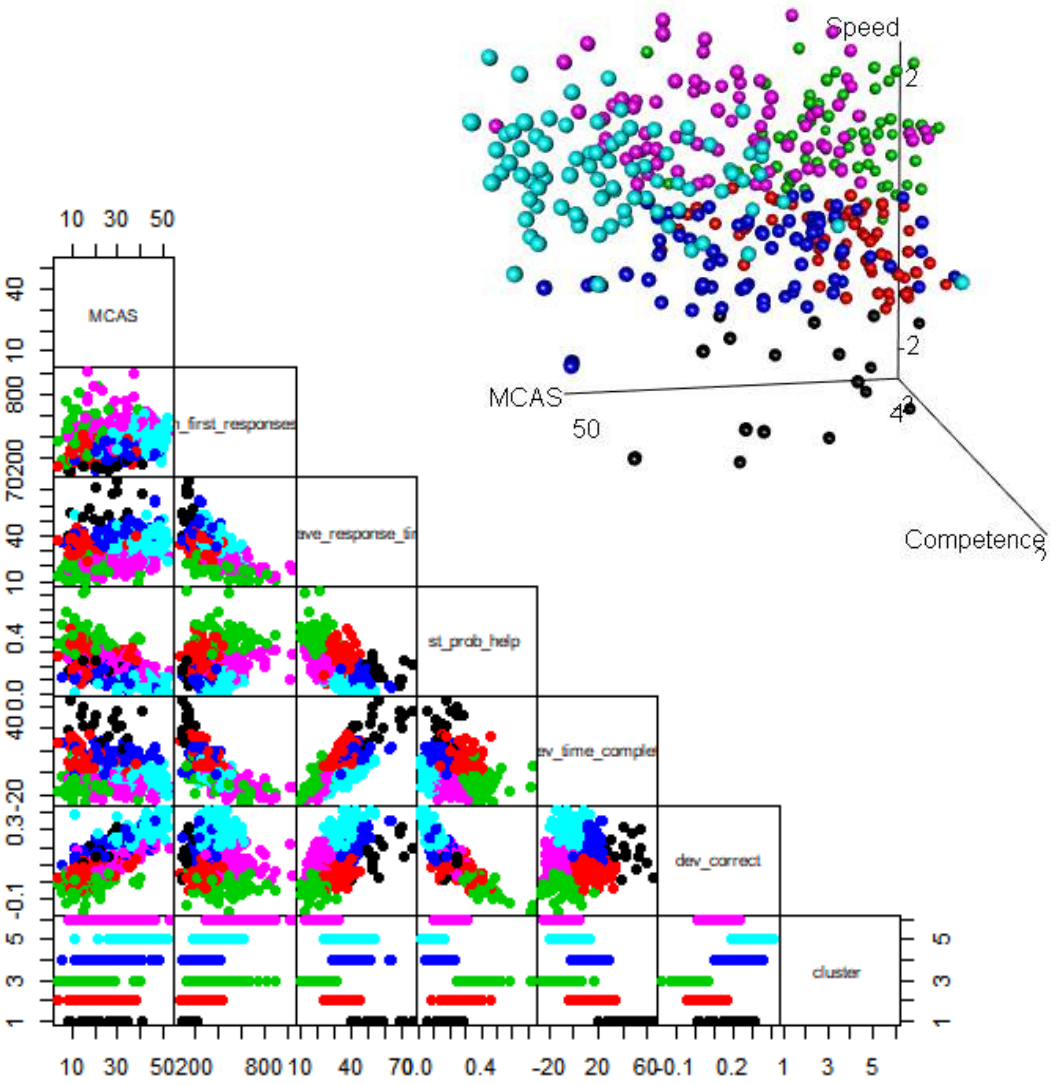


Figure 4.16 Scatter Matrix and 3D plot with clusters obtained by K-Means on Subset PC

4.7 Hierarchical Agglomerative Clustering

In this second analysis, a hierarchical agglomerative clustering algorithm is employed. This method is called hierarchical because it starts from a situation where each observation represents a unique cluster, and then the most similar clusters are aggregated. The process goes on until all of the observations are included in a single cluster. This process does not include a random initialization. [24]

For this analysis, subset A and subset PC were used. In addition to those, a subset SF, which contains all of the selected features except MCAS and practice level (n_first_responses) was employed as well.

Five different types of linkage are employed. However, only two of them seem to provide balanced results: "complete" and "Ward" methods. In particular, as it can be seen from the dendrograms, the "average" method gives highly unbalanced clusters, while the "centroid" method shows inversions, and the "single" method suffers from chaining. These phenomena occur for each of the subsets.

