



HAL
open science

Maximizing students' engagement through effort-based recommendations

Barbara Moissa

► **To cite this version:**

Barbara Moissa. Maximizing students' engagement through effort-based recommendations. Computer Science [cs]. Université de Lorraine, 2021. English. NNT : 2021LORR0230 . tel-03564001

HAL Id: tel-03564001

<https://hal.univ-lorraine.fr/tel-03564001>

Submitted on 10 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : ddoc-theses-contact@univ-lorraine.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4

Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

http://www.cfcopies.com/V2/leg/leg_droi.php

<http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm>



Maximizing students' engagement through effort-based recommendations

THÈSE

présentée et soutenue publiquement le 06 décembre 2021

pour l'obtention du

Doctorat de l'Université de Lorraine
(mention informatique)

par

Barbara Moissa

Composition du jury

Président : Davy Monticolo

Rapporteurs : Agathe Merceron
Julien Broisin

Examineurs : Leandro Krug Wives

Encadrants : Anne Boyer
Geoffray Bonnin

Mis en page avec la classe thesul.

Remerciements

I would like to thank several people without who this thesis would not be possible. My family and friends for all their support, advice, and encouragement. In special, I would like to thank my husband, my biggest supporter during this entire PhD, and the person who went out of his way to make my days better.

My supervisors, Anne Boyer and Geoffray Bonnin, for sharing their time, experience and knowledge with me; for all of the patience, enthusiasm, and research passion; and also for the opportunity to pursue this PhD in France and all of the help they gave me while settling in a new country.

The members of the jury – Dr. Davy Monticolo, Dr. Agathe Merceron, Dr. Julien Broisin, and Dr. Leandro Wives – for accepting our invitation and their contributions to this work.

All the members of the KIWI/BIRD team (LORIA) and the ERPI team (ENSGSI) for the knowledge, discussions and good moments shared. You sure made the journey easier and more joyful.

Finally, the Ministère de l'Éducation Nationale, de l'Enseignement Supérieur et de la Recherche for funding this thesis.

To my family.

Contents

Chapter 1	
Introduction	1
1.1 General context	1
1.2 Problem definition	3
1.3 Contributions and thesis structure	5
1.4 Scientific publications	6
1.5 Important note	7
Chapter 2	
Recommendation systems overview	9
2.1 Recommendation systems	9
2.1.1 Formal definition	10
2.1.2 Filtering techniques	10
2.2 Educational recommendation systems	13
2.2.1 Need for personalization: one size does not fit all	13
2.2.2 Multiple stakeholders' perspectives	14
2.2.3 Good recommendations to foster learning	15
2.2.4 Need of explanations	16
2.2.5 Ethical challenges	16
2.2.6 System acceptance	17
2.2.7 Evaluation challenges	17
2.3 Chapter conclusion	18
Chapter 3	
From the theory of commitment to the cognitive load theory	21
3.1 The theory of commitment	21
3.2 Students' engagement vs. students' effort	23
3.3 Cognitive load theory	26

3.3.1	Cognitive load measurement	27
3.3.2	Cognitive load in educational scenarios	34
3.4	Chapter conclusion	36

Chapter 4	
Multimodal data collection and analysis	39

4.1	Preliminary study	39
4.2	Multimodal data collection	41
4.2.1	Experimental protocol	41
4.2.2	Data collection	43
4.2.3	Initial analysis	45
4.3	Ethical disclaimer	52
4.4	Chapter conclusion	53

Chapter 5	
Measuring and predicting students' perceived effort	55

5.1	Measuring students' perceived effort	56
5.1.1	Related work	56
5.1.2	Comparison with the state-of-the-art	58
5.1.3	Added-value of combining the session data	64
5.2	Predicting students' perceived effort	69
5.2.1	Related Work	69
5.2.2	Methodology	71
5.2.3	Results	71
5.3	Chapter conclusion	72

Chapter 6	
Towards engaging recommendations: formalizing the foot-in-the-door technique	75

6.1	Recommendation models	75
6.1.1	Foot-in-the-door functions	76
6.1.2	Recommendation algorithms	76
6.1.3	Selection criteria	80
6.2	Evaluation protocol	82
6.3	Results	85
6.3.1	Students' effort	86
6.3.2	Students' compliance	87

6.3.3	Students' performance	90
6.3.4	Students' engagement	93
6.4	Limitations	95
6.5	Chapter conclusion	96

Chapter 7

Conclusions and perspectives	97
-------------------------------------	-----------

7.1	Research perspectives	99
-----	---------------------------------	----

Chapter 8

Thesis summary in French / Résumé de thèse	101
---	------------

8.1	Contexte général	101
8.2	Définition du problème	103
8.3	Contributions	105
8.4	Publications scientifiques	107
8.5	Note importante	107

Bibliography	109
---------------------	------------

Appendix A

Proposed exercises	125
---------------------------	------------

A.1	Exercice 1	125
A.2	Exercice 2	125
A.3	Exercice 3	126
A.4	Exercice 4	126
A.5	Exercice 5	127
A.6	Exercice 6	127
A.7	Exercice 7	128
A.8	Exercice 8	128
A.9	Exercice 9	128
A.10	Exercice 10	129
A.11	Exercice 11	129
A.12	Exercice 12	130
A.13	Exercice 13	130
A.14	Exercice 14	130
A.15	Exercice 15	131

Appendix B	
Informed consent form	133
B.1 Original version in French	133
B.2 Translation to English	134
Appendix C	
Questionnaires	137
C.1 Questionnaire 1: Original version in French	137
C.2 Questionnaire 1: English translation	138
C.3 Questionnaire 2: Original version in French	138
C.4 Questionnaire 2: English translation	139
C.5 Questionnaire 3: Original version in French	139
C.6 Questionnaire 3: English translation	140
C.7 Questionnaire 4: Original version in French	140
C.8 Questionnaire 4: English translation	141
Appendix D	
Participation Certificate	143
D.1 Original version in French	143
D.2 English translation	144
Appendix E	
Effort indicators	147
Appendix F	
Correlations	151
Appendix G	
Effort features	155
Appendix H	
Engagement features	157
Appendix I	
Full results of the effort measurement using the effort features	163
I.1 Classification results	163
I.2 Ordinal regression results	164
I.3 Regression results	165

Appendix J**Full results of the effort measurement using the engagement features** **167**

- J.1 Classification results 167
- J.2 Ordinal regression results 168
- J.3 Regression results 169

Appendix K**Full results of the effort prediction using the engagement features** **171**

- K.1 Classification results 171
- K.2 Ordinal regression results 172
- K.3 Regression results 173

Appendix L**Full results of the recommendation models performance** **175****Appendix M****Full comparison results of the recommendation models that formalize the foot-in-the-door technique** **181**

List of Figures

3.1	Human cognitive architecture	26
3.2	Cognitive load over time (adapted from [30])	28
3.3	Cognitive load vs. mental effort and performance (adapted from [30])	30
3.4	Scenarios based on the relationship between effort and grades	31
4.1	Phases of data collection session	42
4.2	Distribution of the number of exercises solved by each student	46
4.3	Distribution of the session time (in minutes)	47
4.4	Mean time solving an exercise vs. session time (all in minutes)	48
4.5	Probability of having a good grade with a given effort level	50
4.6	Grade vs. perceived effort	52
5.1	Effort ratings distribution	59
5.2	Machine learning pipeline	61
5.3	Machine learning approach adopted in our approach to measure the students' effort with engagement features	66

List of Figures

Chapter 1

Introduction

This thesis is part of the PIA e-Fran METAL project, a learning analytics project to teach foreign languages to secondary school students via several personalized digital tools. More specifically, this thesis is part of Item 1.3 whose goal is to reduce the learners' failure rate by providing a dashboard for the teachers. This dashboard shows:

- The learners global activity with a special focus on their effort while executing the proposed educational tasks; and
- A personalized view of learner profiles that includes engaging recommendations of educational tasks with the aim of helping the teacher in his/her new role as a coach for the learners.

1.1 General context

The use of data is a growing phenomenon in many fields such as political science, medicine, economics, physics, social science, etc. [44]. This phenomenon is also present in the learning science, where it holds the promise of advancing our understanding and improving the learning process [61]. From this promise emerged the learning analysis research field, defined as “the measurement, collection, analysis and reporting of data about students and their contexts, for the purpose of understanding and optimizing learning and the environments in which it occurs” [158].

According to Suthers et al. [164] and also to several other researchers [159, 11, 10, 176, 85, 164, 67, 61], learning analytics is about analytics, but mostly about learning. Therefore, learning analytics should ideally take advantage of the available technology to collect and process the collected data while respecting educational theories.

Learning analytics is about analytics

On the analytical side of learning analytics, where this thesis is inserted (i.e., computer science), the data are collected and processed in order to generate useful information for the stakeholders [161]. Among the wide variety of methods and data applications found in the literature, we can find the recommendation systems [148]. These systems are defined as “software tools and

techniques that provide suggestions for items that are most likely of interest to a specific user” [144].

In the educational context, the main objective of such systems is to support the learning process. In order to achieve this goal, the recommendations must be adapted to each student and/or group in order to meet their needs. For example, Fotopoulou et al. [55] propose a recommendation system that recommends activities to help the teacher to improve the social and emotional skills of their students, while Pineda et al. [12, 13] propose a recommendation system that recommends learning resources in order to help students to overcome their difficulties in a given topic.

The recommended items, the target audience and the objectives of educational recommendation systems are diverse, as well as their application contexts [172]. However, to the best of our knowledge, educational recommendations systems have not explicitly aimed at engaging students yet. For this reason, respecting the context of the PIA e-Fran METAL project, we propose in this thesis an educational recommendation system whose objective is to engage secondary school students through recommendations made in the context of learning a foreign language.

Learning analytics is about learning

From a pedagogical point of view, the students’ engagement is essential to take advantage of what the school has to offer to acquire the necessary skills to succeed [56]. In other words, engagement is considered a factor that positively influences learning outcomes because it can increase the students’ learning performance, increase their maturity and reduce the tendency to drop out [160, 57].

Due to the benefits attributed to this construct, it has become a much studied aspect of the learning process. However, there is no consensus on its definition and several similar definitions related to desired behaviors have been adopted, such as commitment. In the social psychology literature, commitment is defined as the “binding of the individual to behavioral acts” [94]. The theory associated with this concept – the theory of commitment [93] – argues that it is not our beliefs and convictions that commit and/or engage us, but our actions. That is, people who act, become committed.

One of the various reasons for a person to commit is the free will compliance. This paradigm can be defined as the study of influencing procedures, allowing someone to freely do what you want him to do [87]. Several studies carried out under this paradigm show that it is possible to influence someone to change their beliefs, choices and behaviors without resorting to argumentation, rewards or punishment. This characteristic makes the paradigm interesting for pedagogical use [87]. More specifically, Joule and Almeida [87] suggest five free will compliance techniques – out of more than 107 techniques found in the literature [137, 40] – to use as a pedagogical tool. Among them is the foot-in-the-door technique, which consists in making consecutive requests with an increasing cost [59]. This technique seems particularly relevant for educational purposes because it is compatible with the the zone of proximal development [177], which states that activities should increase the challenge little by little.

In this thesis we propose to use the foot-in-the-door technique to influence students to carry out their activities. When considering the foot-in-door technique and the theory of commitment together, we expect that the use of the foot-in-the-door technique will influence the students to carry out activities (actions) and, as a consequence, to become committed (engaged).

Learning meets analytics

From a technological point of view, we are interested in recommendation systems as they recommend the best activities for students and can be used to improve many aspects of the learning process, including the students' engagement.

From an educational point of view, we are interested in the foot-in-the-door technique because we believe it can be used to engage the students and, as a result, improve their learning outcomes.

Therefore, by uniting both points of view, we propose a recommendation system that exploits the foot-in-the-door technique to choose the best activities to recommend. To do this, we define the cost of each activity as the students' effort because each task requires a different level of effort to be completed. Furthermore, effort has also been cited as a key factor of learners' success [28, 64, 169, 115] and, therefore, seems suitable to identify which learning tasks are more demanding for a given student and also to foster learning.

1.2 Problem definition

Given the elements previously discussed, our research question can be defined as: **To what extent can the formalization of the foot-in-the-door technique in a recommendation system influence the students' effort, compliance, performance, and engagement?**

To answer this question, we seek to accept (or reject) the following hypothesis: H_1 – *Formalizing the foot-in-the-door into a recommendation system can improve the students' effort, compliance, performance and engagement.*

According to this hypothesis, we have four evaluation criteria to consider:

1. **Effort:** The formalization of the foot-in-the-door proposed in this thesis is based on the effort required to perform each task. Given that the application of the technique implies a gradual effort increase, we expect to see higher effort levels.
2. **Compliance rate:** As already mentioned, the foot-in-the-door technique can increase the compliance rate of the requests that are made. Therefore, we assume that by using the technique to make recommendations, we can increase the acceptance rate of the recommended learning activities and, as a consequence, influence them to do more activities.
3. **Performance:** Given that the application of the technique implies a gradual effort increase and that the effort is related to better learning outcomes, we further hypothesize that the use of this technique can improve the students' performance on the exercises.
4. **Engagement:** According to the theory of commitment, an individual who acts becomes engaged. Since we intend to exploit the foot-in-the-door technique to influence students to do more learning activities, we also expect them to become engaged.

The evaluation of our hypothesis requires, at the very least, a dataset adapted to our objective and to the context of the PIA e-Fran METAL project, and the implementation of a recommendation system that applies the foot-in-the-door technique. This scenario also raises some specific questions:

- *What is effort? What are the differences between effort, engagement and cognitive load?*

The students' effort and their engagement are constructs associated with desired behaviors. Both constructs are often addressed without a definition [39, 72, 140] and, when a definition

is proposed, they often overlap [57, 72, 28, 152]. In addition, we can find in the literature another concept, the cognitive load, that is often referred to as being equivalent to the mental effort [129, 105].

Therefore, defining and distinguishing these three constructs from one another is an important step to answer the next specific questions raised.

- *How to estimate the effort that students have exerted in previous learning activities?*

One of the first requirements for implementing the foot-in-the-door technique is to define the cost of the tasks to propose them in the proper order. As previously defined, in this thesis we consider the amount of effort the students needed to exert to execute a given task as the cost of a task. This means we need to be able to quantify this effort, which is not a trivial task.

The most popular measure of effort, the (self-)assessments performed by students and/or teachers [72, 58], are time-consuming and require manual input. In some scenarios, such as e-learning, asking teachers to rate their students' effort is impossible as their interaction with students is virtually non-existent. Furthermore, this type of data is not ideal to be used in a recommendation system because it can compromise the quality of the recommendations due to problems such as cold start and data sparsity [144].

On the other hand, the objective measures also present problems. For instance, some metrics, such as the time spent on a task, can lead to conflicting results [74, 155, 78]; while other metrics, such as the number of materials consulted [152], are captured over a longer time period and not just during a given task, and therefore, are not suitable for our purpose.

- *How to estimate the effort that students will exert in future learning activities?*

In the context of recommendation systems, one of the requirements to formalize the foot-in-the-door technique is to define the cost of an activity a at a future time t for a user u using only past data in order to recommend the best activities. In other words, how could we use the effort measures to predict the effort an activity will require from a student?

In the literature, we have found only two works that seek to predict some form of effort in future tasks. The first work seeks to identify whether or not students will continue exerting effort in their next task [156]. Meanwhile, the second study attempts to predict how much effort a student will exert in a future task using data strongly related to programming tasks and a different model for each type of task [89]. Both of these works present a few limitations regarding our main goal. The first study does not allow us to distinguish between different effort levels, while the second is deeply related to programming tasks and cannot be easily transposed to the context of our study.

Finally, in order to achieve our goal in the context of Item 1.3 of the PIA e-Fran METAL project while respecting the current laws and ethical principles, we established the following constraints:

- **C1 (Privacy and ethics):** As imposed by the General Data Protection Regulation (GDPR) applied to (but not limited to) France, the country where this thesis was carried out, the privacy and ethical aspects must be respected throughout the thesis. Besides the current law, another motivation for this constraint is the growing awareness of the worldwide population about such issues, especially after the Cambridge Analytics scandal in 2018 [116].

- **C2 (Implicit data)**: As already mentioned, asking students and/or teachers to evaluate the effort made consumes a lot of time and might compromise the recommendations' quality. Thus, in order to guarantee the users' satisfaction and the reliability of the proposed models, we establish that we should rely on implicit data.
- **C3 (Real life application)**: Due to the context imposed by the PIA e-Fran METAL project, we need to ensure that our recommendation system can be used by secondary school students and that its recommendations are adapted to them. This implies that we should consider the specific characteristics of these students and ensure that the chosen data sources are consistent with real-life scenarios.

Main research question: To what extent the formalization of the foot-in-the-door technique in a recommendation system can influence the students' effort, compliance, performance, and engagement?

Considering the context of the PIA e-Fran METAL project and the constraints established, we can rewrite our specific research questions as follows:

- What is effort? What are the differences between effort, engagement and cognitive load?
- What implicit data can be used in real-life scenarios to estimate the effort secondary school students have exerted while undertaking a learning activity?
- What implicit data can be used in real-life scenarios to estimate the effort secondary school students will exert when undertaking a future learning activity?

1.3 Contributions and thesis structure

Generally speaking, each chapter of this thesis constitutes a contribution. Chapter 2 addresses the technological aspect of learning analytics and presents a literature review on recommendation systems. We introduce the domain and its ramifications in sequence-aware recommendation systems and in educational recommendation systems, as well as the main problems addressed by each one of these sub-domains.

Then, Chapter 3 presents a literature review on the pedagogical aspect of learning analytics. That is, Chapter 3 discusses and distinguishes the concepts of effort, engagement, and cognitive load in order to answer our first specific research question. We also identified theories to support our analysis and several possible ways to measure the students' effort and moved towards answering the second and third specific research questions. Unfortunately, these measures were mostly proposed based on experiments carried out with adults in non-educational settings, which prompted us to validate them in the context of this thesis.

For this reason, as presented in Chapter 4, we build our own dataset in compliance with the French law. This set contains subjective, performance, physiological, and behavioral data captured from 120 seventh grade students who solved English exercises. Therefore, this dataset reflects all of the constraints previously established. We end the chapter by presenting an overview

and a preliminary analysis of the data collected considering the students' effort, their grades and the time they spent solving the exercises.

This dataset is then exploited in Chapter 5, where we present three contributions related to the effort measurement and prediction. The first contribution validates the data collected by training a few machine learning models to estimate the students' effort following the same methodology adopted in several cognitive load studies. These models consistently outperform chance and the state-of-the-art models. Therefore, we consider that our dataset can reliably measure the effort exerted by teenage students during English exercises.

The second contribution presented in Chapter 5 explores the combination of the collected data – which we call engagement features because they reflect the definition we adopted in Chapter 3 – to measure the students' effort. These models consistently outperform chance and the state-of-the-art. These results suggest that the engagement features are uniform enough to measure the students' effort in a past task.

The third contribution of Chapter 5 is the adaptation of the previous contribution to train the effort prediction models. That is, since the engagement features are uniform enough to measure the students' effort, we investigate whether or not those features are also uniform enough to predict the students' effort in a future task. The resulting models perform consistently better than chance and are as good as the previous effort measurement models, which corroborates our assumption.

In all of the three approaches presented in Chapter 5, the interaction features (i.e., a combination of behavioral and performance data easily captured in a virtual learning environment through activity loggers) have as much predictive power than other types of data when it comes to estimating the students' effort. This is particularly interesting because this type of data fully respects all of the three constraints we established in this thesis and allows our models to be fully exploited in real-life applications. Another interesting finding related to our effort models is that we were able to achieve high accuracy levels without the need of subjective ratings, suggesting that our measures are reliable enough to estimate the students' effort using only implicit data.

In Chapter 6, we address our main research question. For this, we use the effort measurement and prediction models proposed in Chapter 5 and propose different formalizations of the foot-in-the-door technique. These formalizations are then combined with different recommendation algorithms and different scoring functions to allow a comparison between the recommendation models that apply the technique and those that do not, and also with the path that was actually followed by the students. The results show that the exercises recommended by the models that were combined with the formalization of the foot-in-the-door technique had a positive influence over the students' effort, compliance, performance, and engagement.

Finally, in Chapter 7 we conclude this manuscript and present several research possibilities.

1.4 Scientific publications

This work resulted in five scientific publications:

- Moissa, B., Bonnin, G. and Boyer, A., 2019. Exploiting Wearable Technologies to Measure and Predict Students' Effort. In *Perspectives on Wearable Enhanced Learning* (pp. 411-431). Springer, Cham.
- Moissa, B., Bonnin, G., Castagnos, S. and Boyer, A., 2019, March. Modelling students' effort using behavioral data. In *Technology-enhanced & Evidence-based Education & Learning Workshop at LAK (TeEL'19)*.

- Moissa, B., Bonnin, G. and Boyer, A., 2019. Building a student effort dataset: what can we learn from behavioral and physiological data. In Learning & Student Analytics Conference (LSAC'19).
- Moissa, B., Bonnin, G., and Boyer, A., 2020, Towards the exploitation of multimodal data to measure students' mental effort. In Proceedings of the 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT'20).
- Moissa, B., Bonnin, G., and Boyer, A., 2021, Measuring and predicting students' effort: a study on the feasibility of cognitive load measures to real-life scenarios. In Proceedings of the 16th European Conference on Technology Enhanced Learning (EC-TEL'21).

1.5 Important note

We would like to note that during the last two years of this thesis, we were in the middle of the Covid-19 pandemic. This situation had an impact on this research work. The closed schools, universities, research labs, and other related services made it difficult to contact project stakeholders; compromised the planned collection of additional data (i.e., student sociodemographic data, exercise metadata, and system activity log) with the teachers, the rectorship and the company responsible for the virtual learning environment; and also the evaluation of the recommendation systems with teachers and/or students.

Chapter 2

Recommendation systems overview

As stated in the Introduction, the learning analytics domain has two sides: the technological and the educational. In this chapter, we review the literature related to the technological side, more specifically, that of recommendation systems. This technology interests us because it can help students, teachers and other stakeholders to find the best learning resources in order to achieve their learning goals and also to develop competences in less time [60]. By doing so, such systems can help to increase the students' performance, promote the student collaboration and increase motivation [60].

As recommendation systems were mostly developed with commercial applications in mind [116], we will from now on refer to such systems as traditional recommendation systems. We also review the related literature to have a broad overview of what are recommendation systems, what problem they solve, their advantages and limitations, as well as how these systems choose the items to recommend.

Such systems are not limited to commercial purposes, and have been further adapted to the educational domain where they are referred to as educational recommendation systems. With educational purposes in mind, new goals and challenges must be addressed to provide meaningful recommendations. In the context of this thesis, these goals and challenges are of our interest and, therefore, we review them focusing on state-of-the-art solutions.

This chapter is organized as follows: Section 2.1 presents the traditional recommendation systems and their recommendation approaches; Section 2.2 presents the challenges imposed by educational recommendation systems; and Section 2.3 presents our final thoughts regarding the literature review presented in this chapter and links it to our proposal.

2.1 Recommendation systems

In the beginning of this thesis, we stated that the use of data is a growing phenomenon in several fields [44]. What we did not mention is that there is already a large amount of data available and that this amount grows continuously and exponentially [3]. This huge amount of available data can be a challenge to users searching for relevant/useful information who, without the appropriate tools, might feel as if they are searching for a needle in a haystack. This unfruitful search not only makes it hard for users to accomplish their tasks, but can also cause an information overload (i.e., difficulty to understand an issue and to make a decision due to the excess of information) [3, 80].

One way to overcome this problem is to use recommendation systems, which became an independent research field in the mid-1990s [144]. They can be defined as software tools and

techniques that provide recommendations of items that might suit the users [100, 144]. In other words, recommendation systems filter the available information in order to recommend the best items to the users taking into account their interests, and sometimes the interests of the provider. For instance, a travel recommendation system can have as goal to maximize the number of room reservations, while a recommendation system provided by a destination management organization would like to increase the number of visitors in the city [144]. It is important to note that there must be a balance between the needs of the provider and the user in order to make the recommendations valuable for both sides.

2.1.1 Formal definition

In order to make recommendations, these systems track the users' interactions and stores their data. The data are then fed to the recommendation system model, the core of the system, to infer the utility of the item to each user (i.e., choose the items that will be recommended) [144]. In this context, there are four essential elements:

1. **The users:** People who use the system and also who will receive the recommendations. They can be tourists, customers, students, etc. They can be represented only by a unique identifier (e.g., customer 1, ..., customer n), by a set of attributes (e.g., gender, age, country, etc.), as a concept in an ontological representation of a specific domain, etc. [144].
2. **The items:** What will be recommended to the users. They can be movies, songs, products, learning activities, etc. Just like the users, the items can be represented only by a unique identifier, by a set of attributes, as a concept in an ontology, etc. [144].
3. **The transactions (the data):** It is the set of interactions between the user and the items. The most popular transaction example is the feedback a user provides when he provides a rating for an item (explicit rating) [144]. However, these transactions can be anything that will later be fed to the recommendation system in order to infer the suitability of the items (implicit rating). As shown in the next sections, recommendation systems can exploit the sequences of user activities, the user interactions (and their contexts) in e-commerces, in social media, etc. [144].
4. **Filtering technique:** The filtering technique is the approach chosen to decide how well an item fits the user or, in other words, whether the item is worth recommending or not [144]. We can, for instance, decide if we are going to recommend an item based on the satisfaction it would bring to the user. This means that the recommendation system would apply the chosen recommendation technique on the users' data to predict how satisfied a given user would be and, based on it, decide whether to recommend it or not.

These elements can be further formalized in a recommendation system as follows: a user u belongs to a set of users U , an item i belongs to a set of items I , a transaction t belongs to the set of transactions T , and the function $C(\dots)$ is responsible for choosing the best items to recommend [20, 80, 47, 1]. The function $C(\dots)$ is considered to be the core of a recommendation system and, because of that, we can find several approaches to implement it. Those approaches are presented in the next section.

2.1.2 Filtering techniques

Several recommendation system classifications have been proposed considering different aspects, such as the information used to make the recommendations and their type [42]. In this section,

we describe the classic recommendation system categorization proposed by Burke [23]. This classification categorizes the recommendation systems according to their techniques, and comprises the most popular ones, namely, content-based, collaborative filtering and hybrid. As the foot-in-the-door is a *sequential* free will compliance technique that requires consecutive requests to be made, we also describe sequence-aware recommendation systems, a recommendation approach that takes into account the order in which the items are used and recommended.

Demographic recommendation systems exploits the demographic information of users (e.g., nationality, level of education, gender, age, etc.) to categorize them and make recommendations [84, 16, 23]. It assumes, for instance, that teenagers have similar tastes that are different than the tastes of seniors, that all French people have similar tastes that are different from the Brazilian people tastes, etc. In other words, this type of recommendation system exploits the users' stereotypes [23] and, thus, it is not individually personalized.

Utility-based recommendation systems make recommendations to the user u by taking into account the result of an utility function defined for him [23, 2]. In this approach, the central problem is to create the utility function and the main advantage is that it allows the inclusion of additional attributes, such as the vendor reliability, the product availability, etc. [23]. This allows, for instance, a trade-off between price and delivery dates for a user identified as having an immediate need [23].

Content-based recommendation systems choose which items to recommend by taking into account how much they have in common with the characteristics of the items the user liked in the past (user profile). This approach has some limitations such as the need of a sufficient number of features describing each item and the assumption that two items are the same if they have exactly the same features (limited content analysis problem) [22]; the high similarity between all recommended items that prevents the user to discover new items (over-specialization or serendipity problem) [22, 80]; and the system's incapacity to recommend items when a user has not rated enough items to enable the creation of his profile (new user problem) [22].

Collaborative filtering-based recommendation systems are inspired by human social behavior, where people often take into account the experiences, preferences and opinions of their acquaintances to make a decision [20]; and tackle the limitations of the content-based approaches because they can make recommendations regardless of the content of the item and the user profile [16]. This approach mimics the human social behavior through two approaches: neighbourhood-based and model-based approaches, both usually combined with the Top N algorithm to choose the best N items according to some predefined criteria (c.f., Chapter 6.1).

The *neighbourhood-based collaborative filtering* (also known as memory-based) searches for the best neighbourhood users/items who share similar interests [16]. This similarity can be computed between users (user-based) or between items (item-based) and express a distance between two users/items taking into account the available ratings explicitly provided by the users [20]. The mains limitations of this approach are related to the lack of user ratings from a user or to an item (cold start problem) [22], to the insufficient number of available user ratings to provide quality recommendations (data sparsity problem) [80, 22], to the unique preferences of a user that prevents the systems to recommend good resources to this particular user (grey sheep problem) [62], and to the need for computational power capable to deal with the growing number of users, items and ratings available (scalability problem) [80, 18].

On the other hand, the *model-based collaborative filtering* approaches exploit a model – such

as Bayesian classifiers, neural networks, fuzzy systems, and, more recently, matrix factorization models [98] – to identify the best items to recommend [20] and, therefore, they can process large sparse data sets better and faster than the memory-based methods [151] tackling some limitations of the neighbourhood-based approach (e.g., scalability, data sparsity).

Knowledge-based recommendation systems differ from the previously presented systems because they have functional knowledge or, in other words, they have knowledge about how an item fits the needs and/or preferences of a user and, because of that, they can decide whether this item should or not be recommended [23]. This type of recommendation systems can better address the recommendation of complex products, such as cars, real-estate, etc. [51, 2]. For instance, financial services are not contracted very often making it impossible to have enough ratings to use collaborative filtering approaches and, even if after some years there are enough ratings, the customers of such services might not be satisfied with recommendations based on old ratings because the products might have become outdated.

There are two main knowledge-based recommendation approaches, case-based and constraint-based [144], that differ from each other due to the way they choose the items to recommend: while the *case-based approaches* choose the best items based on similarity metrics, the *constraint-based approaches* exploit knowledge bases containing explicit rules to choose the best items. Although these approaches do not suffer from the cold start problem and neither with the grey sheep problem, they require lots of work to build their underlying knowledge base (knowledge acquisition problem) [22, 51], and might still not find items that completely match the users' requirements (i.e., no items to recommend) [51].

Hybrid recommendation systems combine one or more of the previously mentioned techniques to exploit their advantages and avoid their limitations [16, 22]. For instance, it will use the results of a collaborative-filtering and of a content-based collaborative filtering at the same time [16]. There are several ways to combine the chosen methods, such as [23]:

- **Weighted:** combines the ratings of the chosen systems to obtain the final rating for each item;
- **Switching:** the system changes the approach according to the context;
- **Mixed:** the recommendation of all approaches are presented at the same time;
- **Feature combination:** the data used in all approaches is combined in a single algorithm;
- **Feature augmentation:** the output of one approach is used as input for another approach;
- **Cascade:** one approach refines the resulting recommendations of another approach;
- **Meta-level:** the model learned by one system is used as input for another system.

Sequence-aware recommendation systems have been gaining more attention recently [138] because they address the sequential limitations of the previously presented approaches. They take as input transactions that are chronologically ordered, allowing the recommendation system to focus on recent data, make repeated recommendations, and/or to identify temporal dependencies in the data [82]; and recommends a new sequence of items. Such characteristics can be interesting in several application domains, such as cultural heritage, health, music, and education [188].

Despite the lack of a clear definition, sequence-aware recommendation systems can be roughly categorized as session-based and as sequential recommendations. In a broader sense, we can understand that the *session-based recommendations* (e.g., music) predict the next items considering only the ongoing user session and, therefore, considering only short-term preferences (and smaller datasets); while *sequential recommendations* (e.g., points-of-interest) consider all of the previously captured data, which means it considers long-term user preferences and bigger datasets to make the recommendations [82].

Given the different input format, sequence-aware recommendation systems require a different set of approaches to make recommendations. According to Wang et al. [180], the three main approaches are traditional sequence models (e.g., sequential pattern mining, Markov chains), latent representation models (e.g., factorization machine-based and embedding-based), and deep neural network models (e.g., recurrent neural networks, convolutional neural networks, and graph neural networks). However, we can also find approaches based on reinforcement learning. These approaches are particularly interesting because they can continuously update the recommendation strategy according to the user interactions, and maximize the expected long-term cumulative reward [189, 107]. One example of such approaches is the Q-learning algorithm (c.f., Chapter 6).

2.2 Educational recommendation systems

As shown in the previous section, recommendation systems can recommend different types of items, including learning resources, such as learning paths, papers, web pages, courses, lessons, disciplines, places to study, etc. As a matter of fact, one of the first noteworthy educational recommendation systems appeared in the early-2000s [111]. Since then, several recommendation systems for education were proposed based on, but not limited to, collaborative filtering, content filtering, and hybrid approaches [193, 145].

Given the specificities of the educational domain and its resources (e.g., different goals, data structures, user preferences), the recommendation approaches used for commercial purposes cannot be simply transferred to the educational context where recommendations must support the learning process [145, 41, 86]. For instance, a recommendation system based on users' satisfaction might be good in order to recommend items to buy, but might not help students to learn because, even though students might be satisfied with the provided recommendations, these recommendations might still be not good enough to help them to succeed in their course. Since recommendations made in the educational context affect the users' learning, such recommendations are potentially harmful, especially if they are made to children [121].

According to Murgia et al. [121], recommendations for children in educational contexts have seven layers of complexity that should be addressed. As we believe those layers could be addressed in every educational recommendation system regardless of the target audience, in each one of the following sections we discuss one of them, and make a parallel between educational recommendation systems for adults and for children.

2.2.1 Need for personalization: one size does not fit all

When making educational recommendations the students are the main protagonist, which means that the recommendations made should support them. In order to do so, we must consider that each student is unique and has his own skills, needs, personality, cognitive development, level of engagement, degree of interest, etc. [121]. By considering such differences, it becomes clear that a one-size-fits-all recommendation system will unlikely produce a great learning experience [46]

and, therefore, personalization becomes a critical requisite of any educational recommendation system.

How such personalization is achieved might depend on the beliefs of who is developing the recommendation system, the target audience, on the adopted learning theories, on the recommendation goals, etc. For instance, to design an educational recommendation system for children, Murgia et al. [121] propose to revisit and adapt the seven different roles children play in the search process – visual, developing, power user, distracted, domain specific, non-motivated, and rule-bound users [43] – while also considering that every children can play different roles according to the context in which they are. More specifically, personal preferences define the more visual users who look for non-textual information, different levels of experience/familiarity with the recommendation process result in the development or in the power user, personality has a higher impact on the distracted user, the preferences of the specific user must be taken into account (e.g., by considering what the student wants to accomplish) to avoid reducing his motivation, and the rule-bound user follows the influence and the guidance provided by older and more experience stakeholders (e.g., teachers, parents). On the other hand, when developing a recommendation system for adults, Klašnja-Milićević et al. [95] propose eight students' characteristics (instead of roles) to be considered when making recommendations: the students' goal and task, students' prior knowledge, students' background and past experiences, students' preferences, students' learning paths, students' learning strategies, the group students belong to, and the rated learning activities.

Although several personal characteristics could be taken into account while personalizing educational recommendations, additional attention must be given to recommendations to children due to their characteristics that oppose those from adults. For instance, in some user studies, Landoni et al. [102, 101] observed that the previously cited elements have a stronger impact on younger users causing more extreme reactions. For instance, they observed in their study that children may lack experience and fail to engage with the recommendation system, while adults are used to it; children exhibit low engagement with non-interesting tasks, while adults take advantage of the recommendations to deliver some sort of result no matter what; finally, contrary to the adults, children tend to assume the rigid rule-bound and mistrust recommendations without any explicit information about their source [121].

Personalization is one of the challenges that must be tackled when developing an educational recommendation system. As will be shown in Chapter 3, it reflects the learning science, which deeply acknowledges learners' unique characteristics and their effects onto learning (i.e., each learner is unique and has his own needs and preferences). Therefore, this constitutes one of the requisites of our recommendation system.

2.2.2 Multiple stakeholders' perspectives

Recommendation system for children should take into account not only different user characteristics (e.g., age, knowledge level) and preferences, but also the opinions of multiple stakeholders (e.g., teachers, parents, researchers) [121].

This issue can be tackled by involving the stakeholders in the development of the recommendation system, but also through multi-stakeholder recommendation systems, such as the one proposed by Burke et al. [25] to take into account the preferences of different users (and not just the end user). An example of such system in the educational context is the one proposed by Burke and Abdollahpouri [24] to promote off-line programs in an online platform called City of Learning. Since programs are offered by organizations, have limited capacity and also other pre-defined requirements (e.g., gender equity, age, education level), the recommendation system

considers the users' interests and the programs' requirements to make the recommendations. As this is still an ongoing project, the authors did not provide any further information regarding its implementation or evaluation.

Another example is the work from Zheng et al. [192, 191]. They propose a dataset recommendation system to be used during a data analysis project. They developed a utility-based recommendation system that balances the student's and the teacher's perceived utility in order to recommend datasets that are appealing for both of them (i.e., teachers do not want students to use the easiest datasets, but they do not have an upper limit). Their offline evaluations suggests that the proposed approach can capture the different perceptions students and teachers have.

Unfortunately, none of these studies describe user studies or real-life experiments to provide further insights regarding their approach due, not only to the approach novelty, but also to the complexity of developing such systems and carrying such experiments. The main advantage of this approach is the exploitation of the teachers' role to transmit confidence and trust to the end users. This is especially true when the end users are children, who have less autonomy and less critical sense. Therefore, accounting for the preferences of multiple stakeholders while making recommendations is a promising way to balance the preferences of students with learning (c.f., Section 2.2.3). Although this aspect will not be directly considered in this thesis because we are mainly interested in how to exploit the foot-in-the-door technique to engage students, it can be indirectly addressed by the inclusion of the teacher in the recommendation process and its role to forward the best recommendations to the students.

2.2.3 Good recommendations to foster learning

From a technological perspective, Klašnja-Milićević et al. [95] state that a good educational recommendation system is highly personalized, recommends at the right time and location (i.e., context-aware recommendations), is non-disruptive (i.e., the student has the option to follow the recommendation or not), is socially situated (i.e., it recognizes and exploits students' social networks, role models, levels of trust, etc.), includes the adoption phase, supports the learning process, is highly interactive, and recommends appropriate items according to the students' characteristics.

From a pedagogical perspective, we can roughly say that appropriate/good recommendations are those who help students to learn. However, there is no clear and balanced view of what are such recommendations [78], especially when we consider the different learning/teaching approaches and learning goals [146]. For instance, to Murgia et al. [121] a good recommendation system would meet the four dimensions of teaching outlined by Fadel et al. [49]. More specifically, the recommendation system would facilitate the acquisition of new knowledge, encourage the development of skills, enable reflections on the actions taken and their effects, and boost the development of the emotional side of learning (e.g., develop confidence, autonomy, self esteem in a pleasant environment). On the other hand, Barria-Pineda et al. [12], who propose a recommendation system for college-level students, believe that a good recommendation system should seriously address the current level of learner knowledge rather than their preferences, while also addressing students' goals and teachers' instructional practice for each course and/or discipline. Both points-of-view are clearly concerned with the benefits the recommendation system can offer to the student, but they differ on how to achieve this goal. While Murgia et al. [121] superficially address several aspects, Barria-Pineda et al. [12] address in depth the facilitation of knowledge acquisition.

In this thesis, we propose yet a different point-of-view. Instead of directly focusing on the suitability of the recommended learning resources for a given learning goal, we want to exploit a

free will compliance technique to maximize students' engagement, which as shown in Chapter 3, has been linked to better learning outcomes. We are assuming that, in its vast majority, students engage with learning resources that help them to achieve their learning goals and, therefore, by recommending engaging learning resources we are also fostering learning.

2.2.4 Need of explanations

Explanations of the provided recommendations can help in the adoption success of a recommendation system because they explain how the system works. They increase the system's transparency (a possible solution to the opacity ethical issue mentioned in the next section), they allow the user to tell the system it is wrong (scrutability), they increase the user confidence in the system (trust), they help users to make good decisions (effectiveness) faster (efficiency), they can convince the user to try/buy an item (persuasiveness), and they increase the ease of use and/or enjoyment (user satisfaction) [143].

Explaining recommendations is a complex problem on its own and may require diverse data sources and perspectives of what will make them meaningful to the users [121, 187]. In the educational context, these explanations can help students to assess the quality and the relevance of the recommended resources [75], and are particularly important for children. As previously mentioned, Landoni et al. [102, 101] observed that children, contrary to adults, tend to mistrust recommendations without any explicit information about their source.

Works aiming to provide explainable recommendation (for adult students) include Hosseini et al. [76] and Barria-Pineda et al. [12, 13]. Hosseini et al. [76] propose to justify the value of an item by visually representing the knowledge accumulation associated with it, while Barria-Pineda et al. [12, 13] propose to explain their remedial recommendations (i.e., recommendation of resources to help a student overcome a difficult concept) for college students by using visual and text explanations. The visual explanations highlight concepts that should be further studied, while the textual explanations state if the student is struggling with the content and if he has the required knowledge. They found out that the explanations increased the acceptability of the recommendations (i.e., students who had access to the explanations have shown a higher probability of trying to do the activity).

In our proposal, we do not tackle the students' trust issues through explanations, but by including the teacher in the loop, as proposed by Pera et al. [135] and by Ekstrand et al. [46]. Presenting the recommendations to the teacher who will then forward them to the students if he agrees with them, might be a simple solution to the students' trust issues as the teacher will choose the learning resources and, in a well-designed e-learning platform, the children knows the recommendation comes from him. This presents new challenges because recommendation systems where experts mediate the recommendations are not a well-studied problem [135, 46].