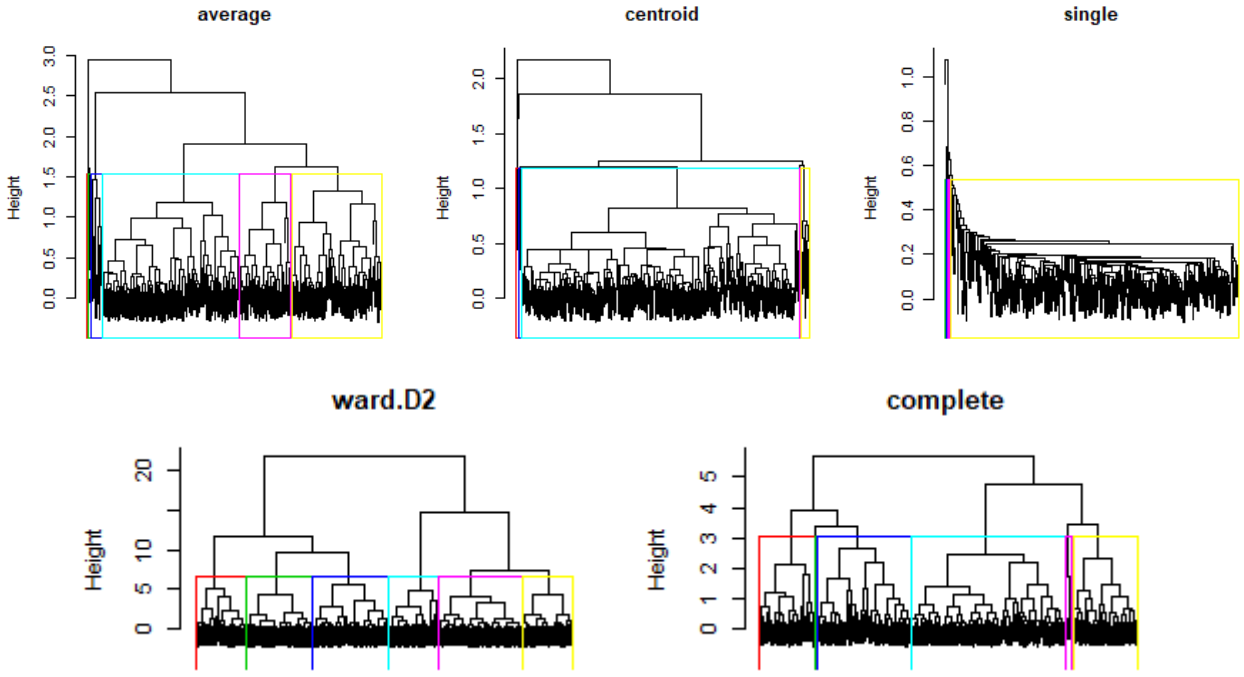
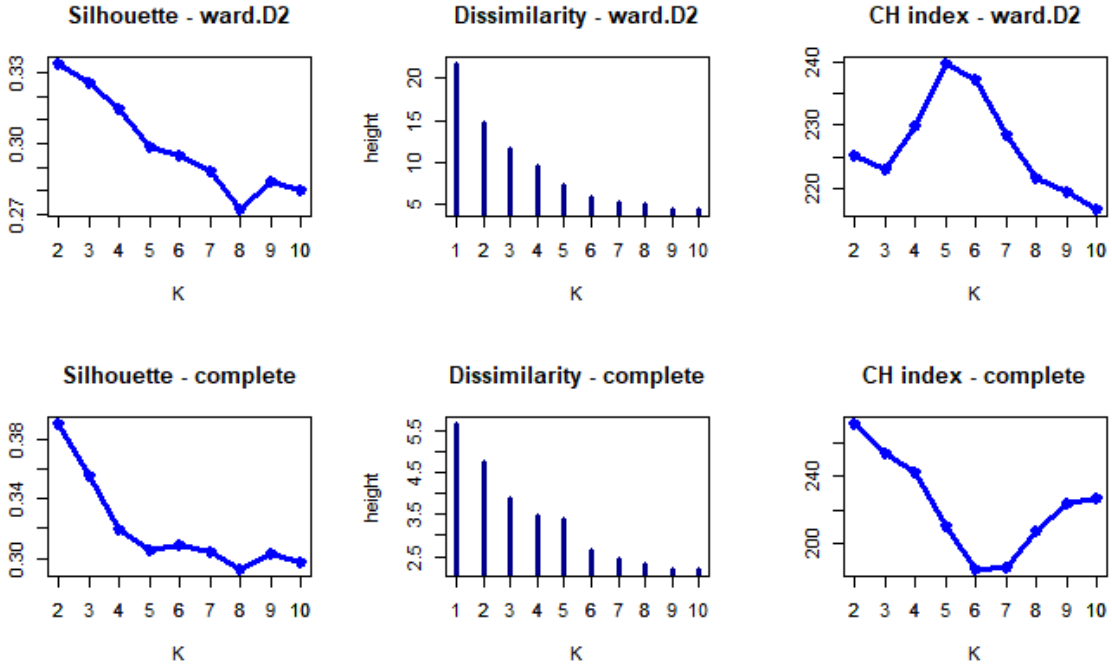


Figure 4.17 Dendrograms with different linkages on subset PC. Clusterization with K=6

Three measures are employed in order to assess the model's goodness, to identify the optimal number of clusters, and to choose the most meaningful linkage: silhouette index, CH index, and the dissimilarity level at which groups merge.

Figure 4.18 Subset PC Indexes for Ward and Complete Linkages



The Ward method always shows more balanced results in terms of clusters' sizes; on the other hand, the "complete" linkage seems to isolate significant groups of outliers. On the basis of the purpose of this analysis, which is to divide students along three dimensions, balanced clusters look like the best choice. Another important goal might be searching for highly-at-risk students, for which outlier detection may be the best route. In view of the above arguments, the "Ward" linkage method is selected. Subset SF shows the worst performance and, therefore, it is discarded.

Table 4.5 - Statistical Measures for Hierarchical Clustering

	Linkage	Optimal K	Silhouette	CH
Subset SF	Ward	5	0.298	220.04
	Complete	4	0.285	202.21
Subset PC	Ward	5	0.315	239.74
	Complete	5	0.306	210.73
Subset A	Ward	6	0.301	281.80
	Complete	5	0.279	257.62

Subset PC

From a number of K equal to 10, clusters organize around a central or average group. As K becomes smaller, these groups aggregate to each other. By looking at the evolution of K and the clusters' variance on MCAS, having 6 groups instead of 5 seems to provide better results for interpretation. The indexes are higher for K equal to 5 but they do not show a big difference when K is 6. In this second case, the groups have more or less equal sizes.

Figure 4.19 Ward Linkage on Subset PC

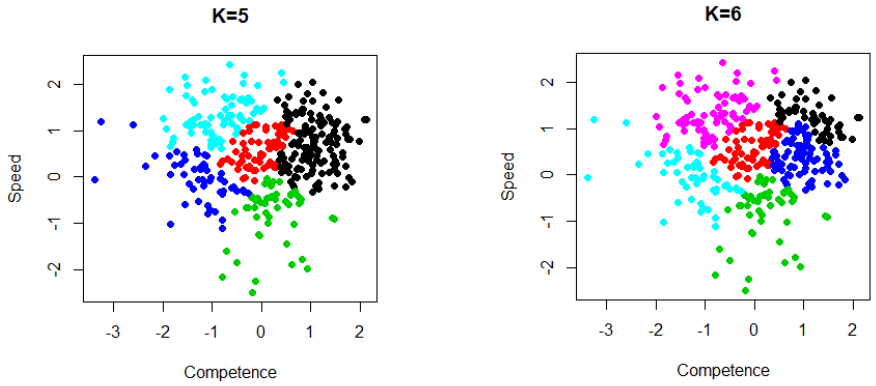


Table 4.6 - Clusters Average Values using Ward Linkage

Subset PC	1	2	3	4	5	6	Total
Size	51	67	51	85	52	76	382
MCAS	43.353	27.104	24.314	34.753	14.750	19.921	27.49
n_first_responses	489.294	365.612	186.353	319.212	238.654	529.724	363.2
st_ave_response_time	29.541	29.231	47.278	37.780	29.855	18.146	31.46
st_prob_help	0.094	0.207	0.149	0.095	0.352	0.330	0.204
fr_st_help	0.086	0.209	0.136	0.084	0.398	0.371	0.213
dev_attempts	-0.467	-0.033	-0.238	-0.462	0.487	0.378	-0.061
dev_time_complete	-10.350	-0.055	28.116	3.740	10.735	-12.124	2.24
dev_correct	0.280	0.133	0.180	0.260	0.009	0.023	0.148
dev_fr_corr	0.188	0.004	0.067	0.164	-0.145	-0.136	0.024
dev_fr_time	-3.737	0.487	19.558	6.058	2.851	-10.551	1.834

- **Cluster 1 and Cluster 4.** These groups contain high-achieving students. Cluster 1 shows the highest average score during homework, the lowest probability of a help request, and the lowest average number of attempts. Although, Cluster 1 shows the highest values, the results are very similar to the one of Cluster 4.

  The differences between the two clusters are evident when speed variables are considered. The former students are generally faster than the latter ones. This speed difference is reflected in the MCAS variable: Cluster 1 has the highest MCAS average score, with a limited variance and a minimum value of 27; on the other hand, Cluster 4 shows an average value almost 10 points behind Cluster 1 and a much higher variance. Although both groups show high correctness percentages, a small difference in the average response time may be responsible for more uncertainty about students' final results: in particular, slower students of Cluster 4 are more likely to perform worse on test day than faster students from the first group.