2.2.5 Ethical challenges

The data collections and the extent to which it can be manipulated is object of several laws in the European Union, the United States, and in other places. These laws are even stronger when it comes to children (i.e., they need more protection because they are potentially less aware of the risks than adults) [121, 45].

Recommendation systems are not only the object of these laws, but also of some concerns inherent to this technology. Milano et al. [116] describe in depth six ethical challenges posed by recommendation systems, namely, inappropriate recommendations, privacy issues, threats to

the autonomy and personal identity of the user, opacity (i.e., lack of transparency), fairness, and the undesired social effects.

We recognize the importance of ethical and legal aspects related to the use of student data. Therefore, we established the constraint C1 (Privacy and ethics). To respect it, we asked for and followed the advice of experts in the field during all of the stages of this thesis. However, as these aspects are beyond its scope, we will not discuss them in detail.

2.2.6 System acceptance

An important aspect of any system is its usability, which is defined by the ISO 9241-11 [81] as “the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”. Some aspects of usability are mainly tackled by a good system design, but in the context of recommendation systems, the quality of its recommendations plays a major role since irrelevant recommendations will probably be ignored and will not offer a better user experience.

Theoretically, if all the challenges posed by the previous layers of complexity are tackled, this layer is also met. Unfortunately, the system usefulness can only be considered to be met based on user studies and/or real-life evaluations (c.f., Section 2.2.7), which are not carried out in this thesis. However, as we want to recommend learning resources to increase the students’ engagement, the usefulness of our recommendation system can be measured through the amount of recommended learning activities accepted by the students, and also by the amount of recommended learning activities that students have truly engaged with (as opposed to doing nothing at all, pretending to study, and other undesirable behaviors).

2.2.7 Evaluation challenges

The evaluation of recommendation systems is already a hard problem on its own, but when combined with the specificities of the educational context, it becomes even more difficult due to the different recommendation goals and scenarios [50]. For instance, studies target different students audiences (e.g., child vs. adults) in different learning modalities (e.g., formal vs. informal learning) in different learning scenarios (e.g., classroom, blended learning, e-learning). Besides, each application might need a different data granularity (e.g., task-level vs. course-level data) and different data types (e.g., log data, heart rate data, eye gaze data).

If we consider proposals where the expert is included in the loop or multi-stakeholder proposals, the difficulty to evaluate the recommendation system increases even more as there are more variables to be considered and more diverse users to satisfy. This difficulty also increases when recommendations are made to children as their opinions and perceptions are not the same as those of adults. In addition, recommendation systems aimed at children are a recent research interest (as observed by the first KidRec Workshop held in 2017). Children seem to be more interested in items that are interesting, amusing and/or informative rather than in items that are precise and relevant [77]. Thus, simply relying on metrics commonly used to assess recommendation systems might not be sufficient and, in some cases, not even possible [121]. For instance, it has been reported that children do not take full advantage of Likert scales, which makes the use of some metrics, such as the root mean squared error (RMSE), questionable [66]. Furthermore, children typically overestimate their own skills while reading and understanding a text (and probably in other contexts too) and tend to give socially desirable answers [78].

As we need to know if our approach can engage the students, it is inherent of this work to evaluate our recommendation system. Therefore, this layer of complexity is addressed in this

thesis (c.f., Chapter 6). Recommendation systems can be evaluated following four approaches [27, 122]:

1. **System performance:** This approach consists in measuring the system performance to check, for instance, how fast they are to provide the recommendations (response time) or if they work well with large datasets (scalability), etc. [122].
2. **Offline evaluation:** This approach uses existing datasets to evaluate the recommendation system. These datasets can contain real interactions of real users in real utilisation scenarios, or they can be synthetic datasets built to test the recommendation system in specific scenarios and conditions [27].
3. **User studies:** This approach consists in a controlled experiment where participants are asked to perform a set of predefined tasks, and allows researchers to collect quantitative and qualitative information about the recommendation system being assessed [122].
4. **Real-life evaluations:** This approach is similar to the users studies, however, they are longer and there is no interference from the research team during the data collection (i.e., the users are using the recommendation system in their everyday life and data is being collected to make the evaluation) [27].

According to Cañizales et al. [27], the number of educational recommendation systems that were somehow evaluated over the last five years increased. However, most of the evaluations (i.e., 48 out of 91 papers, or 53%) were done through offline approaches. Such approaches are limited by their metrics (e.g., RMSE, accuracy, precision, recall) that characterize algorithmic performance, but overlook requirements from other stakeholders and from the context [121]; and also because it may not properly capture the dynamics of the learning process and lead to failed user studies [50].

Despite being limited in its ability to reflect the actual user response, offline experiments remain a good alternative to test and optimize a recommendation algorithm before deploying it [45]. Given the challenges imposed by user studies and real-life evaluations, especially during the Covid-19 pandemic, our evaluation (c.f., Chapter 6) also relies on an offline approach.

2.3 Chapter conclusion

Recommendation systems were born in the mid-1990s in order to filter the high amount of information available and solve the information overload problem. Nowadays, as the amount of data has significantly increased, they are even more relevant. As proof we can cite the extensive research carried in the domain and the recent related proposals, such as multi-stakeholder recommendations, recommendations for children, and explainable recommendation systems.

Since the emergence of the field, a lot has changed. We moved from recommendations based only on explicit ratings and item content to hybrid recommendations. We developed methods to make recommendations using implicit data, which is already defined in the Introduction as a requirement of our recommendation system in order to lighten the workload of the stakeholders and ensure data availability. We developed model-based recommendations to overcome problems identified within the previous approaches (e.g., cold start, data sparsity) and enhanced them to achieve higher accuracy values (e.g., matrix factorization models). We expanded the recommendations items from movies to music, to products, to services, to people, to learning resources, etc.

The expansion to other types of items caused the emergence of recommendation system specializations, such as the educational recommendation systems which aims to address the special needs of the educational scenario. As any other technology enhanced-learning tool, its ultimate goal is to enhance students' learning. Given the inherent complexity of this goal and its dynamic nature (i.e., the learning processes are evolving), developing a good educational recommendation system is not a trivial task. To reach good recommendations, seven complexity layers should be addressed: personalization, multiple stakeholders perspectives, good recommendations to foster learning, explanations, privacy and ethics, system acceptance and its evaluation.

In this thesis, we will directly address four of these aspects, namely, personalization, good recommendations to foster learning, system acceptance and its evaluation. It is important to note that privacy and ethics aspect will be considered during this entire thesis, but will not be a main goal of our recommendation system.

The personalization aspect will be taken into account by considering each student individually, which as shown in this chapter and also in the next, is highly relevant when it comes to students' engagement and learning. We intend to foster learning by recommending engaging learning resources following the assumption that if the recommended resources are engaging, they are helping students to achieve their learning goals. The system acceptance will be considered during the evaluation of our recommendation system, meaning that the recommendations must be in its majority of engaging resources, which as previously discussed, means students accepted the recommendations made. As an inherent requisite of this work, our evaluation will address the influences of the foot-in-the-door technique over the students' engagement.

Having reviewed the recommendation system domain and its specialisation in the educational context, we have accounted for the technological side of our proposal. In the next chapter, we review the learning science literature to better understand the theories we could use to guide the development of an educational recommendation system and tackle the aforementioned aspects.

In this chapter, we provided an overview of recommendation systems with two different perspectives. The first was a broad perspective showing why and how these systems provide recommendations. The second was specific to the educational domain and addressed seven challenges to make recommendations in this scenario.

These challenges are: personalization, multiple stakeholders, fostering learning, need of explanations, privacy and ethics, system acceptance, and evaluation.

From these challenges, four will be directly addressed in this thesis: personalization, fostering learning, system acceptance, and its evaluation.

Chapter 3

From the theory of commitment to the cognitive load theory

In the previous chapter, we presented an overview of the recommendation system domain, which is comprised in the technological side of the learning analytics domain. In this chapter, we proceed by presenting the educational theories on which this thesis are based, accounting to the educational side of the learning analytics domain.

As stated in the Introduction, our goal is to provide recommendations to engage students into learning by formalizing the foot-in-the-door technique. Therefore, the literature review presented in this chapter has three objectives. The first is to understand the foot-in-the-door technique and its underlying theory, as well as its variants and important aspects for it to work. The second objective is to clarify the differences between students' engagement and students' effort, which often have overlapping definitions and measures, and describe the state-of-the-art measures used for them. Both constructs are often measured through indicators that are either time consuming and cumbersome or not feasible to be used at the task level. The third objective is to identify effort measures that respect the constraints established in the Introduction. To achieve this goal, we propose to rely on the cognitive load construct, which is considered by several researchers as being equivalent to the mental effort and it is often measured at the task level [129, 105]. Moreover, it offers several insights about the learning process.

This chapter is organized as follows: Section 3.1 presents the theory of commitment and the foot-in-the-door technique (and its variants); Section 3.2 discusses students' engagement and effort; Section 3.3 presents the cognitive load construct, its theory and measurements; and Section 3.4 presents our final thoughts regarding the literature review presented in this chapter and links it to our proposal.

3.1 The theory of commitment

In the social psychology literature, commitment is the “binding of the individual to behavioral acts” [94] and can, up to a certain extent, be interpreted as engagement. According to the theory of commitment, people who act become committed [93]. To make people act, it is possible to exploit several social influence tactics to induce people to comply with requests, in other words, techniques to influence people to do what we want them to do. According to Joule and Almeida [87], these tactics are interesting for educational contexts because they can influence students to change their conviction, choices and behaviors without resorting to argumentation, the use of rewards or to punishment.

In the literature, we can find more than 107 social influence tactics in the literature [137, 40]. However, Joule and Almeida [87] cite only a few of them as suitable for pedagogical use: the foot-in-the-door [59], the foot-in-the-door with implicit request [171], the foot-in-the-door with labelling [65], the touch [96], and but-you-are-free-to [68]. We are particularly interested in the foot-in-the-door technique and its variants because the touch and the but-you-are-free-to techniques are not relevant to provide engaging recommendations. As a matter of fact, the touch technique [96] consists in making a request to the subject while gently touching his arm which is not possible to implement in a virtual learning environment; and the but-you-are-free-to technique [68] consists in making a request followed by the sentence “but you are free to say yes or no”, which can be easily implemented in a virtual learning environment, but can also be easily ignored, especially after a few recommendations as it will look like a default message from the system. On the other hand, the foot-in-the-door works by making consecutive requests, which can be integrated to a virtual learning environment.

The *foot-in-the-door technique* was first introduced by Freedman and Fraser in 1966 [59]. In their experiment, housewives were asked to answer a few questions about the kinds of soaps they used over a phone call (small request), and three days later they were asked to allow a survey team of five or six men to come into their homes for 2 hours to classify the household products they used (larger request). When compared to the control group (housewives that were asked only to welcome the male team in their homes), the compliance rate of the group in the foot-in-the-door condition was higher (52.8% vs. 22.2%).

Uranowitz [171] also used this technique, but with an implicit request, a variant called *foot-in-the-door with implicit request*. In the study, women walking alone in a shopping mall who did not appear to be in a hurry were approached by the first experimenter who was carrying five bulky grocery bags. He asked these women to watch his bags for a moment while he went back to the store to retrieve his wallet. After the task was accomplished and the woman left, a second experimenter would drop a bag near the same woman and leave. 80% of the women who were asked to take care of the first experimenter grocery bags alerted the second experimenter about the bag he dropped, while only 35% of the subjects who were exposed only to the second experimenter implicit request did the same.

Another variant was proposed by Goldman et al. [65]. The *foot-in-the-door with labelling* comprises a compliment to the subjects after the first request. More specifically, Goldman et al. [65] analyzed four different conditions: 1) the foot-in-the-door, 2) just the target request (control condition), 3) the foot-in-the-door with a positive adjective, and 4) the foot-in-the-door with a negative adjective. In their experiment, a first experimenter approached university students entering a library (subjects) and asked for directions to a given building. After the student answered, the experiment conditions were applied and the experimenter answered accordingly: in condition 1 the experimenter simply thanked the student: “Ok, thank you.”; in condition 3 the student was more enthusiastically thanked: “Thank you very much. You have been very helpful and I appreciate you taking the time to help me”; and in condition 4 the experimenter said with an annoyed voice: “You are not very helpful, and I can usually understand directions. I’ll have to find someone who can be more helpful.”. Note that subjects in the condition 2 were not approached by the first experimenter for obvious reasons. A second experimenter would then approach the subjects inside the library and ask the subjects to join a list of people willing to donate two hours of their time to answer phone calls for a charity telethon. The results show that 40% of subjects in condition 1 agreed to help, against 17% in condition 2, 67% in condition 3, and 20% of subjects in condition 4. By labelling each subject as 1 when the request was accepted and otherwise as 0, Goldman et al. [65] found a statistically significant difference when comparing the foot-in-the-door and the control condition, the foot-in-the-door with a positive

adjective and the foot-in-the-door, and the foot-in-the-door with a negative adjective and the foot-in-the-door.

As described, the foot-in-the-door (and its variants) consists mainly in making consecutive requests in order to increase the likelihood of having the following request accepted. Thus, it is particularly interesting to us because the cost of the tasks increase from the first request to the second, which is compatible with the zone of proximal development [177] that states that the learning contents must gradually increase in difficulty to encourage and advance their learning. Additionally, it is a technique that can be applied in an e-learning system and, more specifically, in a recommendation system. To achieve this goal, we defined the cost of each task as the effort it requires from the student, which as shown in the next sections is related to learning.

3.2 Students' engagement vs. students' effort

The interest on students' engagement emerged from the wish to make students learn more, and the assumption that students can be influenced to enhance their educational performance attracts lots of interest from researchers [142, 54]. It has been defined as investment, as commitment, and also as effortful involvement in learning [72]. However, despite the different definitions found in the literature, it is common to find the term being used without a definition [52, 72, 160]. As a matter of fact, Bedenlier et al. [15] carried a systematic review about engagement through educational technology in the arts and humanities field and, from the papers they analyzed, only 2 (out of 42 papers, or 4.8%) provided a definition.

Even without a clear definition of what is engagement, the multidimensionality of the concept is acknowledged and three main interrelated dimensions are recognized [57, 56, 19, 5]:

1. **Behavioral dimension:** This dimension is based upon the participation concept and relates to a positive conduct, such as exerting effort, being persistent, concentrating, paying attention, asking questions, school attendance, suspensions, voluntary classroom participation, and extracurricular activity participation (e.g., athletics or school governance), etc. Given the elements comprised in this dimension, it is easily and directly observable [5].
2. **Emotional dimension:** This dimension is related to the students' feelings. In other words, their reactions (whether they are positive or negative) towards their teachers, peers and school. These emotions can be, for instance, interest, boredom, happiness, sadness, anxiety, stress, pressure, etc. It is assumed that if students feel good about the educational context, they will be more likely to do what they have to do.
3. **Cognitive dimension:** This dimension is related to the idea of investment, of incorporating thoughtfulness and willingness to make the required effort in order to understand concepts and to master skills.

More recently, some additional dimensions were proposed, such as the socio-behavioral dimension linked to the students' affect and behavior during collaborative group work [109], and the agentic dimension that addresses how students proactively contribute to the instruction their teachers provide [141]. Yet further research is still necessary to determine the extent to which these dimensions are relevant for the engagement construct [57].

If we look into the description of the three main engagement dimensions, we can see that effort is one of its components, more specifically, engaged students exert effort. Effort is considered a key element for learning, as it is a required factor in order to understand complex ideas, to acquire

knowledge and to develop skills [58]. However, just like the engagement concept, students' effort has several proposed definitions in the literature (when one is provided), but although related, none of them are widely adopted [115]. From the definitions found in the literature, some are simpler and some are more elaborated. For instance, Schuman [154] simply defines effort as "the amount of studying", while Dev [119] defines students' effort as the ability to persist with the task, the amount of time spent on it, the curiosity to learn, the feelings of efficacy related to it, or a combination of these factors. Other definitions are very similar to the engagement definition. For instance, Carbonaro [28] defines effort as the amount of time and energy expended to meet formal requirements established by the teacher and/or by the school. He further classifies effort in three categories:

1. **Rule-oriented effort:** This type of effort is related to students' compliance with school's rules and norms (e.g., school attendance) and is similar to the behavioral dimension of engagement.
2. **Procedural effort:** This type of effort is related to students trying to meet the demands of the teacher (e.g., completing assignments, turning in assignments on time, participating in class discussions). Just like the rule-oriented effort, this type of effort is also similar to the behavioral dimension of engagement.
3. **Intellectual effort:** This type of effort is related to the wish of doing the tasks correctly and is somewhat related to the cognitive dimension of engagement.

The main difference between Carbonaro's [28] definition of effort and the three engagement dimensions, is that Carbonaro's [28] definition does not comprise the students' emotions and how it influences their effort. However, Stables et al. [163] provide different interpretations of effort that include the emotional aspects (addressing also the emotional dimension of engagement). In their study, effort could be interpreted as:

- **Achievement:** good quality assignments imply that students have worked hard;
- **Perseverance:** the more time spent doing the assignment, the harder the student worked on it;
- **Reliability:** students that always deliver the assignments on time and within the established criteria; and
- **Students' interest:** students are interested in the task or they enjoy it. This interpretation of effort is related to the emotional dimension of engagement.

Based on these effort interpretations, Stables et al. [163] studied how students, parents and teachers perceive the students' effort. Their results indicate that teachers perceive their students' effort as achievement and reliability, while parents perceive students' effort as achievement and perseverance. On their turn, students recognize the importance of making effort, but also state that they will not exert effort on all tasks, and that they will exert more effort on tests or on tasks they find interesting (i.e., effort as students' interest).

This variety of effort and engagement definitions and perceptions leads to different interpretations of how they should be measured. Thus, several different measures can be found in the literature. The most popular way to measure effort and engagement is through students' and teachers' subjective evaluations [72, 58, 33, 9]. One example is the work of Swinton [169] who

used teacher ratings to positively relate students' effort with students' outcome. Using the same type of measures, Nagy [123] proposes a learning analytics tool that allows teachers to assess whether their students are exerting more or less effort over time in order to provide feedback to their students. Despite the popularity of these measures due to its ease of use, they are time consuming and cannot be automated. In some scenarios, such as e-learning, teachers cannot assess their students' effort reliably as their interaction with their students, apart from a few exchanged messages, is nonexistent. There are also other types of subjective measurements used to measure effort/engagement, such as interviews and observations [58]. However, these approaches present additional limitations, such as the participation of well trained professionals to collect the data.

A popular objective measure believed to be related to engagement/effort is the time spent on the task. It has the advantage of being easily captured in virtual learning environments, but the studies using it present conflicting results. For instance, Schuman et al. [155] used the time spent on a task as a measure for students' effort. Their results do not show a correlation between the collected measures and the learning outcomes. Patron and Lopez [134] study shows that consistency (i.e., lower variation of the time spent on tasks) is a statistically significant explanatory variable of higher grades, while the time spent online is not. Hill [74] also used the time spent as a measure of effort, but he was able to positively relate the time spent studying during weekends with the learning outcomes. Although not directly linked to the concept of effort and in a different scenario, Huibers and Westerveld [78] noticed that the students who took longer to complete a task provided better answers.

Using several objective measures, some models have been proposed. For instance, Scariot et al. [152] proposed a model that combines information about the amount of resources and assignments available with the students' interaction with a virtual learning environment in order to positively link students' effort with their grades at the end of the course. Liu et al. [110] used log data to measure engagement and relate it to the students' grades. Fincham et al. [52] extracted several engagement measures from log data in relationship to the learning outcomes during the whole course duration. D'Mello et al. [39] proposed a model that consists of data coming from several sources, such as videos, audios, data logs, eye gaze, and other physiological measures to determine the students' state of mind (e.g., interest, boredom, mind wandering).

In this plethora of overlapping engagement and effort definitions and measures, perhaps the most distinguishable definitions are the ones proposed by Rozo [147]. In this study, engagement is defined as a bi-dimensional construct composed of the behavioral and the cognitive dimensions (which matches the effort definition proposed by Carbonaro [28]). On its turn, effort is an engagement component defined as a factor related to the actions taken to overcome a difficulty. Based on these definitions and on the work of Linnenbrink and Pintrich [108] that distinguishes the quality of effort from its quantity, Rozo [147] further proposes three indicators to assess whether a student exerted enough effort or not while creating mental models – a graph representing concepts (nodes) and their relationships (edges). The first is the ratio between the number of nodes created by the student and the number of nodes expected by the teacher, the second is the number of nodes containing the minimum requirements defined by the teacher (e.g., title, description, etc.), and the third is a pertinence indicator given by the teacher in order to assess the quality of the exerted effort.

For several reasons, we believe that the effort and engagement definitions proposed by Rozo [147] are the most interesting ones. First, they remove the ambiguity found in the literature by defining effort as a component of engagement. Second, her definition of effort does not follow the assumption “the more effort exerted, the better the learning outcomes will be”. According to this definition, if a student is only solving an exercise because he was told to do so and if he masters the required content, then the student is not exerting effort at all (i.e., there are no difficulties to

overcome), but will probably reach a high grade. Third, the engagement definition adopted by Rozo [147] allows it to be used at several levels (e.g., task level, session level, course level, etc.), while students' effort is linked to a single task. Therefore, in this thesis, engagement will refer to long-term measures (i.e., a set of tasks), and effort will refer to the task level measures (i.e., a single task). Finally, this effort definition is compatible with the mental effort definition found in the cognitive load literature (c.f., Section 3.3), which provides a few interesting properties to our study: its measurement is made considering that it is equivalent to the mental effort, it is measured at the task level, and it further provides several insights about the learning process. Therefore, we propose to rely on the cognitive load literature to measure students' effort.

3.3 Cognitive load theory

The cognitive load theory was proposed in 1988 by Sweller [167] and is based on a representation of the human cognitive architecture, shown in Figure 3.1, which is composed of the working memory (or short-term memory) that interacts with the long-term memory. The working memory can only process a limited amount of information [168, 113, 8], and its capacity is strongly influenced by individual differences (e.g., knowledge levels, self-efficacy, motivation, preferences, expectations, cultural biases, resistance to change, etc.) [113]. Hence, each individual has his own working memory capacity and, consequently, his own maximal capacity of cognitive load. On the other hand, the long-term memory is considered to have an almost unlimited storage capacity [168, 113, 8]. During the execution of a task, the working memory processes the information coming from the long-term memory and also the new information coming from visual and auditory sources, creating the cognitive load. More formally, the cognitive load is a multidimensional construct representing the load imposed on the working memory during the execution a cognitive task, or the amount of cognitive resources being used [129, 88].

In the cognitive load theory, the interactions between the working memory and the long-term memory are of the upmost importance to understand how learning takes place and how complex problems are solved [8]. According to the theory, learning takes place when the student receives new information, processes this information in the working memory to create learning schemes, and then stores it in the long-term memory. Thus, it is important to control the amount of cognitive load imposed by this task because higher levels of cognitive load makes learning more difficult due to the lack of cognitive resources available to process the new information in the working memory [8]. This lack of cognitive resources is also known to decrease the performance of subjects in other types of tasks, such as driving, traffic control, etc. [31].

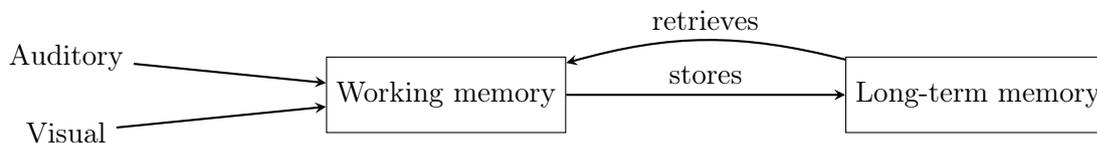


Figure 3.1: Human cognitive architecture

In the early developmental stages of the theory, cognitive load was considered to be the total working memory load experienced during learning [167]. However, as more studies were completed, the theory evolved and three different types of cognitive load have been identified [7]:

1. **Intrinsic cognitive load:** It is caused by the intrinsic nature of the materials to be learned and is dependent upon element interactivity, which dictates how complex the learner finds

the materials [8]. For learners with low levels of prior knowledge, element interactivity will be high for novel materials, and hence high levels of intrinsic cognitive load will be generated [8]. In contrast, learners with more expertise (high prior knowledge) will experience less element interactivity and corresponding intrinsic cognitive load [8]. For instance, this is the case of individuals who are learning how to read and individuals who already mastered this skill [88]. Beginners process every word as a collection of separate letters, meaning that each letter is an element to be manipulated in the working memory. On the other hand, a more advanced reader is capable of processing whole words (and even phrases) at a time (i.e., each word or phrase is an element to be manipulated). This means that the load the same text imposes is different for each reader according to their personal characteristics.

2. **Extraneous cognitive load:** It is the load caused by the processing of unnecessary information, which is created by the instructional designer and is not intrinsic to the materials [8]. A typical example of a situation that results in extraneous cognitive load is when the learners need to split their attention (and working-memory resources) between two or more related sources of information (e.g., information located in different pages or in different places of the same page) [88].
3. **Germane cognitive load:** It was originally considered to be the load directly invested in learning. However, it is now considered to be the resources devoted to process the information relevant for learning [8].

All these types of cognitive load combine to form the *overall cognitive load* (as shown on the left side of Figure 3.2), which can occupy (or not) all of the available working memory capacity [30, 129]. These different types of load, together with their personal differences, explain why some students, despite their higher levels of effort, have low grades and other students, with low effort levels, achieve higher grades. For instance, if the extraneous load is too high, students might become overloaded, achieving low grades. On the other hand, a student who already mastered the required skills is able to handle the same information with more ease and achieves higher grades. Another example is that, while solving a complex mathematical problem, students who mastered some mathematical concepts experience lower levels of intrinsic cognitive load because they could rely on the automatic retrieval of schemes stored in the long-term memory (i.e., knowledge automaticity) [190]. Meanwhile, students who did not master such mathematical concepts have to deal with the required information without any adequate schema, and may experience a cognitive overload. This point of view is compatible with the idea that exerting lots of effort does not necessarily lead to good grades, contributing to the idea that effort can be seen both as quality and as quantity.

3.3.1 Cognitive load measurement

The cognitive load is a highly dynamic construct, and can change from second to second, even within the execution of a given task [181]. As shown in Figure 3.2, according to its fluctuations over the execution of a task, the cognitive load can be referred to as:

- **Instantaneous load:** the current cognitive load level;
- **Peak load:** the highest cognitive load level experienced during a task;
- **Accumulated load:** the total cognitive load experienced in a task, or the area under the curve represented in Figure 3.2;

- **Average load:** the average value of instantaneous load during a task, or the accumulated cognitive load divided by the time spent on the task; and
- **Overall load:** the experienced load based on all the tasks, or the mapping of instantaneous load or accumulated and average load in the learner’s brain. Since this type of load has a clear psychological basis, it is often measured by subjective ratings considering one or several tasks.

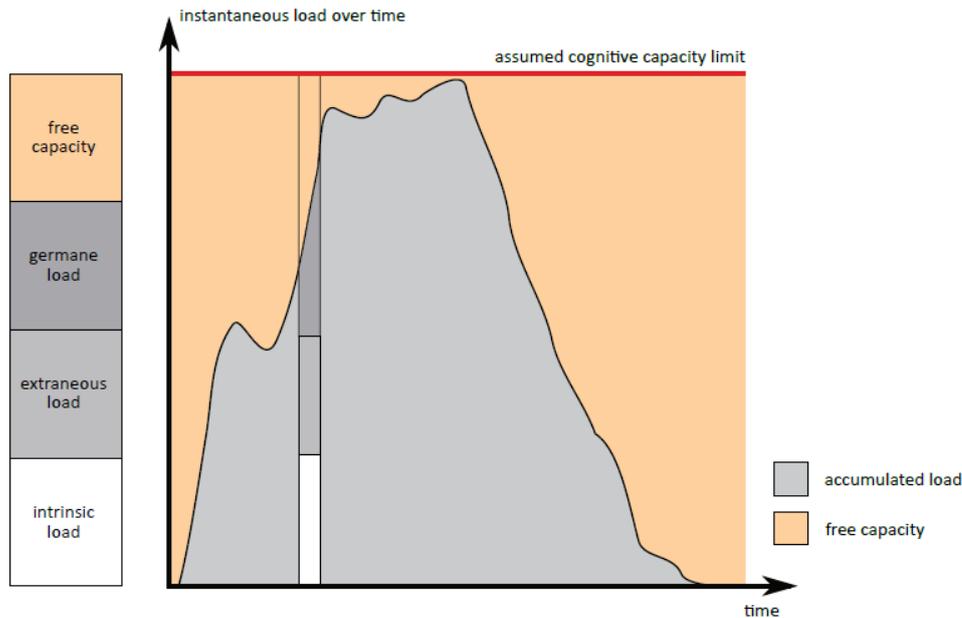


Figure 3.2: Cognitive load over time (adapted from [30])

As previously mentioned, high levels of cognitive load not only hinder learning, but also decrease subjects’ performance on their tasks. As a matter of fact, the attempts to assess cognitive load levels were mainly driven by the needs of work environments where human errors are highly undesirable (e.g., emergency response, aviation, incident management) [30]. Obviously, its measurement is not limited to these scenarios since it is considered an essential step towards the understanding of the cognitive capability, user behavior analytics, and applicable technologies that are responsive to the users’ mental status [31].

As will be shown in the following sections, the cognitive load cannot be measured directly, but there are four types of measures that can be used to assess it. Thus, we can find four types of measures to do so: subjective, performance, physiological and behavioral measures. Each one of these measures has its strengths and weaknesses. Therefore, a combination of different measures is widely encouraged in the literature [34, 120, 30].

Subjective measures

Subjective measures are a popular way of measuring the cognitive load [129, 157]. They are based on the assumption that people are able to introspect on their cognitive processes and report the mental effort exerted [105, 130], which has been demonstrated to be sensitive to relatively small differences, valid, reliable and unobtrusive [132, 129]. This approach consists in asking users to

self-assess their cognitive load by answering a set of questions in the middle of the task [157] or immediately after the task [30], being unsuitable for applications that require real-time data.

Those questions usually rely on Likert scales with five [26, 149], six [32, 71], seven [7], or nine [48, 130, 106] points. According to Chen et al. [30], these questionnaires can be unidimensional or multidimensional scales. The unidimensional scales are criticized because they capture only the overall and the average cognitive load, but it is nevertheless recognized that they are a good indicator for these types of cognitive load [30]. Furthermore, according to Ayres [8], the use of single item measure (difficulty or mental effort) has been used extensively and successfully since 1992 and contributed to probably more than a hundred studies.

On the other hand, multidimensional scales can capture several aspects of the cognitive load. For instance, the NASA Task Load Index (NASA-TLX) [71] is based on six dimensions: performance, mental effort, frustration, task demand, physical demand, and temporal demand. Other examples are the works from Cierniak et al. [32] and Leppink et al. [106], who proposed questionnaires to allow the distinction between the three types of cognitive load.

Given the simplicity of collecting such ratings, they are often used as the ground truth in cognitive load experiments [30]. However, it must be noted that their strength is in showing differences in cognitive load between tasks, and not measuring it [8]. In other words, it allows to identify which task was harder for a subject, but not to directly compare it between subjects due to its relativistic nature (i.e., what is easy for an individual might not be easy for another) and to the individual working memory differences (e.g., reading a text is easy for adults but not for children). Furthermore, this method increases the number of tasks to be executed, potentially annoying the subjects or even conditioning them to always give the same answers, which impacts its reliability [97].

As the subjective measures are often used as the ground truth in the cognitive load literature, we also use such measures as our ground truth. We acknowledge that this approach does not meet the constraint C2 (Implicit data), therefore it will not be used as input for any of the models (except for comparison purposes). Furthermore, we also considered the associated issues while collecting such measures to avoid reducing its reliability (c.f., Chapter 4).

Performance measures

Performance measures – such as time, number of errors, and grades – are based on the assumption that the experienced cognitive load will reflect on the task outcomes. More specifically, in a perfect match of theory and practice, the assumption “the higher the cognitive load, the lower the performance” holds true [132, 8]. As shown in Figure 3.3, these measures tend to remain stable when the cognitive load is in a medium level. However, when the cognitive load levels are high, the mental effort – here defined as compensatory actions/behavior – takes place in order to sustain the performance for a given time period [174], but once the subject experiences a cognitive overload, it does not make any difference and the performance can no longer be sustained.

As also shown in Figure 3.3, two subjects can achieve the same performance and experience different cognitive load levels. Therefore, the performance cannot be directly used to infer the experienced cognitive load. As a matter of fact, researchers found a negative correlation between brain activity under cognitive load and intelligence, which suggests that individuals with higher intelligence levels may process information quicker and possibly in different ways (e.g., use of better mental strategies, possibly due to differences on inhibiting irrelevant processes, and on controlling, switching, and focusing attention) [53, 63, 83]. Thus, due to these differences, the performance measures are often used with the dual-task technique [30].

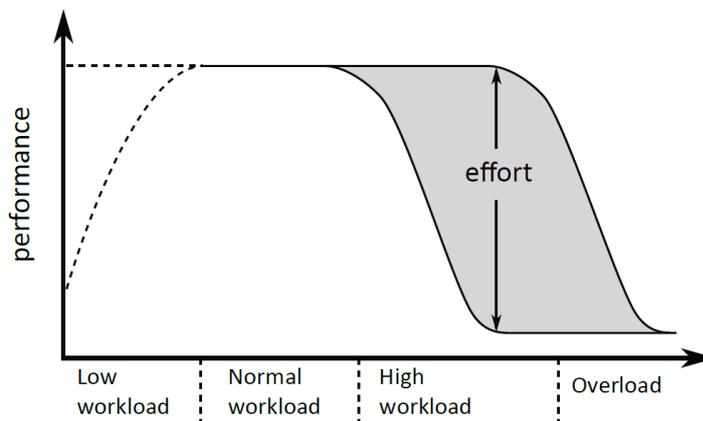


Figure 3.3: Cognitive load vs. mental effort and performance (adapted from [30])

The dual task technique consists in asking the subject to execute two tasks simultaneously and to analyze his performance on the secondary task. If the performance on the secondary task decreases, it means the tasks interfere with each other and are competing for the same type of mental resources [30], and that the primary task is demanding more cognitive resources from the subject. In other words, differences in a student's resource consumption can be measured by performance differences on the secondary task [133]. This method poses the challenge of designing adequate secondary tasks (e.g., that shares cognitive resources with the primary task, but does not demand more resources than the main task). Furthermore, it does not allow the measurement of temporal aspects of cognitive load (e.g., cumulative and instantaneous load) and its use in real-life scenarios is not feasible [30, 190].

In a different approach, Paas and van Merriënboer [131] argue that performance measures combined with other cognitive load measures allow a meaningful interpretation of the instructional conditions. According to them, a high performance with low effort means high-instructional efficiency, whereas a low performance with high effort means low-instructional efficiency.

The cognitive load approach regarding the performance measures corroborates Rozo's [147] point-of-view about the quality of students' effort and highlights the drawbacks of measuring students' effort assuming that more effort leads to better grades (i.e., the amount of effort changes according to other factors such as previous knowledge). Although the dual-task paradigm cannot be applied to our study because it does not respect C3 (Real life application), the performance measures and their relationship with the cognitive load still can, and were made available in our dataset (c.f., Chapter 4). More specifically, the relationship between effort and performance enable us to roughly distinguish four different scenarios, as shown in Figure 3.4:

1. **Student is learning (in green):** This scenario is the most important one as it represents the cases where students exert lots of effort, overcome their difficulties and successfully complete their task (high grade). This can be interpreted as if the student has learned something new, or at least practiced his skills.
2. **Student is struggling (in yellow):** In this scenario, students exert lots of effort but do not overcome their difficulties (low grade). Some possible reasons for this outcome is that the task was not adapted to the students because it required some knowledge they did not have or was ill-structured.

3. **Student has previous knowledge (also in yellow):** In this scenario, students do not exert lots of effort, but still achieve good grades. This can have two interpretations: the positive interpretation is that they already mastered the required skills, and the negative is that they successfully guessed the answer or that they copied the answers from their classmates.
4. **Student is disengaged (in red):** This scenario represents the cases where students do not exert lots of effort and, as a consequence, do not achieve good grades. Possible reasons for the lack of engagement with the task are related to the task boredom or lack of appeal, to the student identifying he has not enough knowledge to execute it, etc.

The ideal scenario is the one where students are learning (green scenario), while the worst scenario is the one where they are disengaged (red scenario). On the other hand, the intermediate scenarios (yellow scenarios), those in which students are struggling or have already mastered the content, call for a trade-off. Sometimes it might be preferable to propose a task where the exerted effort will be low, but the grade will be higher in order to make the student practice without demotivating him; but sometimes it might also be interesting to propose a task that requires a high effort, even though he struggles, to push him to study different contents or to seek help. Furthermore, between all of these dimensions exists a blurry line because it can be hard to say what exactly composes a good and a bad grade/effort, and because sometimes the students are at the frontier of two different scenarios.

	low grade	high grade
high effort	student is struggling	student is learning
low effort	student is disengaged	student has previous knowledge

Figure 3.4: Scenarios based on the relationship between effort and grades

Behavioral measures

Behavioral measures capture objectively and implicitly the subjects' deliberate/voluntary activity [30]. The rationale of this type of measures is that the human behavior changes according to the situation and subjects' interactions with the world can reflect it [31]. In other words, a subject experiencing high levels of cognitive load shows symptoms related to the management of such load [31].

According to Chen et al. [31] the mouse clicking and the keyboard keypressing behavior can be an ideal choice for cognitive load measurement. For instance, the movements made with a

mouse might reflect changes in cognitive processes. According to the study from McKinstry et al. [114], the mouse usage presents a greater curvature and the lowest peak velocity when the right answer was ambiguous or more complex. Using other types of peripherals, Dale et al. [35] asked the participants to classify some items using the Nintendo[®] Wii[™]. The results show that participants' arm movements were faster and smoother after they learned the content. Another related study is the one from Arshad et al. [6], who analyzed the relationship between pauses in the mouse activity and cognitive load. Their results show that subjects experiencing higher cognitive load tend to make more and longer pauses. The use of a digital pen (another input peripheral) can also be tracked to infer cognitive load levels. For instance, Yu et al. [184, 185] show that the writing pressure, the writing velocity, the pen orientation, the pen-tilt, and the curvature involved in the written strokes information can indicate different cognitive load levels.

Another way to infer the cognitive load through behavioral measures is by tracking subjects speech and extracting speech and linguistic features. Khawaja et al. [92] carried a study to investigate the suitability of speech features (e.g., length of silent and filled pauses, frequency of silent and filled pauses, and response latency) to measure cognitive load. Participants were asked to read a story out loud and to answer a few questions about it. The results show that silent and filled pause lengths, and response latency are significantly higher when the cognitive load levels are higher. In a similar study, Yin et al. [183] used spectrum features measured by mel-frequency cepstral coefficients, and prosodic features measured by fundamental frequency (or pitch) and speech intensity as inputs to a gaussian mixture model, achieving a correlation coefficient of 71.1% with the subjective ratings. Linguistic features – which can also be extracted from texts written by the subjects – can be found in the work of Khawaja et al. [91]. They analyzed participants' linguistic features during a firefighting task using a multitouch tabletop screen. From the audio transcripts, they extracted several linguistic features, such as word count, words per sentence, negative emotions, positive emotions, swear words, etc. Almost all of these features increased together with the cognitive load.

In a different approach, Verrel et al. [175] studied the effects of cognitive load on gait patterns during a treadmill walk based on the assumption that the experienced cognitive load influences on the sensorimotor responses. They used the residual variance (i.e., the relative amount of variance in the residual pattern) from the principal component analysis method to measure how regular the subjects' body movements were. They found different relations between gait and cognitive load on different age groups. Gait was more regular (i.e., reduced residual variance) on those whose age was 20-30 years old, more irregular (i.e., increased residual variance) on those whose age was 70-80 years old, and inconclusive in those whose age was 60-70 years old.

Other behavioral measures that have been explored are head movements and mouth openness. According to Guhe et al. [69], when the cognitive load increases the subjects tend to move their heads and also to open their mouth more often.

Probably one of the main advantages of some of the behavioral measures presented (i.e., keyboard, mouse, digital pen usage) is its collection ease. As a matter of fact, such measures are widely used in the learning analytics domain (e.g., several of the previously described effort and engagement measures). Therefore, this is a measure that fully meets constraint C3 (Real life application).

Physiological measures

Physiological measures are based on the assumption that changes on the psychological state lead to a physiological change [129]. In other words, the increase of the experienced cognitive load affects body properties (e.g., temperature, heart beats, etc.) [99, 8]. These measures have been

used for decades to assess the cognitive load [4] and might be the most diversified approach to do so [190]. The explored measures are related to the brain activity, to the cardiovascular system, to the nervous system, to the ocular system, and to the skin.

Brain activity measures can be categorized as high-spatial resolution methods that track slow changes in the brain, such as functional magnetic resonance imaging (fMRI) and functional near-infrared spectroscopy (fNIRS); or as tools that provide high temporal resolution and detect fluctuations in the electrical activity of the brain, such as electroencephalography (EEG) that is often used to derive event-related potentials (ERP). As these measures cannot be captured by largely available equipment, they are not relevant for our study and, therefore, we will not further describe them.

Other type of physiological measures are those related to the cardiovascular system, such as the heart rate, the heart rate variability, and the time between each heart beat (also called R-R interval or inter-beat interval). Some studies show that when the cognitive load increases, the heart rate increases, but the heart rate variability decreases [120, 90].