  The two clusters also differ in terms of practice on ASSISTments. Although it has already been said that this variable may be biased, this gap may be part of the reason why Cluster 4's pupils are slow on average.

  Members of Cluster 1 can be described as excellent students, with high assignment score and high speed: except for some students, they usually perform well on the final test.

  Cluster 4 also contains well-prepared students, whose results in the final test are more unpredictable. The variance of their final score could be a consequence of their slower response times, since MCAS is a time-limited test.

- **Cluster 2.** This is the central cluster shown in the PC scatterplot. Its average values always fall near the middle of the distributions. The performance on final test is highly unpredictable but still rotates around the middle score. Although they are usually faster than the average, these students typically spend more time evaluating the problems than high-achieving students, their score is lower and they have a higher probability of asking for help.

  This group includes average students, whose performance is neither excellent nor unacceptable. They show signs of motivation since they usually take their time to solve the problem. However, they may not be prepared enough to perform well during homework; the fact that they are more likely to access

help strategies may not be a sign of carelessness but rather lack of command on the skill being tested.

- **Cluster 3.** It is characterized by students with slow response times, typically more than 15 seconds longer than the school average. This peculiar feature is also found in other time-related variables. The number of attempts, as well as the correctness percentage variables, present high variances: therefore, the cluster includes both high-achieving and average students. This is supported by the high variance in MCAS, which is by far the highest among all groups. The probability of asking for help is quite moderate.

  The speed seems to be correlated with the practice level. In fact, this group shows the lowest level of familiarity with the platform, which may explain their slow performances. Moreover, the high value could be distorted by the presence of unidentified outliers.

  In conclusion, this cluster contains slow students which cannot be further identified with any other peculiar characteristic. In spite of their slowness, their performance on final test is highly variable.

- **Cluster 5.** It includes the lowest-achieving students in the whole school. Its performance measures are extremely low, and this is reflected on the results of the final test, whose variable has a very low mean as well as a restricted variance. Their response time is highly variable but never excessive. Students may try to answer the questions and solve the problems, but they seem to lack completely the skills and the knowledge needed to answer correctly.

  Their average number of attempts indicates that they lack a good preparation on the subject that they are tested on. They are also very likely to request help, and this probability is even higher in first attempts. They may act out of carelessness or just because they cannot solve the problems: not much can be said about students' motivation or carelessness.

  However, since this cluster includes both students with extremely fast response times and very high probability of asking for help, both careless and demotivated students (not well prepared) are included: their attitude towards the homework may damage their ultimate performance but still in the first place, they seem to perform badly. Therefore, the typical feature of this group can only be described as lack of preparation and knowledge.

- **Cluster 6.** It is characterized by very fast responses and overall high speed in solving the problems. Whether the answer is correct does not seem to be important to these students, since their average correctness percentage is below the average. They access help on average one third of the times. Their number of attempts is generally high. In spite of their supposed carelessness (given the average time they take to reflect on a question), some of these students surprisingly perform well on test day: the group average is higher than the one of the previous cluster with a maximum of 40 points.

  The results of the clustering algorithms show that even those who are identified as careless students during homework are able to succeed and get high scores in the final test. Therefore, the analysis seems to fail in accurately dividing at-risk students and careless students. This may be due to the inappropriateness of the variables employed, which all regards homework, or the inappropriateness of the algorithm.

  In both cases, the groups which most could benefit from the groups identification (unprepared and demotivated students) do not yield accurate results. Therefore, the analysis does not provide major advantages.

Subset A

In comparison to the previous subset, clusters show higher variance in MCAS scores.

The second major difference regards clusters' sizes which are also much more unbalanced: first, there is an outlier group in terms of speed (cluster 6), which contain all of the students with an average response time higher than 1 minute. Given the size and the high variances that this cluster presents, it is daring to say more about these students.

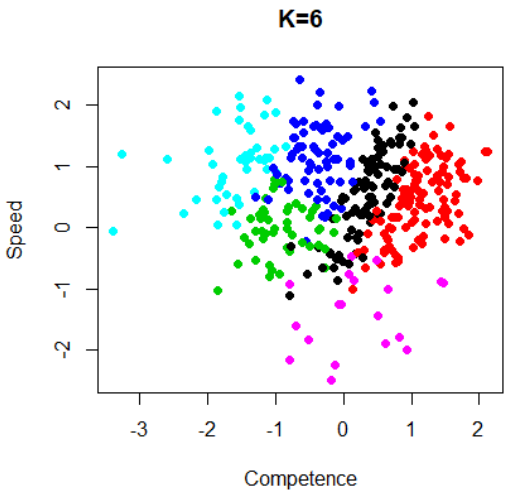


Figure 4.20 Ward linkage on Sub-

The fact that outliers are isolated may advantage the homogeneity of the other groups.

Cluster 4 and 5 identify students with extremely low scores in time-related variables, generally lower than the high-achieving group. However, there are differences between these two relatively careless groups of students. On the one hand, Cluster 4's results in terms of correctness are undoubtedly better than the ones of Cluster 5, as well as the number of attempts needed to solve a problem. These distinctive traits are perhaps reflected on the final test score, which is on average easier for these students. Cluster 5 may contain not prepared individuals who also lack self-confidence and motivation.

Other low-achieving students are included in Cluster 3. Unlike Cluster 4's members, students from this group consider the problem through and seem to prefer attempting the problem on their own instead of asking for a help request, even though their answers are often wrong, especially in first responses. Despite the demonstrated effort, their average performance on test day is well below the middle.

Cluster 1 represents the average students' group with a tendency of avoiding help requests. This suggests acceptable levels of self-confidence, together with lack of carelessness proven by a moderate average time considering problems.. In conclusion, Cluster 2 represents excellent students in terms of correctness, who are also very fast on average.

4.8 Model Selection

In terms of statistical measures whose value does not depend on the number of features, K-means models perform better than the hierarchical agglomerative ones.

Moreover, subset A always yields better results subset PC. The first model uses the variables that best represents the three aspects that the analysis aims at investigating (knowledge, speed, motivation). On the other hand, the second model employs principal components which contains more than 85 % of all of the information which has been retrieved from students' log actions during the exploratory analysis: for the purpose of this study, it is also important to evaluate the informative value of log-data.

With regard to the interpretation, K-means clustering also shows more robust results, according to the initial assumptions about students, than hierarchical clustering.

In view of these arguments, the non-hierarchical clustering results on subset A and subset PC are chosen as the most significant models of the analysis. These models are going to be implemented in the other schools, in order to prove whether similar groups can be identified.

Table 4.7 - Clustering Methods Comparison

K=6	Method	CH	Silhouette
Subset A	K-Means	315.07	0.328
	Ward	281.80	0.301
Subset PC	K-Means	271.02	0.335
	Ward	237.3666	0.295

4.9 Generalization

In order to prove the validity of the models and their conclusions, it is considered relevant to employ them in other contexts. If the algorithms correctly identify actual categories of students, then the same patterns should be evident in other samples of the same population.

Generalization is particularly difficult in the social sciences, such as the educational field, since there are social and psychological effects which can cause major discrepancies even within very limited environments.

In this particular study, the dataset that has been employed contains data from different schools located in the same city, thus allowing to control for all of the factors that affect students' performance at higher levels. In particular, some of these factors are related to education and are exactly the same for each of the school of the dataset, such as: syllabus, standardized final test, test evaluation methods, education system policy. Other such variables are economic or social factors as: public expenditure in education, town median income, official language, and mainstream culture.

The reference population is therefore the city's population of students and the samples are identified in the schools. Applying the selected models on the other

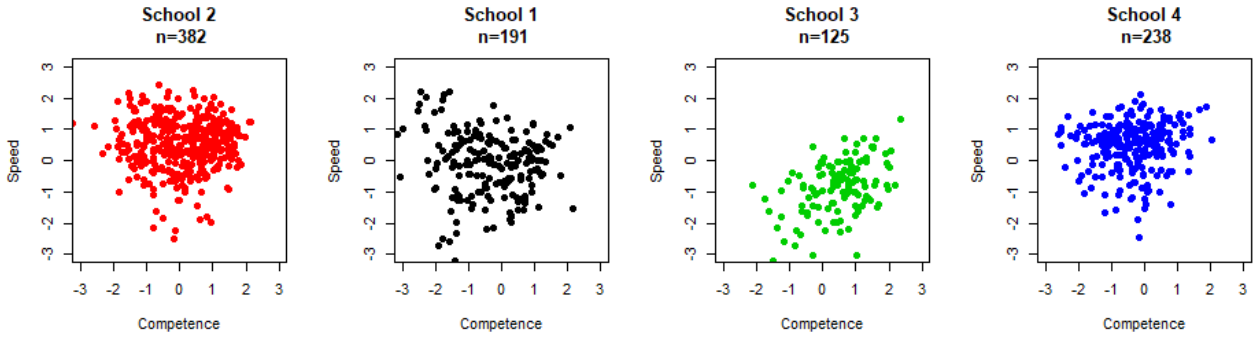
samples allows to prove the validity of the model, at least in the limited environment of a town. The models will pass the test if similar interpretations can be given to the clusterization on the other schools.

Before implementing the model, the differences among the reference school and the other samples are assessed: a two-sample t-test is carried out for each couple and the p-values state that in each case the samples are significantly different for almost all of the variables.

Nevertheless, School 4's scatterplot on PC shows a triangular-shaped distribution, similar to the one of School 2. While these two schools have similar distributions, School 1 and 3 do not show the same patterns. Given these results, the chosen models are expected to yield similar clusters when applied on School 4, while the interpretation for School 1 and 3 is likely to diverge.

It is important to point out that School 3 is characterized by a substantial lower practice level on the ASSISTments platform than the other schools: this is likely to bias the results, especially in terms of speed and time variables. Therefore, School 2 and School 3 cannot be compared, since most of the variables depend on the student's practice level.

Figure 4.21 Scatterplots on PC of Schools

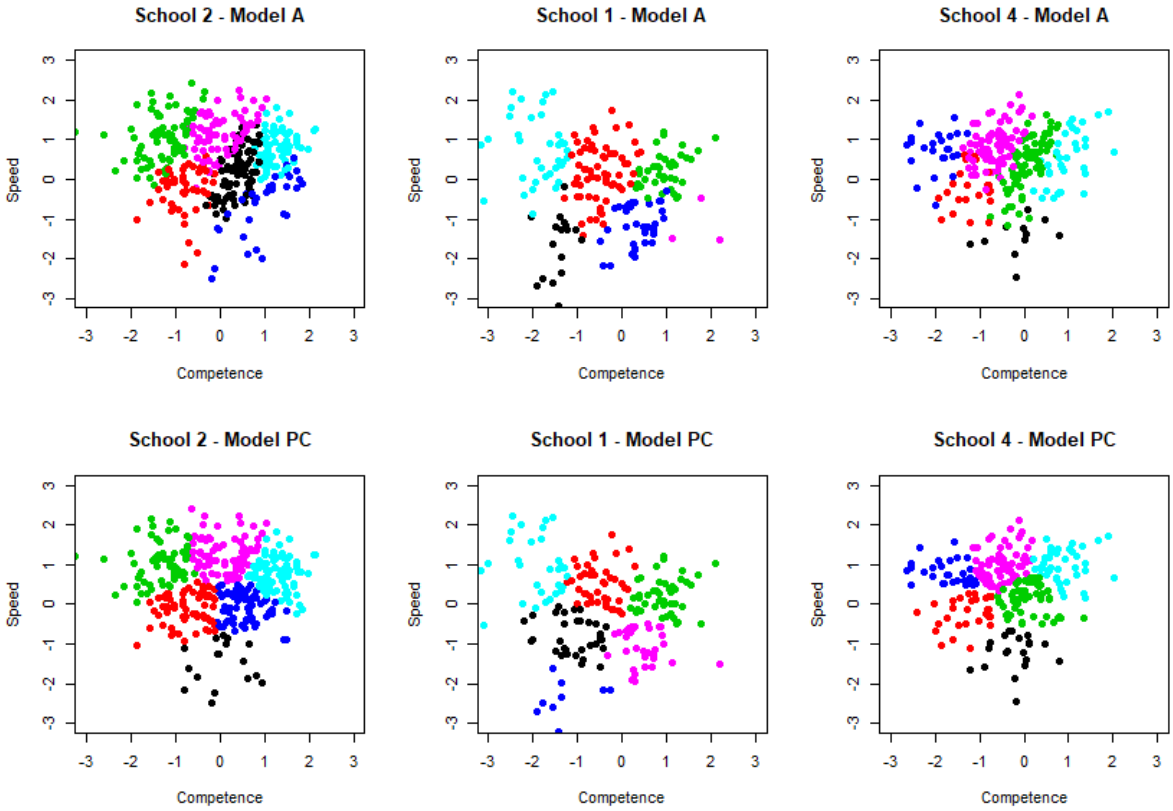


The validation samples that are used are School 1 and 4. When the models are applied to these schools' dataset, the statistical measures indicate that they perform well, even though the optimal number of clusters may vary. In order to compare the models, the number of clusters is arbitrarily set to 6.

Model A does not show the desired results: although students are still divided along the principal components (even though this model employs only three features), the clusters hardly follow the pattern traced by the reference school (School 2). Although the interpretation of these clusters may be interesting, the

purpose of this validation analysis is to assess whether the supposed categories of students identified in School 2 are recognizable in other samples of the same city.