These measures could be affected by blood pressure variations, which are themselves affected by factors such as sleep deprivation, ambient noise, and respiration (which can also be influenced by other factors, such as speech) [120]. Therefore, Mulder [120] argues that these measures should not be the only cognitive load measures adopted. Another issue of cardiovascular measures is that some researchers do not consider them valid because they are intrusive and insensitive to subtle fluctuations of cognitive load [132, 127, 124]. However, this measure can no longer be considered intrusive given the emergence of wearable technologies and the fact that several smartwatches available can measure the heart rate. Some studies such as the ones from Marandi et al. [186] and Phan et al. [136] show positive results regarding this aspect.

Skin response measures can also be used to measure the cognitive load. They refer to skin temperature and electrical conductance (also called skin conductance, galvanic skin responses and electrodermal activity). For instance, Or and Duffy [128] conducted experiments to assess the correlation of nose and forehead skin temperatures with the cognitive load. The results showed that the drop of the nose temperature was related to a higher level of cognitive load. Another related study is the one from Shi et al. [157]. They analyzed the total amount of skin electrical conductivity that changes due to sweat secretion (i.e., more sweat, more electricity) [36]. Their results show that the skin conductivity increases when the cognitive load increases.

Probably one of the most popular cognitive load measures is related to the eye activity. Some measures obtained through this method are related to the eye blinks, to the eye gaze (fixations and saccades¹), and to the pupil dilation [118, 34]. Blink frequency and blink duration decreases as the cognitive load increases; while blink interval, fixation frequency, fixation duration, saccades distance, the percentage change in pupil size, the mean pupil dilation, the peak dilation and the latency to the pupil dilation peak increase when the cognitive load increases [118, 104, 34, 8, 37, 14, 173].

However, it is important to note that the task-evoked pupillary responses are sensitive to changes in lighting (e.g., computer screen, sun, room lighting), dilating in its presence [99, 34]. For instance, Schultheis et al. [153] found out that, while reading some texts in a hypertext environment, the pupil size did not change when the text was easier or harder. Furthermore, pupil measures are also responsive to fatigue, emotional arousal, and pain making it hard to tell to what extent the captured changes are related to the experienced cognitive load [34]. Another

¹Fixations happen when the eye is kept aligned with something for a given time, allowing for the image details to be processed. On the other hand, saccades are an eye movement that happens between fixations to change from one point of interest to another [170].

disadvantage of such measures is that, although they can capture cognitive load changes over time, they do not capture the overall cognitive load [34].

The eye activity measures are usually captured through eye trackers, but they can also be obtained through electrooculography (EOG) devices, and video records. Even though eye trackers and EOG devices are still expensive and sensitive equipment, which makes their use in real-life scenarios complicated [34], the eye tracking technology is becoming mainstream, increasing the population access to it and the feasibility of its use. As a matter of fact, it is already possible to buy eye trackers to play games. Furthermore, as already mentioned, these measures can be captured through video recording (although maybe not with the same quality), which can be easily captured.

The described physiological measures are prominent in the cognitive load literature because they can be captured at a high rate and with a high degree of sensitivity [30, 129, 190]. However, when we consider their feasibility for real-life scenarios, they become considerably less interesting due to their specificity (i.e., their usefulness is restricted to research and health facilities), high cost, size, and their obtrusiveness [29, 190]. Fortunately, some of these measures can already be captured through wearable technologies (e.g., smartwatches, smartshirts) and other accessible devices (e.g., smartphones, webcams). Therefore, as will be explained in Chapter 4, our dataset includes eye activity, pupillary response, and heart rate data.

3.3.2 Cognitive load in educational scenarios

As previously stated, the measurement of cognitive load advanced mainly due to the need of work environments where errors must be avoided at all costs. However, we can also find related works in the educational scenarios and, in this section, we review recent studies carried in this scenario.

One of these studies is the one from Larmuseau et al. [103]. They investigated cognitive load differences in well- and ill-structured problems using subjective ratings, physiological data (i.e., skin temperature and skin electrodermal activity) and consultation of support resources collected from 15 future primary school teachers who were between 18 and 24 years old. Their results indicate that there are in fact differences in cognitive load levels when considering these two scenarios. However, this work does not attempt to identify different cognitive load levels, which is our main goal regarding the students' effort.

On the other hand, Mock et al. [117] sought to identify different cognitive load levels. For that, they used touch sensor data (i.e., interaction patterns while using a multitouch device) from 30 fourth grade students to predict user states (defined according to the task difficulty) on 96 math problems. Their results show that the touch patterns can predict high cognitive load with an average classification accuracy of 90.67%. However, as their cognitive load levels were defined according to the task difficulty, it does not reflect the participants' experienced cognitive load, but rather the experts evaluation of each task. This implies that their effort measurement model is not personalized and, therefore, does not seem adapted to be used in an educational recommendation system.

In a similar approach, Spüler et al. [162] trained a linear ridge regression model with brain measurements collected from EEG sensors from 10 students (17-32 years old) in order to identify the task difficulty level of performing 240 addition problems. Two approaches were considered: cross- (trained the model with data from a group and used a single model for everyone) and within-participant (each subject has his own model). The cross-participant approach reaches a correlation value of 0.82 and an rooted mean square error (RMSE) of 1.34, while the within-participant achieved a higher correlation value (0.90) and a smaller RMSE (0.95). However, as

discussed in another study [179], even though the within-participant has a better performance, it might be preferable to use the cross-participant models because they are easier to train and require less time to deploy. In a further study, Walter et al. [178] exploited these models to provide adaptations in a learning environment. They evaluated the adaptations by asking participants (ten students between 17 and 32 years old) to solve arithmetic additions in the octal number system, and the results show that students' performance increased from the pre-test to the post-test when using their adaptive system. However, they use the same criterion to distinguish between different cognitive load states (i.e., the task difficulty), which as already discussed, does not reflect the participants' experienced cognitive load. Furthermore, they used EEG devices, which do not respect constraint C3 (Real life application).

In a different approach, Yen [182] presented an analytical method (i.e., a method to estimate cognitive load without empirically measuring it) for measuring the intrinsic cognitive load of learning content based on the level of element interactivity and on a graph. This graph relates each course module to its required modules (e.g., a link from node X to node Y implies that element X must be learned before element Y) following the assumption that an element X needs to be loaded into the working memory when learning Y. In other words, this model considers that the more requirements a learning module has, the higher is the intrinsic cognitive load. Just like the previous studies, this work does not reflect the participants' experienced cognitive load.

Borys et al. [21] collected the data – i.e., eye gaze, pupil size and EEG – of twenty male graduate students (computer science Master degree) with an average age of 22.8 years old during arithmetic exercises. They retained the data from 13 of the 21 participants, and classified it following two approaches. In the first approach they classify the data into two categories – no task vs. cognitive load – and achieved an accuracy of 90.4% with the support vector machines classifier. In the second approach they classified the data in three categories: no cognitive task, low cognitive load and high cognitive load. In this second approach, their accuracy ranged from 50% to 73%. Different from the previous studies, Borys et al. [21] did not rely on the task difficulty to define their cognitive load states and relied on the mean response time to distinguish between the low and high cognitive load states. However, as already discussed in this chapter, time constitutes a performance measure and, therefore, might not be suitable to measure the cognitive load. Furthermore, they used EEG devices, which, as already mentioned, do not respect constraint C3 (Real life application).

Herbig et al. [73] collected data – students' perceived effort, skin temperature, eye gaze, pupil size and heart rate – of 21 students (17 of them were male) aged between 20 and 22 years old, enrolled in a computer science-related course, and who successfully passed the mathematics lectures covering the selected topics. Participants were asked to watch some videos and answer a quiz about vectors, integration, and eigenvectors. Their approach is similar to the one used by Borys et al. [21], but instead of classifiers they used several regressors to estimate students' perceived effort. Their results show a mean squared error (MSE) considerably smaller than the baseline (a dummy regressor predicting always the mean). In comparison to our work, the main difference of their work is the equipment used. They used several cardiovascular devices in order to collect their data. This means that their approach does not fully respect constraint C3 (Real life application).

Given that the works from Borys et al. [21] and Herbig et al. [73] seek to estimate the cognitive load level, they are the ones that present the highest similarity with our study. Given this similarity and the fact that some of our data are similar, these works will be used as our baselines in Chapter 5, where their methodology and results are described in more depth.

From all of the discussed works, we can see several differences regarding our goal. First, none of the presented studies explored the cognitive load with teenage students learning a foreign

language. This emphasizes the need to validate the effort measures in this context since the generalizability of the cognitive load measures to different contexts has already been questioned [125]. Second, all of the previously described studies have a small number of participants, which might not be too representative (as will be explained in Chapter 4, we carried an experiment with 120 students and retained the data of 102). Third, none of the described studies assess the cognitive load through web log data (as will also be explained in Chapter 4, we investigate the use of such measures). Fourth, several studies use specific equipment, such as EEG devices, which does not respect C3 (Real life application) and, therefore, the results cannot be exploited in real-life scenarios. Finally, only one of the presented studies exploits their models in virtual learning environments, but using EEG-related data, which as already discussed, is not suitable for real-life scenarios. Finally, in this thesis, we are not only interested in training models to identify different cognitive load levels, but also in exploiting these models to propose a recommendation system that formalizes the foot-in-the-door technique.

3.4 Chapter conclusion

In this chapter, we reviewed the human factors linked to our goal, namely, the theory of commitment to which the foot-in-the-door is related, and the concepts of students' engagement and students' effort that pose some challenges regarding their definition and measurement due to the overlapping definitions found in the literature.

After reviewing the literature, we adopted the following engagement definition: a multidimensional construct with behavioral, emotional and cognitive dimensions. Effort is further defined as a factor linked to the actions taken to overcome a difficulty [147], a definition that is compatible with the mental effort found in the cognitive load literature: a compensatory behavior to handle higher amounts of imposed cognitive load. On its turn, the cognitive load is defined as "the amount of cognitive resources required by a task".

We further defined that the engagement is a long-term construct (e.g., a set of tasks executed during a session, a course, a week, a month, a year), while effort is a short-term construct related to a single task.

As the cognitive load is considered equivalent to the mental effort and is a widely studied construct, it becomes a relevant construct to our study. It is particularly interesting because it provides insights about how to measure students' effort on the task level, and offers explanations as to why some effort measures (e.g., time) do not fully relate to the learning outcomes. For instance, the related literature establishes a link between the cognitive load and performance measures (e.g., time, grades), describing the issues that can arise from measuring students' effort assuming that more effort leads to better grades. As the exerted effort can change according to other factors, such as previous knowledge, it means we can have good grades exerting a little effort and bad grades exerting lots of effort. This limitation highlights the differences between the quality and the quantity of effort. According to the performance and to the effort we can roughly distinguish four different scenarios: students who are disengaged, students who are struggling, students who mastered the content, and students who are learning.

This relationship emphasizes the need to recommend learning resources that require an appropriate level of effort. In a scenario where the adequate learning resource difficulty (e.g., beginner, intermediate, expert) always corresponds to a medium effort, this would mean that recommending learning exercises that require little effort would probably result in disengaged students or in students' using resources with a content they already mastered. On the other hand, recommending resources that require lots of effort would not necessarily make students learn. Instead of

encouraging learning, the recommended resources could make students struggle so much on the long run that they would eventually become disengaged because they would never achieve the expected outcome. Thus, being able to properly infer students' effort in past and future tasks becomes even more important to address students' personal characteristics and choose the more suitable learning resources for each one of them, avoiding to make harmful recommendations that would have a negative effect on their learning.

Inferring the amount of effort a task requires from a student can be further considered as a personalization issue, given that each student has its own characteristics and will perceive the effort they exerted differently. Therefore, the personalization requisite established in Chapter 2, might be met by properly estimating the students' perceived effort in past and future tasks. Another previously established requisite that we intend to tackle by taking into account the learning science review presented in this chapter is the need to foster learning. In this thesis, we propose to formalize the foot-in-the-door technique to engage students into learning, assuming that by engaging students they will learn more. The literature review carried in this chapter corroborates this assumption, as both engagement and effort are considered key elements to learning.

Having reviewed the learning science side of the learning analytics domain we were able to identify some effort measures feasible to real-life scenarios by relying on the cognitive load literature. However, relying on the cognitive load construct to measure the students' effort poses some challenges. First, the subjective ratings often used as the gold standard indicate whether the amount of cognitive load increased from one task to another. It means that these ratings do not measure the cognitive load (in a sense that the measure is exactly the same for every person, e.g., metric system) and are different for each subject. Second, the physiological and behavioral measures are still being investigated in order to measure the cognitive load and sometimes their generalizability to other domains (e.g., from driving tasks to learning tasks) is questioned [125]. Third, these techniques might require large and expensive equipment, and controlled settings which is not suitable for educational settings and do not respect C3 (Real life application). Finally, such measures are still not validated in real-life scenarios to measure teenage students' effort during foreign language tasks. Therefore, a dataset respecting the constraints of our study is, to the best of our knowledge, non-existent. For this reason, we built our own dataset, which is presented in the next chapter.

In this chapter we presented the theory of commitment and the foot-in-the-door technique (and its variants). The foot-in-the-door technique consists in making consecutive requests with an increasing cost, which we defined as the students' effort.

We defined the students' effort as a component of the students' engagement, and we also consider it to be equivalent to the cognitive load construct.

The cognitive load literature provides valuable information regarding how to measure the students' effort (i.e., subjective, performance, behavioral and physiological measures) at the task level.

Chapter 4

Multimodal data collection and analysis

In the previous chapter, we identified four different categories of measures that can be used to estimate the students' effort, namely, subjective, performance, physiological and behavioral measures. As we could not find any datasets that completely matched our needs and the context of our study, we decided to collect our own data. At the same time, we performed a preliminary study with a dataset that partially met our objective.

This chapter describes this preliminary study in Section 4.1, and the construction of our dataset in Section 4.2 detailing the experimental protocol, the execution, and an overview of the resulting dataset. Finally, after an ethical disclaimer in Section 4.3, Section 4.4 concludes this chapter by presenting our final thoughts.

4.1 Preliminary study

Before we could start our own data collection, we undertook a preliminary study based on another dataset that has been collected by our research group in the context of a study about the link between gaze and memory [112]. This study involved 14 participants, between 18 and 40 years old, who were invited to learn the basic concepts of the Esperanto language on which they had no prior knowledge. Each participant was invited to browse different course pages with no time limitations and, only when he was satisfied, he was directed to an evaluation questionnaire with 21 questions (11 sentences to translate and 10 multiple-choice questions). During the entire session, the participants were tracked by an eye tracker (Tobii X1 Light) and its software (Tobii Pro Lab). The final dataset contains behavioral, physiological and performance data.

We considered all of the tasks the participants executed as being a single task and extracted a few indicators from the collected dataset: time spent on task, total number of page views, total number of clicks, total number of keystrokes, total time of fixations, and total number of fixations. We then combined those indicators into: the average time of fixations, average time between fixations, average time between each page view, average time between clicks, average time between keystrokes, average time between actions (i.e., clicks and keystrokes together), and a combination of the interaction indicators into a new indicator called weighted actions and computed as shown in Equation 4.1, where $\#Hits$ stands for the number of page visits, $\#Clicks$ stands for the number of clicks, $\#Keystrokes$ stands for the number of keystrokes. It is worth mentioning that we sum 1 to the number of clicks and keystrokes in order to always have a value higher than 0, even if the student does not click or type, in order to make pages visited more

important than pages that were not visited.

$$\text{Weighted actions} = \# \text{Hits} \times (\# \text{Clicks} + \# \text{Keystrokes} + 1) \quad (4.1)$$

Even though we discussed in Chapter 3 that the grades are not the best ground truth to use, we assume that in the context of this preliminary study, the grades are a reliable effort measure because participants had access to completely new learning contents and had to exert effort to learn it. More specifically, we assumed that as the participants had no prior knowledge in Esperanto, the results generally do not suffer from the influence of the students who already mastered the content. Therefore, to study how students' effort can be measured and modeled using these data, we analyzed the Spearman's correlation coefficient ρ of the raw and combined indicators with the participants' grades following the assumption of the engagement/effort literature: "the more effort exerted, the better the learning outcomes will be".

The correlations between all of our indicators and the final grades obtained by the participants are shown in Table 4.1. As can be seen, all of the raw indicators have a weak correlation with the grades, while some of the other indicators have medium correlations ranging from 0.35 to 0.54. As the combined indicators have a greater correlation with the grades than the raw indicators, this corroborates the idea that combining different measures is a more accurate manner to estimate the students' effort (c.f., Chapter 3).

	Type	Indicator	ρ
Raw indicators	Interaction	Time spent on the task	0.21
		Number of page hits	0.28
		Number of clicks	0.14
		Number of keystrokes	0.10
	Eye gaze	Fixations duration	0.00
		Number of fixations	-0.08
Averaged indicators	Interaction	Average time between page hits	0.05
		Average time between clicks	0.35
		Average time between keystrokes	-0.27
	Eye gaze	Average time between fixations	0.36
		Average fixation duration	0.54
Weighted indicators		Weighted actions	0.35

Table 4.1: Correlation between the extracted indicators and the grades

We then proceeded to create an effort model by choosing one interaction and one eye gaze indicator to combine, expecting them to lead to an even stronger correlation with the grades. The chosen eye gaze indicator was the average time between fixations, and the chosen interaction indicators were the average time between clicks and the weighted actions (we could not choose only one indicator because both of them have the same correlation coefficient). As each one of these indicators has different ranges and distributions, we normalized them through an exponential function, presented in Equation 4.2, with an upper limit of 1 before combining them according to the model shown in Equation 4.3. In this model, two weights w_1 and w_2 were defined in order to optimize the importance of each indicator.

$$\text{norm}(i) = 1 - \exp^{-\alpha i} \quad (4.2)$$

$$\text{Effort} = w_1 \times \text{norm}(i_1) + w_2 \times \text{norm}(i_2) \quad (4.3)$$

The combination of the average time between clicks with the average time of fixations did not present any improvements since the highest correlation achieved (0.54) is still equal to the correlation of the eye gaze indicator alone. On the other hand, the combination of the weighted actions with the average time of fixations – with $w_1 = 0.7$ and $w_2 = 0.3$ – did increase the correlation to 0.59. Therefore, this model allows to estimate the students’ effort assuming that the more effort exerted the better the grades will be.

Even though this preliminary study suggests that physiological and behavioral measures can be used to measure the students’ effort when using the grades as the ground truth, the same grade can be obtained with different levels of effort, which highlights the need to rely on other measures as the ground truth, such as the subjective ratings, which are considered a valid and reliable measure in the cognitive load literature. Furthermore, as already described, the dataset used in this study contains the data of adults and not of teenagers which does not match the context of the e-FRAN METAL project. As discussed in Chapter 2, adults and teenagers present different behaviors.

4.2 Multimodal data collection

In order to collect subjective, physiological, behavioral and performance data from secondary school students doing foreign language activities, we ran a few data collection sessions. The Nancy-Metz rectorship services, more specifically the DANE (Délégation Académique au Numérique pour l’Éducation), an academic service to allow access to digital tools, was responsible for contacting the schools and asking for their cooperation. Five schools agreed to participate in our study and invited their 7th grade students – who have English lessons at the school – to participate. Eight data collection sessions were scheduled with these schools, which, in our full capacity (6 sessions per day, with 4 students in each), could allow us to collect data from up to 192 students. From all of the invited students, 120 agreed to participate.

The following sections describe the experimental protocol adopted during the data collection sections, how the data was collected, and provide an overview of the dataset we built.

4.2.1 Experimental protocol

All of the data collection sessions were held in the students’ school. The steps followed during each data collection session, with each participant, is shown in Figure 4.1, and has the following phases:

Equipment setup: In this phase, we invited students to create their unique identifier (in compliance with C1 (Privacy and ethics)), wear a smartwatch in their dominant hand and calibrate the eye tracker.

Initial presentation: In this phase, we presented the experiment phases to the students, asking them to answer the initial questionnaire, solve the proposed tasks at their own pace, assess their effort in it and, at the end of the session, to answer a final questionnaire. We also emphasized that the most important part was not their final grade, but their honesty while answering the questionnaire.