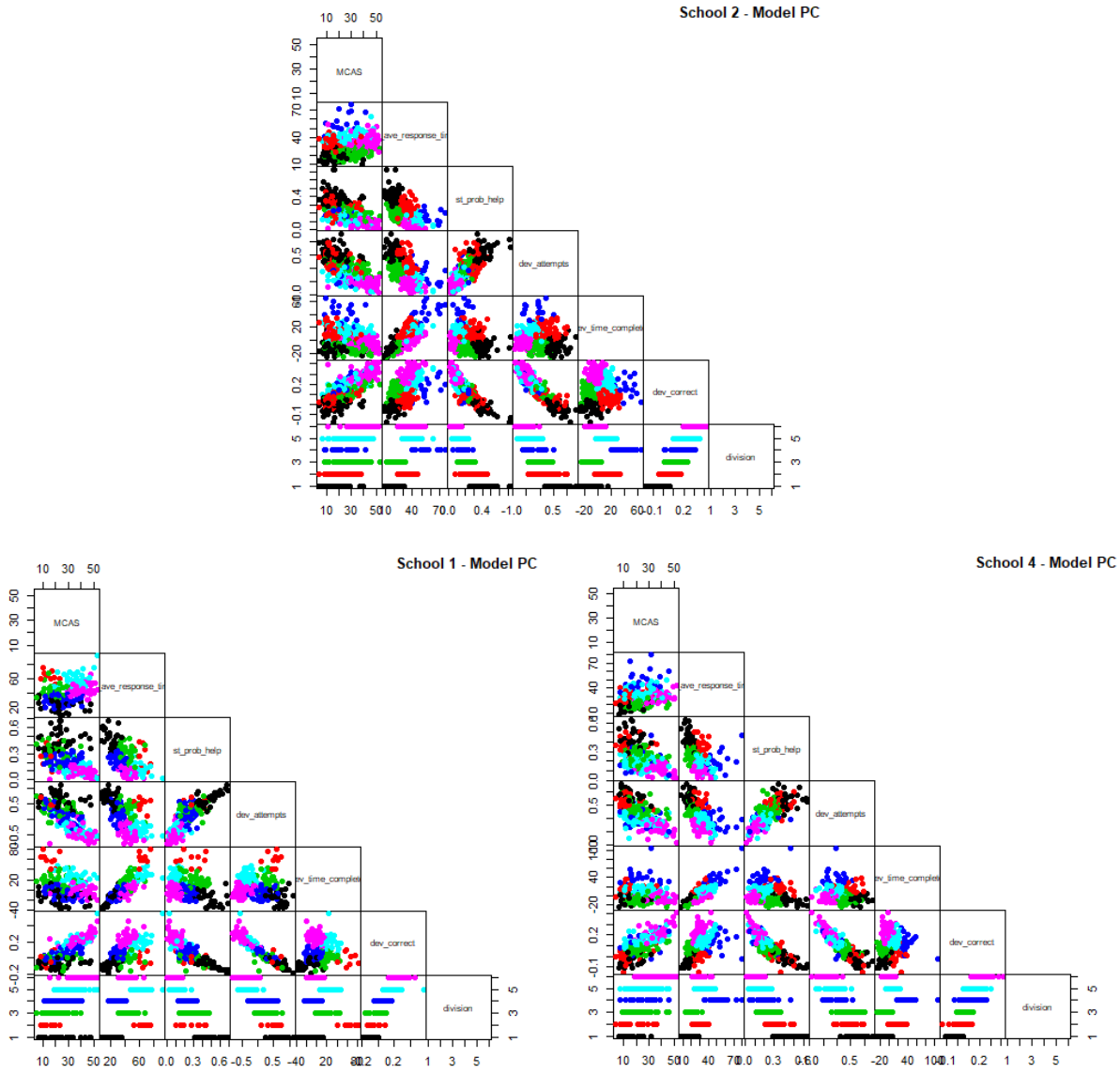
Figure 4.22 K-Means on different subsets in different schools with K equal to 6



On the other hand, some similarity in the patterns is more evident when model PC is employed: in each sample, the speed outliers' group is isolated and the remaining observation are divided into two average-speed and three high-speed groups. The likeness is not only visible on PC, but clusters also show similarities in terms of features' means and variances. The clusters' colours have been changed in order to improve readability: when the groups are ordered by the average correctness percentage, it is clear that the interpretation of clusters from School 4 is similar to the ones of the reference school.

This is not the case for School 1, in which the outliers in terms of speed are not as far from the other groups as in the other schools: in particular, students from this cluster seem less prepared so that they are also the lowest-achieving group in the final test. However, the clusters on the extremes of the distribution can be described as their equivalents in School 2.

Figure 4.23 Scatter Matrix of Model PC in different schools

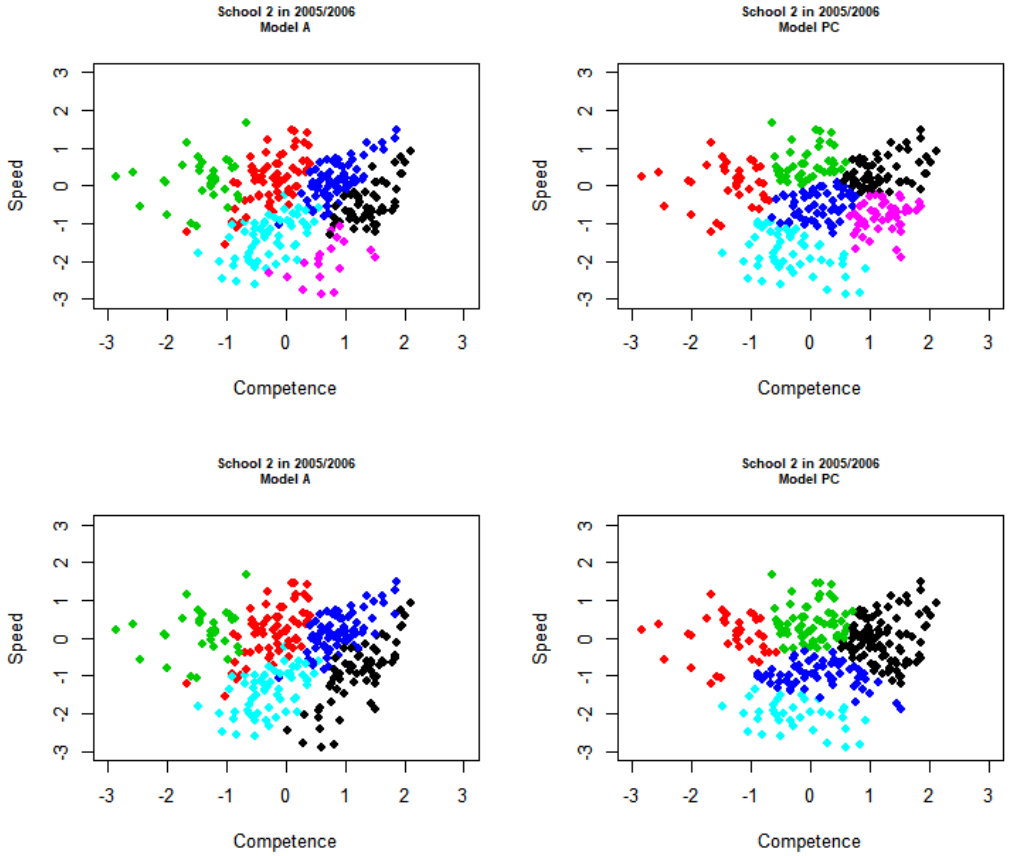


The dataset that was employed allows to perform another type of generalization test, which regards the temporal dimension. Until now, all of the observations were from the 2004-2005 school year: in this second test, the log-actions of School 2's students performed in 2005-2006 are employed. The distributions show that this subset has a tendency towards higher response times and lower speed in comparison to the previous school year. As a consequence, a distinctive group of highly slow students cannot be identified, since observations are more uniformly distributed, and therefore, the subset looks more similar to School 1.

Neither of the models show the same patterns found in the same school using the previous year's data. In the previous analyses, the number of clusters was set equal to 6 because one of the clusters would contain the outliers: since in this case

there are no such observations to be found, the models are also evaluated for K equal to 5 but, even in this case, the models do not yield the expected results.

Figure 4.24 Models A and PC on School 2 in 2005-2006 with K equal to 6 and 5



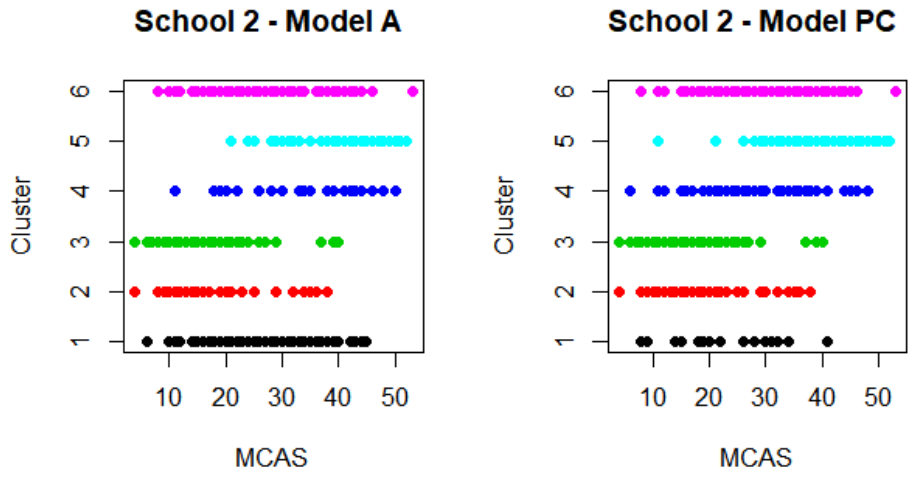
As it was expected, the models that were studied on School 2's students are hardly valid for other schools or in other school years. Clustering algorithms are not predictive models, but rather descriptive, and the results are heavily based on the single instances that make up the sample. However, an attempt has been made in order to prove that generalization of these models is not possible. Nevertheless, School 4 has shown a surprising resemblance to the reference school. The dataset does not contain enough information about the schools so that it is not possible to find out what causes the similarity between these two schools. Possible factors may be a similar approach to the use of the ASSISTments platform, but also localization or students' families median income.

Although the results may be different, the chosen models offer useful and readily understandable information to teachers about their students' approach towards problems.

Using these insights about students can help them improve throughout the school year and enhance their probability of succeeding in the final test. A label can be assigned to each of these categories so that teachers have an immediate sense of a students' performance in the homework assignments.

However, although there is a slightly positive correlation between assignment performance and final performance, categories are not good predictors for students' final test scores, given the high variance each of them has on the MCAS variable. Therefore, the clusters should not be used as such, but rather as the description of a student's typical behaviour towards assignments.

Figure 4.25 Models A and PC's clusters against MCAS scores in School 2



4.10 Conclusions

The analysis about the identification of the clusters has produced different models that yield different results. The models themselves yield different clusters when hyperparameters are changed. All clustering algorithms' results strongly depend on the data they are implemented on: moreover, the K-Means method needs a random component for the initialization of the algorithm. Therefore, the models that have been chosen are bound to change when its parameters or some of the observations vary.

Nevertheless, all of the clustering algorithms suggest that the optimal number of clusters is close to 6. In addition to that, the attempt at generalizing the results of the clusterization led to conclude that a similar composition of groups of students can be found in schools that are close to the one used to choose the best models.

These cluster analyses have not assumed that a priori students' groups actually exist, but rather the goal has been to provide teachers with a reliable representation of their school or class, divided into homogeneous clusters.

Although the models that have been created are descriptive and they are built on the average results obtained by each student by the end of the year, the most advantageous purpose of these models is their implementation throughout a school year, so that teachers have access to easy-to-understand information and evidence about their students' performance and behaviour.

To prove how these models may work, random students (within a subset of students with high numbers of actions) from school year 2005-2006, which were not employed in the implementation of the chosen models, are selected. For each of them, the cumulative average of the variables of interest is computed, so as to have information about their pathway during the whole year. Then, for each of their actions, the respective scores on the principal components are predicted: the variables are not scaled based only their values but rather based on the whole student dataset. Students' pathways are plotted in the PC space: as it can be seen, each student changes their behaviour throughout the school year. In each point, the students' variables depend on the previous actions because of the cumulative means. These types of plots provide teachers with an additional tool to supervise students' progress.

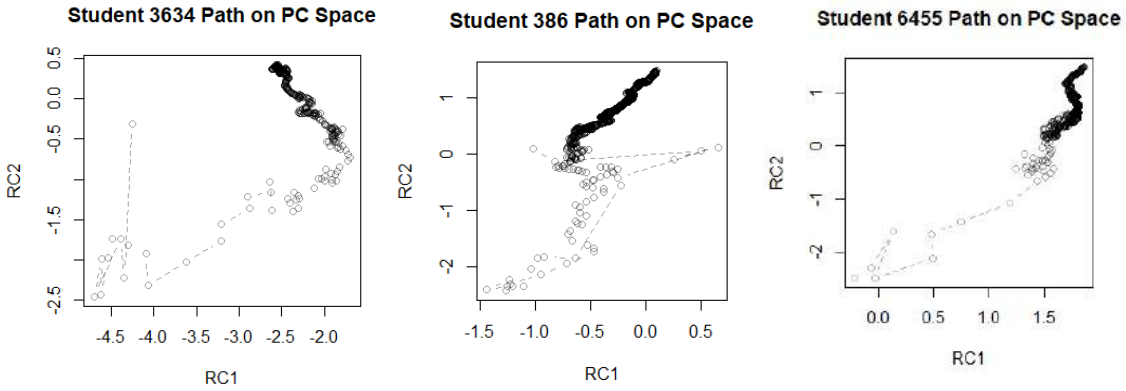
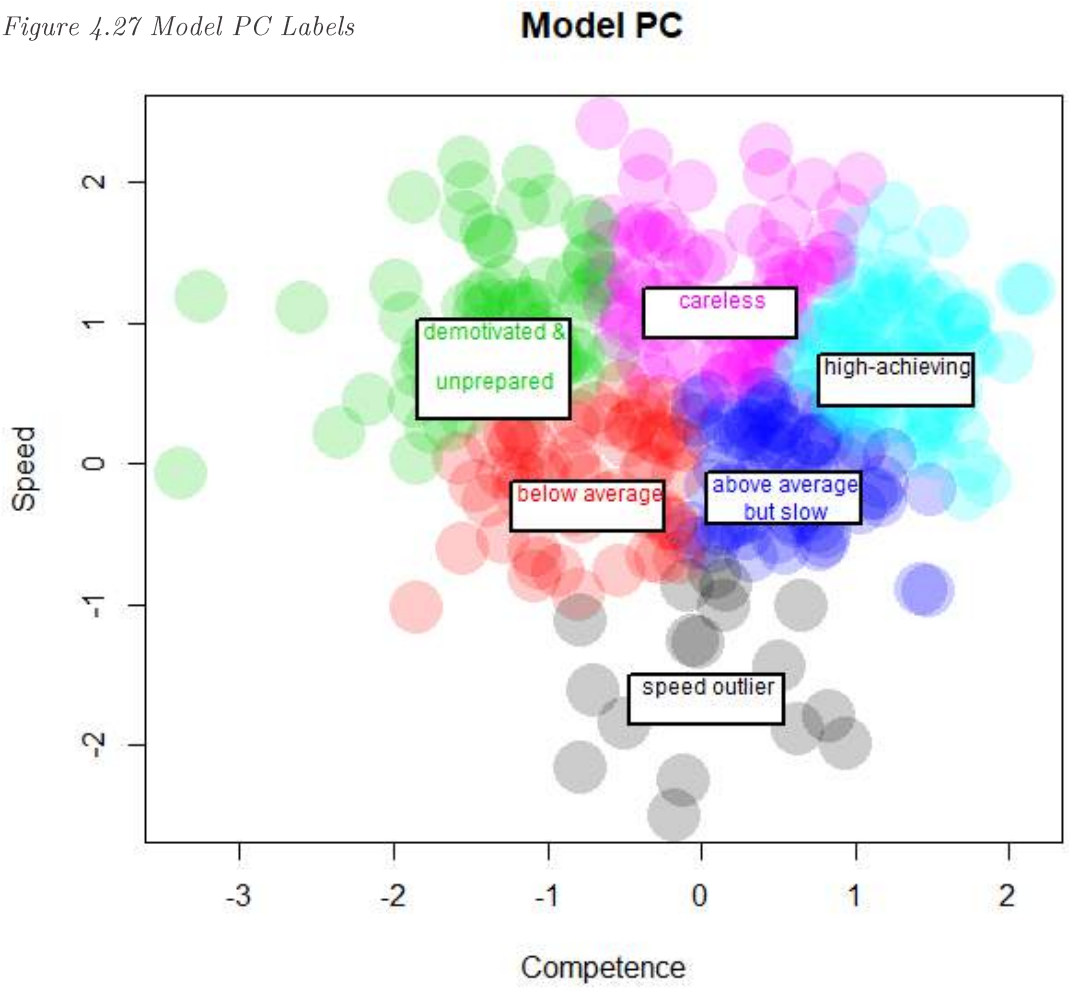