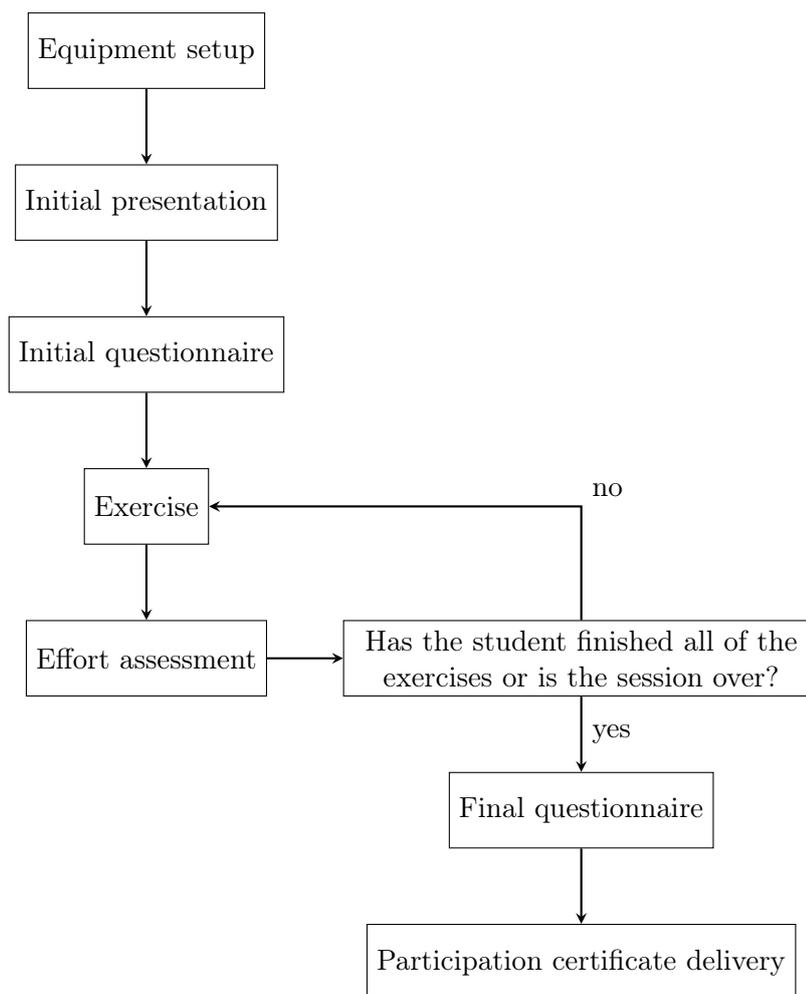


Figure 4.1: Phases of data collection session

Initial questionnaire: Students were invited to answer a few initial questions (c.f., Appendix C, Questionnaire 1) in order to provide us with some personal information (i.e., gender, age, feelings towards learning English, tiredness level). Even though the subjective data does not respect C2 (Implicit data), we only used for comparison purposes.

Exercise: Given the context of the PIA e-Fran METAL project, we proposed 15 English exercises (c.f., Appendix A) with different levels (i.e., five A1+ exercises, five A2 exercises and five B1 exercises) defined according to the Common European Framework of Reference for Languages. The first 5 exercises contained the answers that should be used in each field (e.g., “fill in the fields below with the following expressions”), exercises 6 to 10 made it clear what should be done (e.g., “conjugate the verb between parentheses in the past”), exercises 10 to 15 were more vague in their descriptions (e.g., “complete the sentences below correctly”). Each group of 5 exercises was ordered according to the difficulty level previously defined by the editor (A1+, A2 and B1).

These exercises are available in an online platform called BRNE (Banque de Ressources Numériques pour l’École), and were made available to us by the DANE. As the BRNE is managed by Eduscol, the official French website to support educators, and its exercises are designed to

secondary school students (in France, le cycle 4), we believe they can be considered relevant, reliable and suitable to our study.

We were aware since the beginning that students might not be able to solve all of the proposed exercises and we chose exercises from all levels on purpose in order to allow us to analyze students' effort on exercises with different characteristics (e.g., content studied vs. content not studied, easy exercises vs. hard exercises, drag n' drop exercises vs. fill the blanks exercises, etc.), even in exercises that might not be completely suitable for them as it is important to distinguish whether our recommendation system can identify those exercises as well.

Effort assessment: Students assessed their effort in each one of the proposed exercises through the question *How much did you have to think to solve this exercise?* using a Likert Scale that ranged from 1 (very, very little) to 7 (very, very much) (c.f., Appendix C, Questionnaire 2).

This question was adapted from the work of Salomon [150] and, as shown by Leppink et al. [106], it does not distinguish between intrinsic, extraneous and germane load. Note that in our study we do not intend to distinguish between different types of load or between different related dimensions such as the ones present in the NASA-TLX questionnaire [71] and, for this purpose, the unidimensional scale is relevant and has been used extensively and successfully (c.f., Section 3). We also wanted to ensure the participants would not be annoyed by long questionnaires between the exercises, avoiding to condition them to always give the same answers and/or to skip the assessment questionnaires.

When we adapted this question to our study, we reduced the number of Likert scale points from 9 to 7 because we believe that, given the participants' age, it would be too many points to distinguish from. As a matter of fact, it has been reported that children tend to answer Likert scales using the extreme points [66]. We also changed the wording of the original question – *How much did you concentrate while watching/reading?* – in order to adjust it to the context (i.e., solving exercises and not watching/reading).

Final questionnaire: At the end of the session (i.e., when the time was over or when the student finished the 15 exercises), we asked students to answer a final questionnaire (c.f., Appendix C, Questionnaire 3). This questionnaire comprised questions related to the students tiredness levels at the end of the session and to his feelings during the session.

Participation certificate delivery: We then delivered a participation certificate (c.f., Appendix D), thanking the students for participating in our study. The main purpose of this certificate was to provide students with their unique identifier and our contact to withdraw their authorization (in compliance to C1 (Privacy and ethics)).

Each school imposed its own constraints to the sessions. For instance, some schools reserved slots of 45 minutes for each session, while other schools reserved slots of 1 hour. Furthermore, the first session of the day was smaller because we had to unpack all the equipment and prepare it, but the first students arrived at the same time we did.

4.2.2 Data collection

Besides the questionnaires students answered during the data collection session – which were already presented –, we also collected implicit data by tracking students activity and their physiological responses with the following instruments:

- **Activity logger:** The activity logger, a script developed by us to track students interaction, captured the keyboard and mouse usage, and the pages students accessed and their timestamps.
- **Smartwatch:** All students were monitored by a TicWatch E smartwatch, that allowed us to capture students’ heart rate, the applied force of gravity, the rate of rotation, the acceleration force and the linear acceleration data through an application we developed, which simply captures the signals provided by the sensors and stores it in a file for further analysis.
- **Eye tracker:** The eye tracker allowed us to capture students’ eye activity (i.e., fixations, saccades), their pupil diameter, their click and keystroke data. We had access to four different settings of equipment and software as shown in Table 4.2. This allowed us to collect more data and to observe how using different models influences our final effort models in a real-life scenario where each user has his own hardware and software configuration. We acknowledge that eye trackers might not be an everyday life device disrespecting C3 (Real life application), but as mentioned in Chapter 3, gaming eye trackers are available in the market and alternatives allowing to capture such data through a webcam already exist.

All of the eye trackers and softwares are from Tobii, a company that develops and sells products for eye control and eye tracking, and have a sampling rate between 30Hz and 60Hz. As a higher frequency provides more fine-grained data (i.e., it collects more samples per minute), we prioritized the use of the eye trackers with the higher frequency.

#	Eye Tracker device	Eye tracker software	Sampling Rate	Participants
1	Tobii Nano	Tobii Pro Lab 1.111.19220	60Hz	39
2	Tobii X2-60	Tobii Studio 3.4.8.1348	60Hz	33
3	Tobii X2-60	Tobii Studio 2.0.5	60Hz	16
4	Tobii X1 Light	Tobii Studio 3.1.0	30Hz	28

Table 4.2: Data collection settings

After the data collection sessions were carried out, we invited the students’ teachers to assess the proposed exercises regarding their difficulty (i.e., easy, medium, hard) and their suitability (e.g., content taught or not) by answering a questionnaire (c.f., Appendix C, Questionnaire 4). Unfortunately, none of the teachers answered the questionnaires, and we asked an English teacher from a school that did not participate on the data collection sessions to answer it in order to have at least some information.

From the 120 students who agreed to participate in our study, we had to remove the data from 18 students due to several issues, such as no eye tracker data available, missing answers to the initial questionnaire, eye tracker connection issues, smartwatch placed in the wrong hand, etc. Those issues left us with the data of 102 students, from which we extracted 126 effort indicators that cover all of the four types of measures discussed in Chapter 3 (i.e., subjective, performance, behavioral and physiological). The complete list of computed indicators is presented in Appendix E.

Some of these indicators were inspired by the learning science literature and others by the cognitive load literature. The indicators inspired by the learning science are related to students’ interaction with the system, more specifically to the clicks, time between clicks, keystrokes, time between keystrokes, visible keystrokes (i.e., keystrokes that are not Ctrl, Shift, Alt, etc.), time

between visible keystrokes, backspaces, time between backspaces, actions (i.e., the sum of clicks and keystrokes), time between actions, and the number of attempts to solve an exercise during a single page visit. The other indicators, inspired by the cognitive load literature, are related to the perceived effort ratings, heart rate, pupil diameter, fixations, fixations duration, time between fixations, saccades length, saccades speed. We also have some metadata related to the exercise (e.g., type, difficulty) and to the student (e.g., gender, age).

This means that our dataset has all of the four types of data found in the cognitive load literature (c.f., Chapter 3). Furthermore, with the exception of the subjective effort ratings, all of the data captured could be collected using everyday devices, respecting all of the constraints imposed, namely, C1 (Privacy and ethics), C2 (Implicit data) and C3 (Real life application).

4.2.3 Initial analysis

Our final dataset contains data of 102 participants and the 126 effort-related indicators. In our initial data analysis, we excluded a few cases from our dataset:

- All of the data samples without a subjective effort rating;
- All of the data samples without an answer (identified through the number of clicks in the case of the drag n’ drop exercises and by the number of keystrokes in the case of the fill the boxes exercises); and
- All of the data samples related to an exercise solved in two different page views. More specifically, as we have two samples related to the same student in the same exercise, but only one effort rating, we could not create a link between their activity and the provided effort ratings without potentially biasing our results.

We started our analysis by an overview of the students’ demographics and overall activity. We then moved to analyze the differences in the students’ perceived effort, grades and time spent on each exercise in order to obtain a few more insights regarding the three main effort measures found in the literature (c.f., Chapter 3). We then proceeded to analyze how well the extracted indicators correlate with the students’ perceived effort and with their grades. All of these analysis are described in the following sections.

Data overview

From the remaining 102 participants 52 of them were boys and 50 were girls. All of them were aged between 11 and 14 years old. 70 students claimed to like English, 47 students claimed English is easy, and 62 students claimed to have good grades in it. When we compared their tiredness levels in the beginning and in the end of the data collection sessions, we found a statistically significant difference² showing that they felt less tired at the end of the session (start: mean= 4.17, median= 4; end: mean= 3.69, median= 4; p-value= 7.52×10^{-7}), suggesting that fatigue had no influence over the effort students exerted on the last exercises they solved.

As shown in Figure 4.2, the number of exercises solved by each student ranges from 2 to 15 and, on average, each participant answered 8.7 exercises (out of 15), with a median of 8 exercises,

²All the statistical tests mentioned in this section (except for the tiredness levels comparison) were done using the statistical language R with the unpaired Student’s T-test (for the parametric data) or the unpaired Mann-Whitney U test (for the non-parametric data) after checking for normality with the Shapiro-Wilk test. The tiredness levels comparison used the paired version of the tests. In all tests, the null hypothesis was rejected when the p-value ≤ 0.05 .

and took on average 2.9 minutes to solve each one (median = 2.6 minutes). Given the exercises' difficulty, the different time slots allowed by the schools, and the personal characteristics of the students, we already expected that students would solve different numbers of exercises. However, we believe that the different time slots allowed by the schools did not have a strong influence over the students' behavior and feelings. As shown in Figure 4.3, most session times range from 30 to 40 minutes. As further shown in Figure 4.4, the average time students spent on each exercise is, in general, the same as those who had longer sessions.

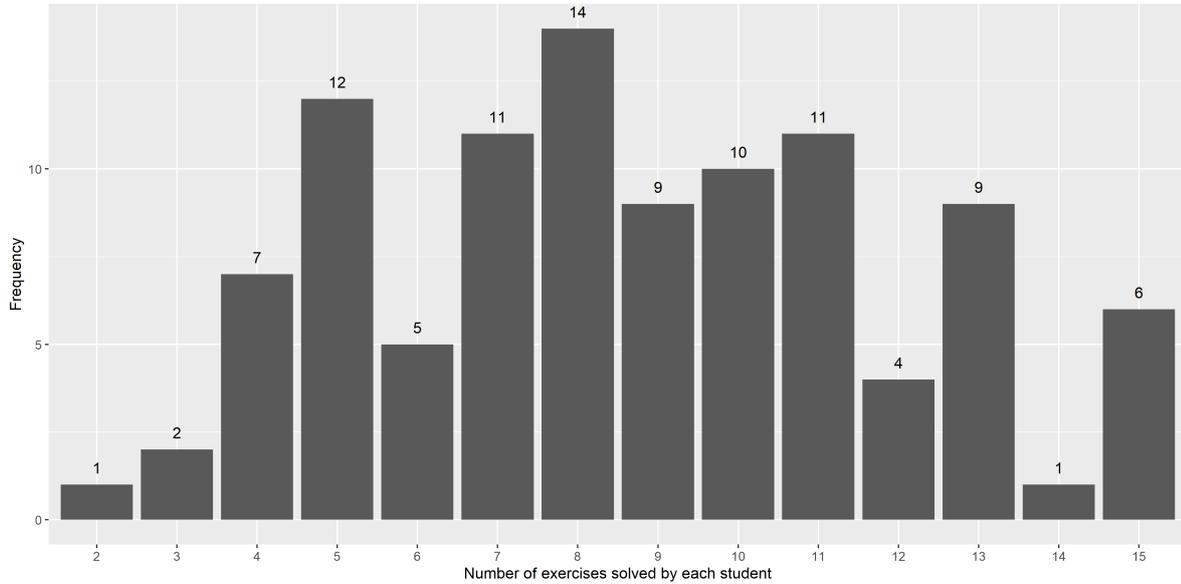


Figure 4.2: Distribution of the number of exercises solved by each student

Effort, grades and time spent

We further analyzed the data to check for differences regarding the answers provided in the questionnaires from three different perspectives: the perceived effort, the grade, and the time spent solving each exercise. We were motivated to analyze the students' effort and the time spent because they are prominent in the learning science literature. As seen in Chapter 3, parents and teachers assess whether students' exert effort or not by observing how high are the grades and if they spend time on their learning tasks [163]. These analysis are presented next.

Boys vs. girls: When we looked into gender differences, we found results that match those presented by Nagy [123]. As shown in Table 4.3, girls achieved grades higher than those achieved by boys, while also exerting a little bit more of effort and spending more time solving the exercises. As all of these values present a statistically significant difference, these results indicate that the students' gender should somehow be accounted for during the effort measurement.

Students who claimed to have good grades vs. those who did not: In this analysis, we could confirm that the students who claimed to have good grades in English are, in general, being honest. Our results show that a statistically significant difference exists between the grades of those who claimed to have good grades in English and those who did not. However, despite not exerting different effort levels, these students have spent more time solving the exercises. This

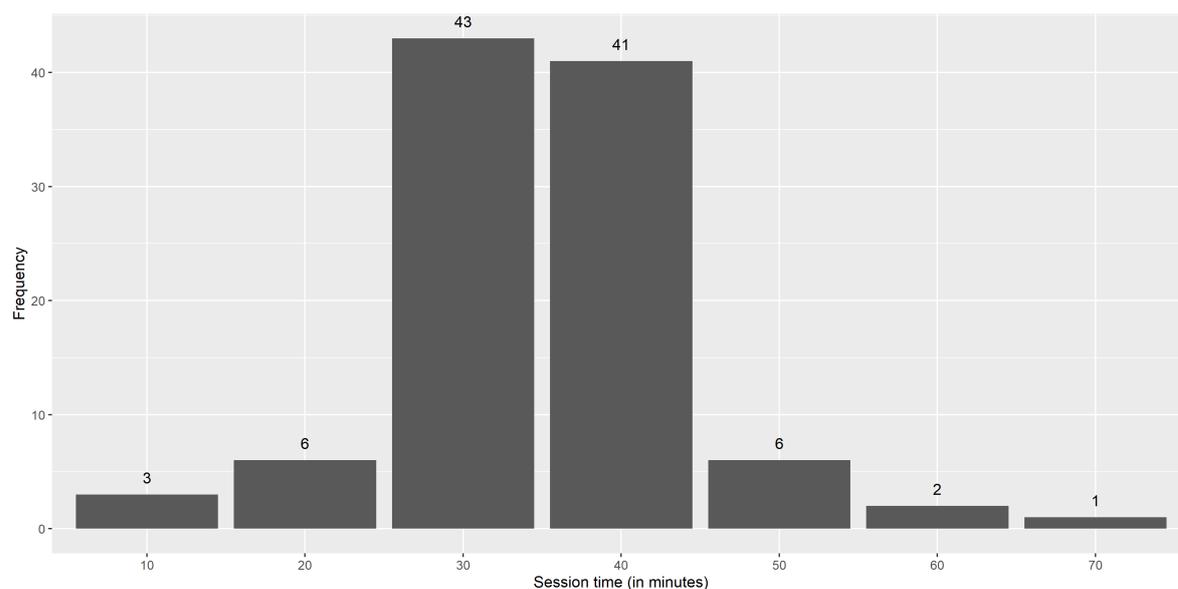


Figure 4.3: Distribution of the session time (in minutes)

Indicator	Metric	Boys	Girls		p-value
Effort	Mean	4.42	4.93	*	5.47×10^{-6}
	Median	5.00	5.00		
Grade	Mean	34.62	42.31	*	0.00115
	Median	30.38	42.86		
Time	Mean	2.58	3.14	*	6.17×10^{-8}
	Median	2.36	2.96		

Table 4.3: Effort, grade and time differences between boys and girls

suggests that the time might be more linked to the quality of the exerted mental effort than with its quantity (e.g., students who are exerting some effort and learning vs. students who exert lots of effort and eventually quit).

Students who claimed to like English vs. those who did not: Students who claimed to like the language had statistically higher grades in comparison to the other students, but exerted the same amount of effort and spent the same time solving the exercises. As presented by Stables et al. [163], enjoyment is one of the reasons why students exert effort. However, it is possible that they do not feel like they are exerting it, explaining why these students achieved higher grades with the same amount of time and perceived effort as the students who do not like English. Another possibility, is that, as they like the language, some of them study the language as an extra-curricular activity and, therefore, generally have a higher prior knowledge.

Students who claimed English is easy vs. those who did not: Students who claimed English is easy spent more time solving the exercises and achieved higher grades than those who did not, while exerting the same amount of effort. As these students think English is easy, they probably have more ability than the other students, which probably makes them achieve higher grades with less effort than the other students. As already suggested when comparing the

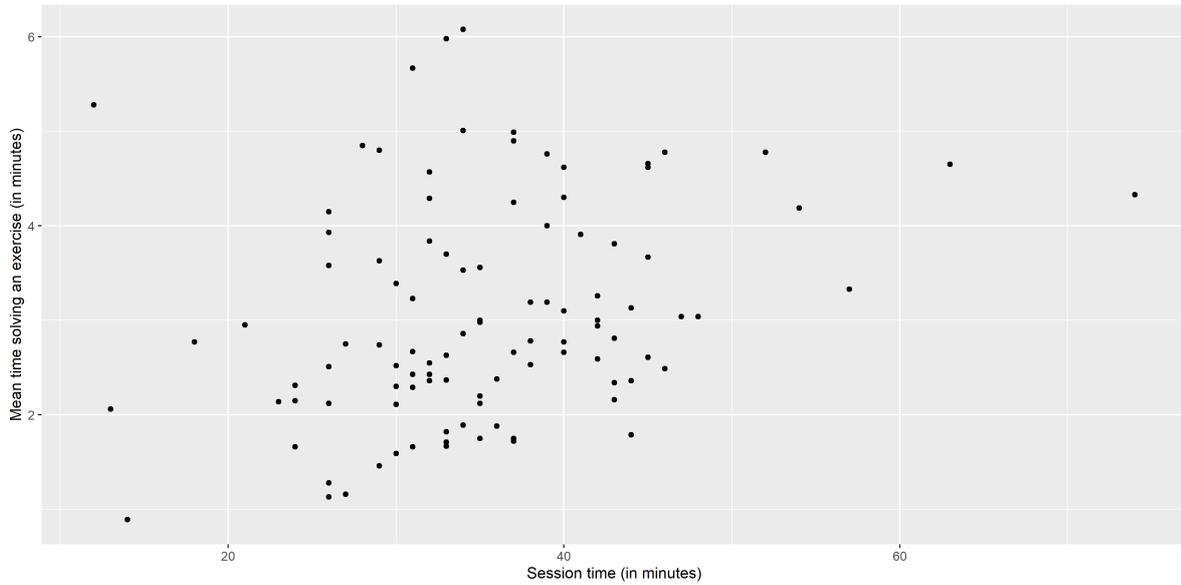


Figure 4.4: Mean time solving an exercise vs. session time (all in minutes)

Indicator	Metric	Good grades	Bad grades	p-value
Effort	Mean	4.61	4.75	0.0532
	Median	5.00	5.00	
Grade	Mean	45.61	28.94 *	8.38×10^{-14}
	Median	45.80	16.67	
Time	Mean	3.07	2.59 *	1.22×10^{-6}
	Median	2.88	2.38	

Table 4.4: Effort, grade and time differences between students' who claimed to have good grades and those who did not

students who claimed to have good grades with those who did not, time might be more linked to the quality of the exerted mental effort than with its quantity.

This analysis corroborates the need for personalization as different groups of students exerted different levels of effort and achieved different outcomes. It also corroborates the use of performance measure as a quality indicator for effort, rather than as a direct indicator.

Effort transitions and probability of good grades

We analyzed the level of effort that all of the students exerted through a Hidden Markov Model (HMM) because this model allows us to understand, through its transitions, how the students' effort varies throughout a session and also which levels of effort are related to better grades (transmission rates). Our HMM was built by considering the effort transitions of each student and is presented in Table 4.7, where each row represents the effort a student has exerted in an exercise and the columns represent the effort he has exerted in the following exercise. The transmissions rates represent the probability of achieving a good/bad grade when exerting a given level of effort in the exercises.

Since the exercises were proposed in an ascending order of difficulty, we expected to find a

Indicator	Metric	Like	Do not like	p-value
Effort	Mean	4.65	4.70	0.546
	Median	5.00	5.00	
Grade	Mean	42.92	29.42 *	3.5×10^{-8}
	Median	42.86	18.18	
Time	Mean	2.93	2.72	0.0553
	Median	2.67	2.58	

Table 4.5: Effort, grade and time differences between students who claimed to like English and those who did not

Indicator	Metric	Easy	Hard	p-value
Effort	Mean	4.57	4.75	0.0413
	Median	5.00	5.00	
Grade	Mean	46.20	32.13 *	5.09×10^{-10}
	Median	50.00	25	
Time	Mean	3.08	2.72 *	0.00545
	Median	2.76	2.54	

Table 4.6: Effort, grade and time differences between students who claimed English is Easy and those who did not

pattern that showed that students increased their level of effort throughout the session. The resulting model shows that, in general, students maintained the same level of effort for some time to eventually increase it. Interestingly, after an exercise that required effort levels 2, 3, or 4, students tended to increase their effort on the next exercise. These behaviors are compatible with the foot-in-the-door due to the increased effort, and also with the observation made by Sharma et al. [156] that students tend to maintain their behavior during data collection sessions; which suggests the need for tools to influence them to change their current state, especially if this state does not favor the learning process. However, we emphasize that the data from neither of these studies were collected under conditions of real use of a virtual learning environment and, therefore, the data may not fully reflect reality. For example, instead of continuing to solve exercises that require a level of effort equal to 7, students could end the session and/or choose exercises that are easier and/or adapted to their knowledge.

Regarding the grades, we can see in the transition and emission probabilities of the HMM presented in Table 4.7 and also in Figure 4.5 that the relationship between the students' effort and performance remembers the relationship presented in Chapter 3. That is, the greatest chances of good performance (i.e., grades greater than 50) are obtained with intermediate levels of effort (i.e., 2, 3 and 4). This reinforces the need to apply the foot-in-the-door considering the zone of proximal development [177] and increase the effort little by little.

Effort and grades correlations

As seen in Chapter 3, some measures of cognitive load are supposed to increase when the cognitive load increases and others to decrease (and vice-versa). Thus, we looked into the correlations between the effort indicators and the perceived effort ratings. Furthermore, the effort/engagement literature and the cognitive load literature have opposite assumptions. While the effort/engagement literature believes that the more effort/engagement exerted the better the

	1	2	3	4	5	6	7	End	Grades	
									Good	Bad
Start	0.05	0.04	0.21	0.31	0.31	0.04	0.04	0	0	0
1	0.36	0.05	0.07	0.11	0.12	0.07	0.05	0.16	0.38	0.62
2	0.11	0.09	0.24	0.24	0.16	0	0.07	0.09	0.51	0.49
3	0.04	0.12	0.18	0.15	0.28	0.07	0.09	0.08	0.56	0.44
4	0.06	0.04	0.14	0.22	0.24	0.15	0.04	0.1	0.54	0.46
5	0.02	0.03	0.09	0.15	0.34	0.19	0.08	0.09	0.35	0.65
6	0.03	0.02	0.04	0.1	0.23	0.28	0.19	0.11	0.22	0.78
7	0.01	0.01	0.05	0.03	0.14	0.11	0.5	0.16	0.07	0.93

Table 4.7: Effort transitions and probability of good grades in each state

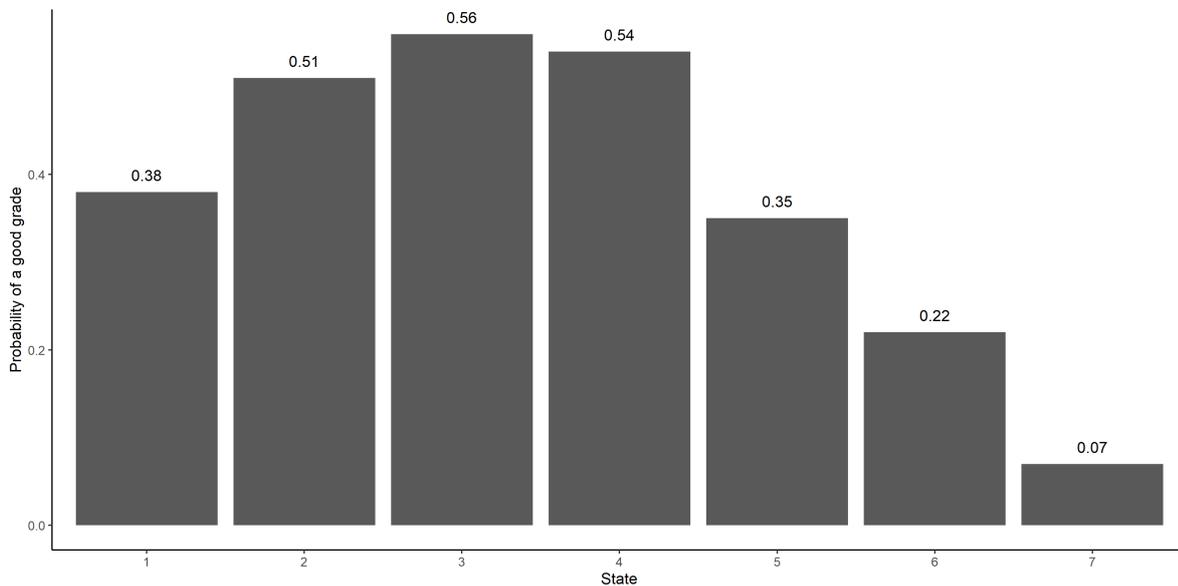


Figure 4.5: Probability of having a good grade with a given effort level

learning outcomes will be, the cognitive load literature assumes that the performance will drop as the experienced cognitive load increases. For this reason, we also looked into the correlations between our effort indicators and the grades.

Given the statistical properties of the effort subjective ratings and the grades (i.e., the effort subjective ratings are ordinal and the grades do not present a normal distribution), we computed the Spearman’s rank correlation coefficient ρ to assess their correlations with our indicators. The 10 highest effort and grades correlations are respectively presented in Tables 4.8 and 4.9 while the full correlation table is available in Appendix F.

We noticed that all of our indicators have a weak correlation with the effort ratings, while some correlations with the grades are moderate (e.g., keystrokes ratio or exercise complexity). This trend has already appeared in our preliminary study (c.f., Section 4.1), where the combined indicators had a stronger correlation with the students’ grades.

The most interesting observation we can make from the correlation analysis is that the effort ratings are negatively correlated with the grades which suggests that, at least at the task level, the cognitive load assumption has a better fit (i.e., the performance drops when the experienced

Indicator	Spearman's ρ			
	Effort		Grade	
Grade	-0.32	*	1.00	*
Smallest time between fixations	0.18	*	-0.11	*
Standard deviation of time between clicks	0.15	*	0.11	*
Mean frequency of clicks	-0.14	*	-0.17	*
Smallest saccade length	-0.14	*	-0.12	*
Highest time between clicks	0.13	*	0.17	*
Range of time between clicks	0.13	*	0.16	*
Average time between clicks	0.13	*	0.20	*
Standard deviation of time between actions	0.12	*	0.23	*
Highest time between actions	0.12	*	0.23	*

Table 4.8: Effort 10 highest correlations

Indicator	Spearman's ρ			
	Grades		Effort	
Keystrokes ratio	0.57	*	-0.08	*
Visible keystrokes ratio	0.57	*	-0.08	
Exercise complexity	-0.50	*	0.08	*
Actions ratio	0.50	*	-0.05	
Standard deviation of time between visible keystrokes	0.46	*	-0.10	*
Standard deviation of time between keystrokes	0.46	*	-0.10	*
Average time between visible keystrokes	0.45	*	-0.09	*
Average time between keystrokes	0.45	*	-0.09	*
Range of time between visible keystrokes	0.41	*	-0.05	
Highest time between visible keystrokes	0.41	*	-0.05	

Table 4.9: Grades 10 highest correlations

cognitive load increases). This assumption is further corroborated by the grades distribution in relationship with the perceived effort ratings. Figure 4.6 shows this distribution in a violin chart³ combined with a boxplot. As we can see, the probability of achieving higher grades when students exert lots of effort is smaller than when they exert a low/medium effort. For instance, when we look at the grades distribution of the perceived effort rating equal to 7, we can see that the few grades higher than 60 are outliers, while in all of the other effort levels the grades were well distributed between 0 and 100. In the effort ratings equal to 6, the grade distribution does not contain outliers, but has fewer higher grades (i.e., the width of the violin shape is smaller at the top). These results match the cognitive load literature (c.f., Chapter 3, where exerting lots of effort might compromise the learning process).

In a further analysis, we computed an effort indicator using the model developed during our preliminary study (Equation 4.3) and the data we collected. These indicators presented a weak correlation with the grades ($\rho = 0.10$) and with the perceived effort ($\rho = -0.02$) when considering all of the exercises at the same time. As we built this model considering a single task, we ran another analysis considering one exercise at a time. The resulting correlations between

³A good explanation about how to read such charts and its benefits over a boxplot can be seen in: <https://www.data-to-viz.com/graph/violin.html>

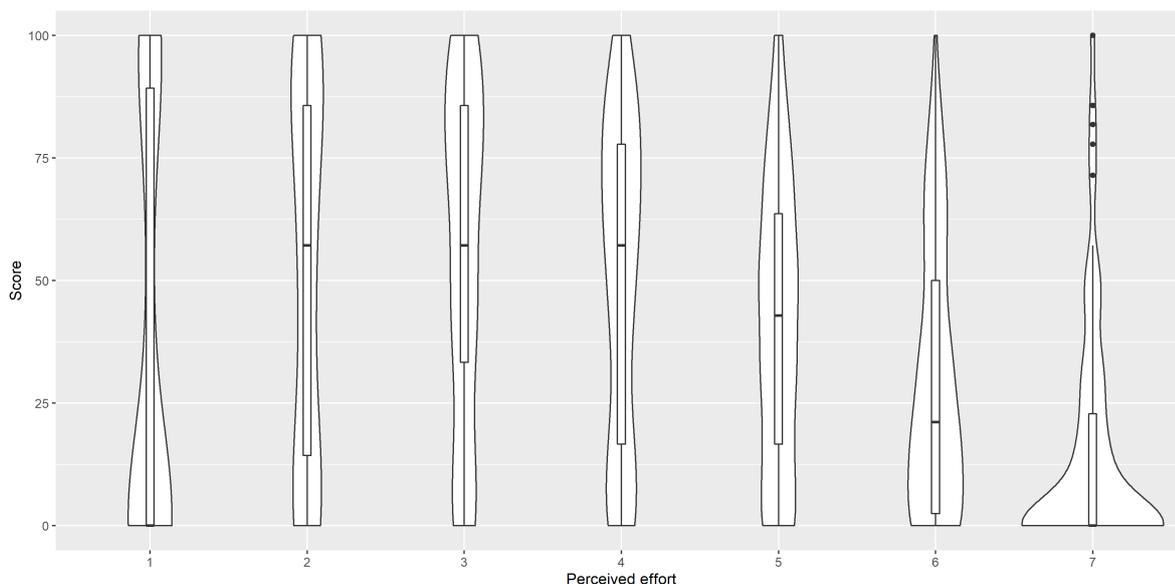


Figure 4.6: Grade vs. perceived effort

the indicators computed by the model and the students' perceived effort ranged from $\rho = -0.27$ to $\rho = 0.36$, while the correlations with the grades ranged from $\rho = -0.24$ to $\rho = 0.57$. These results suggest that the previously proposed effort model might work better when analyzing the exercises individually, and further corroborates the need to combine several indicators in order to achieve more accurate results. However, it might not be the best option in order to analyze a complete set of exercises, as it did not yield neither medium nor strong correlations with the students' perceived effort and grades. Furthermore, as this model was developed considering the students' grades as the effort indicator, and as the highest correlations we obtained are with their grades, this model performs better when estimating the students' grades and not their perceived effort.

4.3 Ethical disclaimer

Given the constraint C1 (Privacy and ethics), we would like to register in this section all of the ethical procedures followed before, during, and after the data collection.

Before inviting the students to participate in the study, we placed a demand to the Ethics Committee of the University of Lorraine to obtain an authorization to collect the data. We then invited the students to participate through the proper channels, more specifically, through the DANE. We also required students to ask permission beforehand and give us an informed consent form signed by their legal guardians (c.f., Appendix B) .

During the data collection session, we generated a unique identifier for each student, using a hash algorithm based on their name, surname, and birth date. This code was generated in order to ensure participants' anonymity by not storing any of their personal information (except for their school, which we converted into a numeric identifier, and for the answers they provided in the questionnaires, such as age, gender, etc.). In the participation certificate we gave to the participants, we marked their code and our contacts in order to allow them to retract their authorization if they wish to.

Therefore, our final dataset is anonymized and does not allow the identification of a participant without his unique code.

This ethical procedure allow us to make the dataset available to the research community, probably through the LOLA project carried out by our research team. However, this is still a work in progress.

4.4 Chapter conclusion

After a preliminary study based on a dataset that was collected for a different purpose and that was not fully compliant with the context of our study, we carried out several data collection sessions in which 120 seventh grade students from French schools agreed to participate.

The resulting dataset contains subjective, performance, behavioral and physiological data that respects all of the constraints established in the Introduction and are compliant with the context of this study. From these data, we were able to keep the data of 102 students, and to extract 126 effort-related indicators.

Our initial analysis shows that these indicators have stronger correlations with the grades than with the students' perceived effort. We also found several statistically significant differences when comparing students who claimed to like English, who claimed English is easy, boys and girls, etc. regarding their grades and the time they spent solving the exercises. However, regarding their effort we only found a difference regarding their gender (i.e., girls exerted more effort than boys).

These results show an alignment between theory and practice, which means that the students' effort is not directly observable through indicators such as time and grades, but rather through a set of indicators. Given the amount of data we have at our disposal, in the next chapter, we exploit machine learning models to combine our indicators and validate our data in the context of our study.

In this chapter we presented the dataset we built. It comprises data (more specifically 126 indicators) of 102 students from 5 different French schools who solved up to 15 English exercises while being tracked by a log tool, a smartwatch and an eye tracker. We have also shown that these data reflect the state-of-the-art and that, given the amount of data we have collected, it would be interesting to exploit machine learning models to estimate the students' effort.

Chapter 5

Measuring and predicting students' perceived effort

In the previous chapter, we described the dataset we built to carry out our study. Although this dataset comprises the four types of measures presented in Chapter 3 – namely, subjective, performance, behavioral and physiological data –, it has two main differences from the cognitive load literature. First, our dataset contains data collected from teenage students while they solved a few English exercises in a virtual learning environment. As shown in Chapter 3, the cognitive load measures were proposed after experiments carried out with adults in non-educational contexts in order to estimate the overall cognitive load in past tasks, or to identify a possible cognitive overload. As discussed in Chapters 2 and 3, children and teenagers can present behaviors and physiological responses different than those presented by adults. Second and last, our dataset includes indicators collected through the use of activity loggers, a tool widely used in the learning analytics literature, but not fully explored in the cognitive load literature. Given these differences and the comparably higher richness of our dataset, in this chapter, we want to answer the following research questions:

- To what extent can the cognitive load and the learning analytics indicators estimate students' effort in a past task?
- What indicators yield the best effort estimation of a past task?
- To what extent can the cognitive load and learning analytics indicators estimate students' effort in a future task?
- What indicators yield the best effort estimation of a future task?

To answer these questions, we exploited machine learning methods to automatically combine our measures and estimate the students' perceived effort in past and future tasks. Therefore, this chapter is divided in two main sections. In the first one (Section 5.1), we focus on estimating the students' perceived effort in a past task t (measurement). In this section, we present in details a few related works and two different approaches to measure the students' perceived effort. The first approach follows the state-of-the-art and measures the students' perceived effort in a task t with the data collected during its execution; while the second approach combines the data collected during all of the tasks already executed (including the task t) to measure the students' perceived effort in task t . In the second main section (Section 5.2), we focus on estimating the students' perceived effort in a future task $t + 1$ (prediction). We present a few related works

and an approach to predict the students' perceived effort in a future task $t + 1$ using the data collected during all of the tasks already solved. We then conclude this chapter in Section 5.3.

5.1 Measuring students' perceived effort

We start this section by presenting two related works in Section 5.1.1, which are used as a baseline for the first approach. Next, in Section 5.1.2, we present the methodology and the results obtained with the first approach. Then, in Section 5.1.3, we present the methodology and the results of our second approach.

5.1.1 Related work

As mentioned in Chapter 3, two closely related cognitive load studies have been recently published, namely the works of Borys et al. [21] and Herbig et al. [73]. As we intend to use them as a baseline, we describe their work in depth, focusing mainly on the aspects related to our work in order to allow a comparison between our work and theirs.

The study of Borys et al. [21]

The aim of the study of Borys et al. [21] was to perform a classification of the participants' mental states. The participants of their study were 20 male graduate students (computer science master degree). While the participants executed the proposed tasks – six arithmetic tasks composed with 17 operations (adding and subtracting integers), and five breaks run alternately –, their data was captured by an eye tracker and by an EEG device in a controlled lab setting. Unfortunately, due to the low quality of the EEG recordings, the final dataset contains the data of only 13 participants.

This dataset is composed of features extracted from the data collected by the EEG device and by an eye tracker. Since our dataset does not contain any EEG data (c.f., Chapter 4), we are mainly interested in the features they extracted from the eye tracker data:

- Number of fixations
- Mean, median, standard deviation, and maximum of fixations' durations
- Mean, median, standard deviation, and maximum of saccades' durations
- Mean, median, and maximum of saccade amplitude
- Mean, median, and maximum of saccade acceleration
- Mean, median, standard deviation, maximum, skewness, and kurtosis of pupil diameter
- Mean, median, standard deviation, maximum, and sum of blinks' durations

Borys et al. [21] adopted two different approaches to carry out their study. The first approach was to classify the data in only two mental states: cognitive load (i.e., the participant was solving an arithmetic task) or no cognitive load (i.e., the participant was on a break). Their second approach consisted in distinguishing between low cognitive load, high cognitive load, and no cognitive load. As Borys et al. [21] did not collect any subjective ratings, the distinction between low and high levels of cognitive load was made according to the mean response time

of the participants, which as discussed in Chapter 3, might not be representative enough of the experienced cognitive load because, as it is a performance measure, two people with the same response time can experience different cognitive load levels. In both of their approaches, Borys et al. [21] tested several classifiers, several hyperparameters and different feature selection methods.

We are particularly interested on the models they trained to classify the data into three cognitive load mental states because they are more similar to the models we trained. Among these models, the best performing one is a K-NN classifier that uses the Euclidean distance algorithm with a distance weight (squared inversed), and $k = 10$. As input, they used the 17 eye tracker features who presented a p-value < 0.05 when the Chi-squared test was applied. The mean accuracy value of this model was equal to 73%, while the accuracy of all the other models trained following the same approach ranged from 50.4% to 73%. Unfortunately, they did not compare their results with any baselines.

In our study, we have several differences when compared to the study of Borys et al. [21]. More specifically, we do not only have a dataset containing different features and collected from a larger number of participants, but the participants are in a different age range, solved English exercises instead of arithmetic ones, did not have a break during the tasks, solved the exercises in their school together with a few peers instead of in a controlled lab setting, etc. However, the most important difference might come from the fact that we do not have a “no cognitive load” option, and that our classes were defined according to the students' subjective ratings and not according to the students' mean response time, which is a performance measure and, therefore, not fully reliable to be used as a ground truth (c.f., Chapter 3).