Figure 4.26 Random Students' Pathways

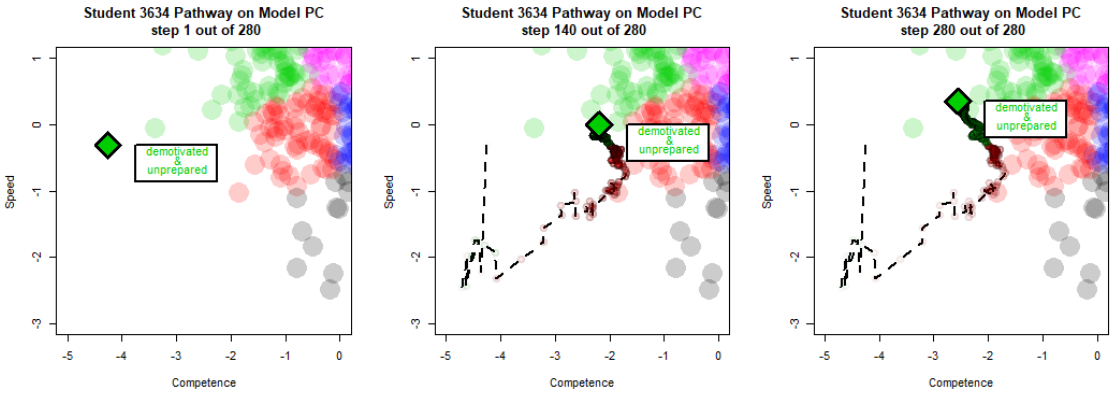
With respect to the model, model PC with a number of clusters equal to 6 is chosen: labels are assigned to each of the clusters ("speed outlier", "below average", "demotivated & unprepared", "above average but slow", "high-achieving", "careless").

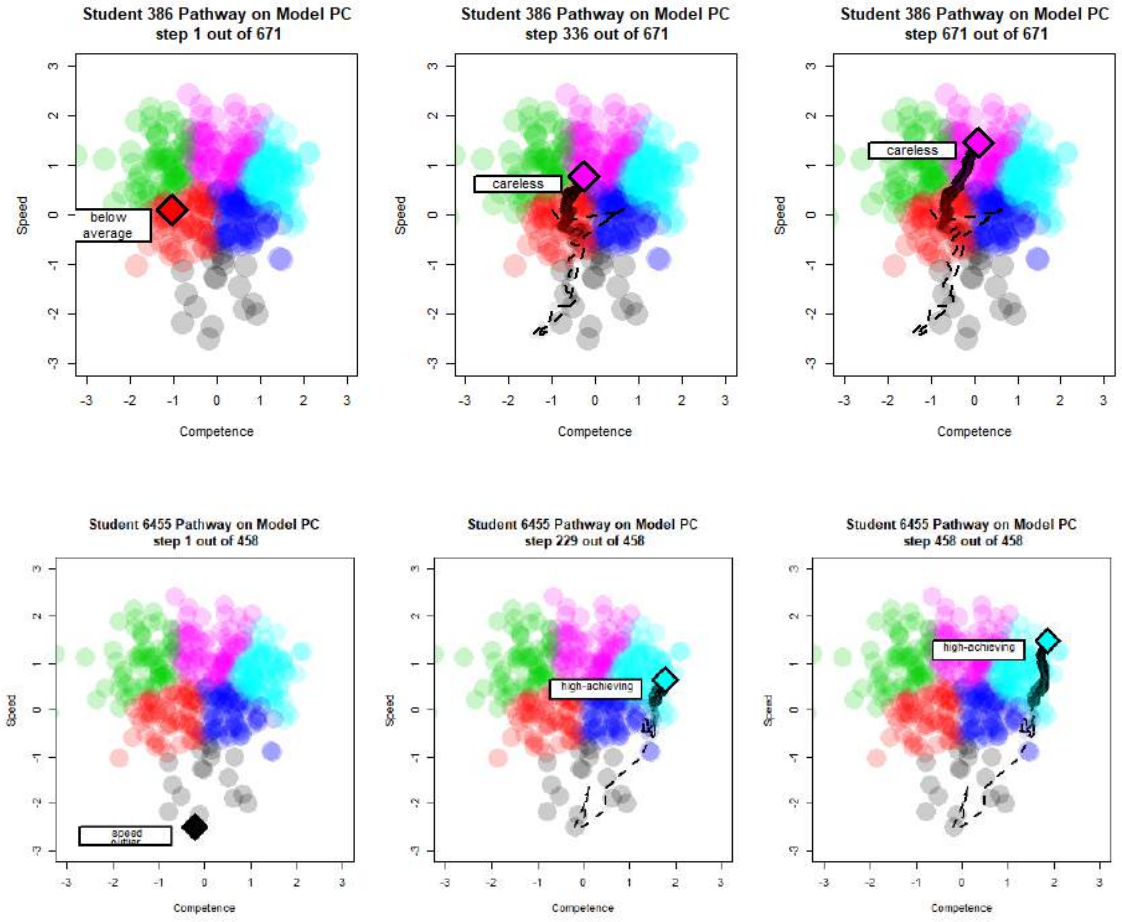
Figure 4.27 Model PC Labels



In order to assign labels to students in each point in time (problems), the K-Nearest Neighbours method is used: the modal category of the subset of the K subjects that are closer to the new data point is assigned as the label of the point itself: the number of K neighbours is arbitrarily set to 10.

Figure 4.28 Students' Labelling over Time





For instance, after the first few actions, the first student (seems to perform worse as the year goes by, since its competence level constantly decreases, while its speed level increases: the label assigned to student 3634 is "demotivated and unprepared", which is consistent with its final score of just 5 points.

On the other hand, the second student (386) shows an improvement in terms of response time but not in competence. Its final label is "careless" because of its high speed. This category is characterized by a high variance on MCAS: this student's final score is 22, which is also consistent with an average results.

Finally, the last student (6455) improves both in competence and speed with astonishingly high scores on PC variables. Its category can only be "high-achieving", which is most definitely supported by its MCAS score of 51.

Once again, these clustering and visualization methods prove their usefulness in helping teacher identifying in advance students' needs and weaknesses.

5 References

[1] Baker, R. (2010) Data Mining for Education. In *International Encyclopaedia of education*, vol. 7, pp. 112-118.

[2] Journal of Educational Data Mining. Editor: Olney, A. https://jedm.educationaldatamining.org

[3] Baker R., Siemens G., (2014) Educational Data Mining and Learning Analytics. In *The Cambridge Handbook of Learning Sciences*, pp 253-272.

[4] International Society of Educational Data Mining: http://educationaldatamining.org/

[5] Society of Learning Analytics Research: https://www.solaresearch.org/

[6] Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010) A Data Repository for the EDM community: The PSLC DataShop. → http://pslcdatashop.org

[7] U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. https://nces.ed.gov/

[8] Baker, R., & Siemens, G. (2012) Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In *LAK 2012: Second International Conference on Learning Analytics and Knowledge*, Vancouver, British Columbia, Canada.

[9] Castro, F., Vellido, A., Nebot, A. and Mugica, F. (2007) Applying Data Mining Techniques to e-Learning Problems. In *Evolution of teaching and learning paradigms in intelligent environment*, pp. 183-221, Springer.

[10] Ostrow, K. S. & Heffernan, N. T., (2014) Testing the Multimedia Principle in the Real World: A Comparison of Video vs. Text Feedback in Authentic Middle School Math Assignments". In Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M. (eds.) *Proceedings of the 7th International Conference on Educational Data Mining*. pp. 296-299.

[11] San Pedro, M., Rodrigo, M., Baker, R. (2011) The Relationship between Carelessness and Affect in a Cognitive Tutor. In *Proceedings of 15th International Conference on Artificial Intelligence in Education.*

[12] Hook, Porter, Herzog, (2018) Dimensions: Building Context for Search and Evaluation. In *Frontiers in Research Metrics and Analytics vol. 3.* https://www.dimensions.ai/

[13] Coelho, O.B., Silveira, I.F. (2017) Deep Learning applied to Learning Analytics and Educational Data Mining: A Systematic Literature Review. In *VI Congresso Brasileiro de Informática na Educação (CBIE 2017)*

[14] Gross, E., Wshah, S., Simmons, I., & Skinner, G. (2015) A handwriting recognition system for the classroom. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pp. 218-222

[15] Heffernan, N. & Heffernan, C. (2014). The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. In *International Journal of Artificial Intelligence in Education.* 24(4), 470-497.

[16] WPI News, U.S. Department of Education Awards Worcester Polytechnic Institute $8 Million to Scale and Expand ASSISTments, an Online Learning Tool Proven Effective at Improving Middle School Math Scores (Oct. 2019)

[17] Problem Set PSABD4E5 on ASSISTments: https://app.assistments.org/FA/v/PSABD4E5

[18] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

a. Package "car". John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage.

b. Package "class". Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York.

c. Package "cluster". Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2018). cluster: Cluster Analysis Basics and Extensions.

d. Package "dummies". Christopher Brown (2012). dummies: Create dummy/indicator variables flexibly and efficiently.

e. Package "ggplot2". H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

f. Package "gridExtra". Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics.

g. Package "leaps". Thomas Lumley based on Fortran code by Alan Miller (2017). leaps: Regression Subset Selection. R package version 3.0.

h. Package "lme4". Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.

i. Package "lmerTest". Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." Journal of Statistical Software, *82*(13), 1-26.

j. Package "manipulate". JJ Allaire (2014). manipulate: Interactive Plots for RStudio

k. Package "moments". Lukasz Komsta and Frederick Novomestky (2015). moments: Moments, cumulants, skewness, kurtosis and related tests.

l. Package "NbClust". Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36.

m. Package "performance". Daniel Lüdecke, Dominique Makowski and Philip Waggoner (2020). performance: Assessment of Regression Models Performance. R package version 0.4.4.

n. Package "plotrix". Lemon, J. (2006) Plotrix: a package in the red light district of R. R-News, 6(4): 8-12.

o. Package "plyr". Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29.

p. Package "psych". Revelle, W. (2019) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA.

q. Package "rgl". Daniel Adler, Duncan Murdoch and others (2019). rgl: 3D Visualization Using OpenGL.

r. Package "scales". Hadley Wickham (2018). scales: Scale Functions for Visualization.

s. Package "tidyr". Hadley Wickham and Lionel Henry (2019). tidyr: Tidy Messy Data. R package version 1.0.0.

[19] RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

[20] Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. In *Proceedings of the Royal Society of London Series* A 160, 268–-282.

[21] Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. In *Biometrika*. 34 (1–2): 28–35

[22] Snijders, T. & Bosker, R. (2012). Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling (2^{nd} Edition). Ed. SAGE Publications.

[23] Berger, Fisher. (2013) A Well-Educated Workforce is Key to State Prosperity. Economic Policy Institute. *Economic Policy Institute*, 22(1), 1-14

[24] G. James, D. Witten, T. Hastie, R. Tibshirani. (2013) An Introduction to Statistical Learning: with Applications in R. New York, Springer.

[25] ASSISTments Data Mining Competition 2017.

[26] Pardos, Baker, R., San Pedro, Gowda, Gowda. Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning. Outcomes. In *Journal of Learning Analytics*, 1(1), pp. 107–128.

[27] Casella, Berger. Statistical Inference. Belmont, CA: Duxbury, 2002

[28] Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. In *Psychometrika*, 23, pp. 187-200. 10.1007/BF02289233.

[29] Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. In *Applied Statistics*, 28, pp. 100–-108. 10.2307/2346830.

[30] Calinski, T., Harabasz, J. (1974). A dendrite method for cluster analysis. In *Communications in Statistics - Theory and Methods*, 3(1), pp. 1-27

[31] Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. In *Applied Mathematics*, 20, pp. 53--65.