The study of Herbig et al. [73]

Herbig et al. [73] claim to have “investigate[d] the so far most diverse set of cognitive load measures in the e-learning domain”. This set of measures is composed by heart, skin, eye, body posture, performance and subjective measures, accounting to all the measure types identified in Chapter 3. They analyzed how well different subsets of captured data can be used to estimate the intrinsic difficulty, the perceived cognitive load, and the perceived difficulty of the task. As our goal is to estimate the students' cognitive load (i.e., students' effort), we will only present this aspect of their study.

Their experiment was carried out in an e-learning setting where participants learn through videos and quizzes. The participants were 21 students aged from 20 to 33 years old (mean=25.2) and most of them are male (17 male subjects). All of the participants were enrolled in a computer science-related course and have already been approved in the mathematics lectures covering the selected topics: vectors, integration, and eigenvectors. Participants were invited to answer an initial questionnaire, and then watched six pairs – two pairs for each topic, one is considered easy, and another considered hard – of mathematics videos and quizzes in a Moodle e-learning platform. After each video and quiz, students assessed their cognitive load level. After each quiz, participants took a break. At the end of the session, they answered a final questionnaire. During the session, the participants' data was captured through several devices. We are mainly interested in the features that can also be found in our dataset (c.f., Chapter 4):

- **Subjective measures:** After each video and quiz, students were asked to provide two ratings. One of these ratings was their perceived cognitive load informed in a 9-point Likert scale.
- **Performance measures:** They captured the time participants took to answer each quiz, but ignored the time required to watch a video because it has the same duration for all

students. They also collected the percentage of correct quiz answers, but they did not exploit them in their models because they were available only for the quizzes (i.e., it was not possible to attribute a grade to the videos).

- **Eye-based measures:** From all of the eye-based measures present in their study, the ones comprised in our dataset are the number of fixations normalized by the content time, fixation durations, saccade durations, and pupil diameter.
- **Heart measures:** As they extracted heart rate variability and blood volume pulse measures (which are not available in our dataset), none of their heart rate measures are comprised in our dataset. It is interesting to note that they used at least three different wearable devices to be able to do so, which is not feasible in real-life scenarios (i.e., the constraint C3 (Real life application) is somewhat disrespected).

They preprocessed the data by replacing the missing values with the participants' mean, normalized the data, and used a recursive feature elimination with cross-validation to select the best 30 features, a ten-fold cross validation process, and a ridge regression (with $\alpha = 2$) to train different models using different subsets of data. They assessed how well each one of these models estimated the participants' cognitive load, and compared the mean squared error (MSE) of each one of their models with the MSE of a baseline regressor that always estimated the participants' mean value.

Their results have shown that, when estimating the cognitive load in all of the quizzes (the aspect of their study with the highest similarity with ours), their baseline yielded an MSE of 2.673, while their model yielded an MSE of 0.696 with a statistically significant difference. They have also shown that a combination of all the features (except for the performance and body posture measures) had the best performance with an MSE of 0.7 while other combinations ranged from 0.95 (eye and heart-related measures combined) to 2.05 (skin-related measures).

The works from Herbig et al. [73] and Borys et al. [21] have two main differences between them: the former used a regressor to estimate the participants' subjective ratings, while the latter used a classifier to estimate the cognitive load status defined according to the participants' response time. Our study also differs from the work from Herbig et al. [73]. More specifically, our dataset contains different features and contains data from more participants with a different age range, solved English exercises instead of mathematics ones, solved the exercises in their school together with a few peers instead of in a controlled lab setting, etc. Furthermore, the participants of our study are more uniformly distributed regarding their gender, since 52 boys and 50 girls participated in our study, while the participants from the present study were in majority men. This gender distribution could have impacts on the model performance, as we have already shown a few differences between the behavior of boys and girls in Chapter 4. The last difference concerns the effort rating scales they used, which are not the same as the ones we have in our dataset (the ratings they collected range from 1 to 9, while the ratings we collected range from 1 to 7) (c.f., Chapter 4).

5.1.2 Comparison with the state-of-the-art

We will now present our first approach to measure the students' effort. It uses the data collected during a task t to estimate the students' perceived effort in task t , to which we will refer to as *effort features*. We present in the following sections the methodology, and the results in comparison to the state-of-the-art.

Methodology

In order to train our effort models we adopted an experimental methodology with two steps.

Step 1 - Method choice: Our first step was to choose the model with the best performance in order to fine tune it later. In the literature, we can find three different categories of machine learning methods:

1. **Classifiers:** These methods are recommended to estimate categorical values, such as, classifying a student as approved or not; or an animal into cat, dog or fish. As shown in Chapter 3, they have been extensively used to estimate the experienced cognitive load. Therefore, we experimented with 24 classification methods available in the Scikit (version 0.22.1) Python library.
2. **Regressors:** These methods are recommended to estimate continuous values, such as, the carbon emission, a person's income, etc. Although we do not believe regression models are adapted to estimate our effort levels (i.e., 7 Likert scale points) because they are not continuous values, the regressors were successfully used by Herbig et al. [73] to estimate the experienced cognitive load (i.e., 9 Likert scale points). Therefore, we experimented with 28 regressors available in the Python Scikit (version 0.22.1) library.
3. **Ordinal regressors or ordinal classifications:** These methods are a special case that lies between the classifiers and the regressors because they are recommended to estimate ordinal values or, in other words, categorical values whose order have a meaning [70]. As this is the case of Likert scales, we experimented with 2 ordinal regressors available in the Python mord library (version 0.6). Note that we did not find any cognitive load studies exploiting such methods.

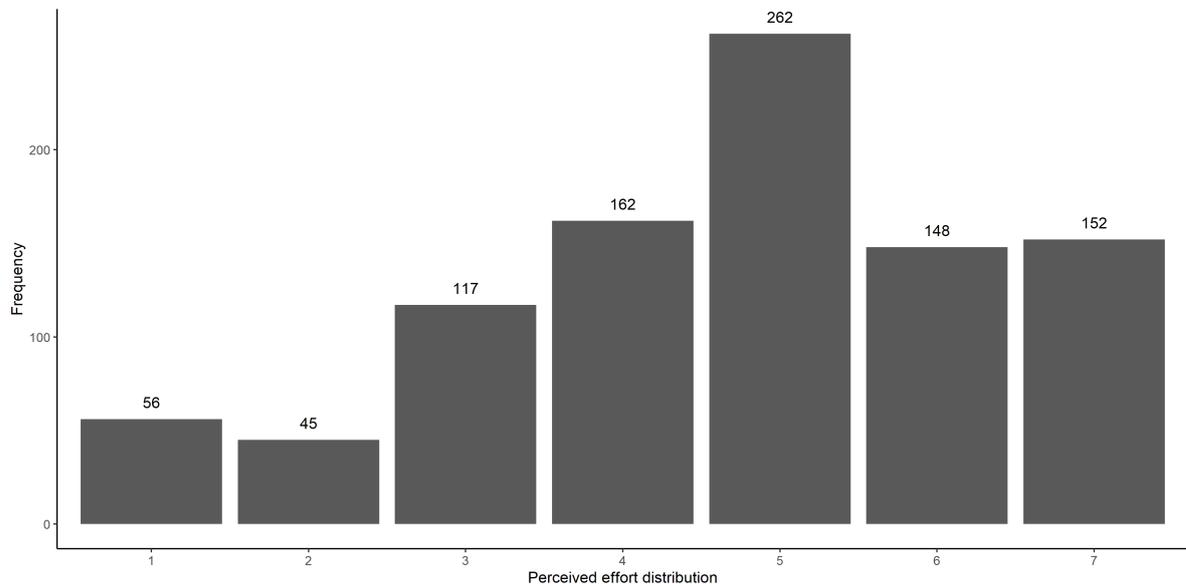


Figure 5.1: Effort ratings distribution

We ran a 5-fold cross-validation with each one of these methods using the complete feature set (i.e., all of the indicators presented in Chapter 4 that do not require a frequent user input as

listed in Appendix G) and the default hyperparameters. As our dataset is imbalanced (Figure 5.1) and contains a large number of ratings equal to 5, we chose the classification and the ordinal regression method that yielded the highest balanced accuracy (or the average recall of each class), and the regression method that yielded the lowest root mean squared error (RMSE). Therefore, the chosen methods were the Extra Trees classifier, the KNN regressor, and the LogisticIT ordinal regressor.

We then checked whether some data preprocessing techniques, such as feature selection and feature scaling, increased the models performance. The removal of highly correlated features (features with a Spearman's correlation coefficient ρ higher than 0.90), the replacement of missing values with the mean, and the feature scaling enhanced the models performance and, therefore, were included in the machine learning pipeline. However, it is important to note that the features scaling is not done in the classification process because the Extra Trees classifier does not require feature scaling in order to perform well, and that the regression and ordinal regression processes use different data scalers (we use the MinMax scaler⁴ for the regressions and the MaxAbs scaler⁵ for the ordinal regressions).

The resulting pipeline is presented in Figure 5.2 and explained in the following step.

Step 2 - Model training: As mentioned in the beginning of this chapter, we do not only want to validate the generalizability of the measures found in the cognitive load literature to the context of our study, but we also want to have more in depth information about the cognitive load measures that have a higher predictive power when it comes to the students' perceived effort measurement. Therefore, we splitted our dataset into different feature sets. These feature sets were mainly defined according to their source (e.g., pupil diameter, heart rate, etc.) and type (i.e., performance, behavioral, and physiological) and are fully described in Appendix G. Moreover, we have also defined a few features sets according to the works of Borys et al. [21] and Herbig et al. [73]. Inspired by each one of these works we built two feature sets: one set containing all of the features from their work that are also available in our dataset (e.g., mean of fixations duration, mean of saccades' length) to reproduce their work to the best extent possible; and another set containing all of the features from types available in their work and in ours (e.g., all the features related to the fixations, all of the features related to the heart rate) to have a richer feature set inspired by their work.

All of these feature sets were combined with the subject and exercise features in order to be used as inputs to train our effort measurement models. The reasoning behind this choice was that the student features (i.e., age, gender, etc.) would account for effort differences regarding the personal traits of each subject; and the exercise features (i.e., difficulty, type, etc.) would account for the effort differences regarding the characteristics of each exercise or, in other words, to changes in the intrinsic cognitive load (c.f., Chapter 3). Following this reasoning, the effort features would account for the extrinsic and to the germane cognitive load, and all of the features together account for the overall cognitive load/students' effort.

Each one of the possible combinations of features was preprocessed in order to remove highly correlated features, scale the values and replace the missing values with the mean. They were then used as input to train an effort measurement model trained using a 5-fold cross validation process and a grid search (i.e., an exhaustive test of pre-defined hyperparameters for a given method) to choose the best hyperparameters. Each classification and ordinal regression model was then evaluated according to its balanced accuracy, and each regression model was evaluated

⁴<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>

⁵<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MaxAbsScaler.html>

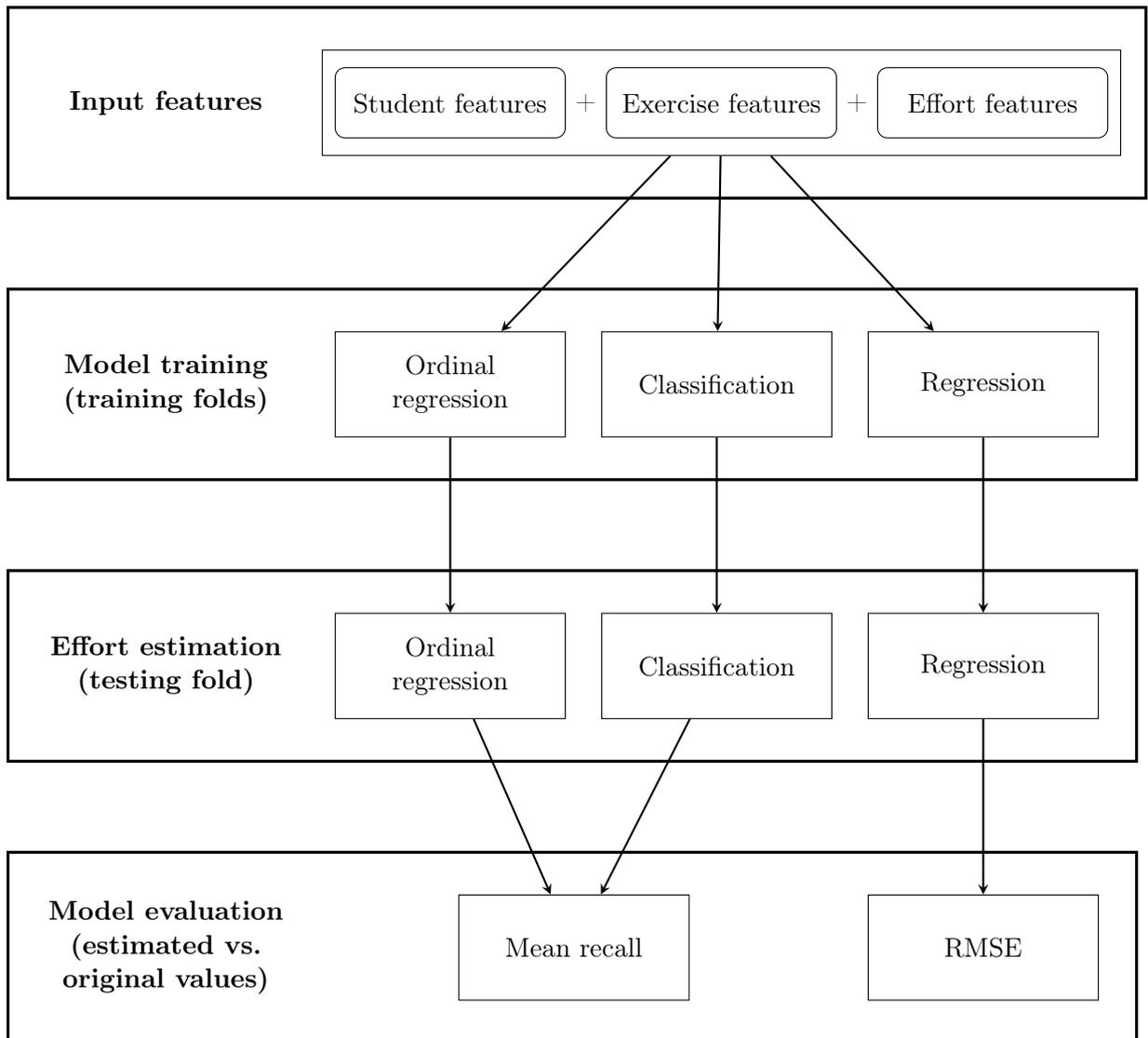


Figure 5.2: Machine learning pipeline

according to its RMSE.

The performance of each model was then compared to a few baseline models. Each one of the classification and ordinal regression models was compared to a dummy classifier that outputs a random value taking into account the underlying distribution of the real students' perceived effort ratings, and to the best reproduction possible⁶ of the classification model proposed by Borys et al. [21]. On the other hand, the regression models were compared to a dummy regressor that always estimates the mean of the real ratings and to the best reproduction possible⁷ of the regression model proposed by Herbig et al. [73].

⁶We used the same machine learning method, the same hyperparameters, and all of the features resulting from the intersection of our dataset and their dataset.

⁷See footnote 6.

Results

We trained several models using different feature sets that were combined with the student and exercise features in order to train the classification, ordinal regression, and regression models using a 5-fold cross-validation approach and a grid search to choose the best combination of hyperparameters. However, in this section, we discuss only the classification models because they not only yielded the best results, but they allowed a more fine grained analysis through their confusion matrices and additional metrics. We did not expect the classification models to outperform the ordinal regression ones, since the latter held the promise of better performance given our type of data, but the resulting models present, in general, a balanced accuracy as good as chance. Regarding the regressions, the results are consistently better than the mean and as good as the model proposed by Herbig et al. [73]. Furthermore, as already mentioned, we do not believe a regression model is the most suitable method for our classes because the subjective ratings are not continuous values. The results obtained with the ordinal regressions and with the regressions can be seen in Appendix I.

Subset	#Features	1-7 Recall		1-7 Accuracy		1-3 Accuracy	
Dummy	10	0.13	#	0.17	#	0.44	#
Borys et al. [21]	22	0.25	*	0.29	*	0.56	*
Empty	10	0.27	*	0.35	*	0.6	#
Grade	11	0.29	*	0.33	*	0.61	*
Time	11	0.28	*	0.34	*	0.6	*
Heart rate	15	0.26	*	0.33	*	0.6	*
Pupil diameter	20	0.29	*	0.36	*	0.61	*
Fixations	14	0.28	*	0.35	*	0.61	#
Fixations duration	14	0.29	*	0.36	*	0.62	#
Time between fixations	14	0.28	*	0.35	*	0.61	*
Distance between fixation points	15	0.27	*	0.32	*	0.6	*
Fixation points change speed	13	0.28	*	0.37	*	0.61	#
Clicks	13	0.27	*	0.36	*	0.63	#
Time between clicks	14	0.27	*	0.36	*	0.61	#
Keystrokes	11	0.27	*	0.34	*	0.6	#
Time between keystrokes	14	0.29	*	0.37	*	0.61	#
Visible keystrokes	10	0.27	*	0.34	*	0.59	*
Time between visible keystrokes	10	0.27	*	0.34	*	0.6	*
Backspaces	14	0.28	*	0.35	*	0.6	#
Time between backspaces	14	0.28	*	0.35	*	0.61	#
Actions	13	0.27	*	0.36	*	0.62	#
Time between actions	15	0.27	*	0.36	*	0.61	#
Attempts	11	0.28	*	0.35	*	0.6	#
Borys et al. [21] (types)	40	0.31	*	0.37	*	0.63	*
Herbig et al. [73] (types)	45	0.3	*	0.37	*	0.62	#
Eye activity	30	0.3	*	0.36	*	0.61	#
Interactions	41	0.32	*	0.36	*	0.66	*
Performance	12	0.29	*	0.34	*	0.63	#
Behavioral	39	0.28	*	0.35	*	0.63	#
Physiological	45	0.31	*	0.36	*	0.63	*
All	76	0.33	*	0.39	*	0.65	*

Table 5.1: Classification's accuracy while measuring students' effort with effort features

Table 5.1 (column *1-7 Recall*) presents the balanced accuracy values of the resulting Extra Trees classification models trained with different subsets of features to estimate the 7 perceived effort ratings provided by the students (c.f., Chapter 4). As already mentioned, all of these subsets are a combination of data collected during the exercises with the students' personal data and the exercises' metadata. Therefore, the empty subset contains only the students' personal data and the exercise metadata, while the other subsets contain a combination of data collected during the exercises (e.g., grades, time, keystrokes, time, etc.) with the students' personal data and the exercises' metadata.

The performance of these models is further compared to two baselines: the first is a dummy classifier that predicts a random class according to the class distribution – with a statistically significant difference denoted by an asterisk (*) – and the second is the model proposed by Borys et al. [21] – with a statistically significant difference denoted by an octothorp (#). If the asterisk and/or the octothorp are in red, it means the model in question yields a value smaller than its baseline, and if it is blue the model in question yields a larger value. It is worth noting that the interpretation of the marker color changes according to the metric being observed. For instance, higher accuracy values mean better models, therefore the blue marker is desirable. On the other hand, smaller RMSE also mean better models, and the red marker is desirable in this case.

In a similar manner, column *1-7 Accuracy* presents the overall accuracy values of these models without considering the imbalanced distribution of our dataset, and column *1-3 Accuracy* presents the accuracy values when considering a mapping of our 7 classes into three classes (low, medium, high). In this mapping the ratings equal to 1 (very, very little) and 2 (very little) represent a low effort level; the ratings 3 (little), 4 (neither little nor much) and 5 (much) represent a medium effort level; and the ratings 6 (very much) and 7 (very, very much) represent a high effort level. Through this point-of-view, a rating equal to 2 classified as 1, or a rating equal to 7 classified as 6 are considered as an acceptable error that happened due to the relativistic nature of the subjective ratings being transformed to an objective measure. Furthermore, it is possible that ratings equal to 1 and 2 have the same meaning to different students. However, the main reason behind this choice was to allow a more direct comparison with the results presented in the work of Borys et al. [21] who, as already mentioned, classified their data into three classes using the accuracy evaluation metric.

From the presented results, we can see that all the models, regardless of their input subset of features, performed consistently better than chance as all of the presented models have a mean recall (0.26 – 0.33) and accuracy (0.32 – 0.39) with a statistically significant difference when compared to the dummy classifier (mean recall of 0.13 and accuracy of 0.17). When comparing our model to the model proposed by Borys et al. [21], only four models have a mean recall that presents a statistically significant difference: the same types of features found in the work of Borys et al. [21], interactions, physiological, and all.

If we look at the general accuracy of the models (column *1-7 Accuracy*), in addition to the already mentioned models, we can also see another 10 models that have a statistically significant difference when compared to the work of Borys et al. [21].

All of these conclusions still hold true when we look into the accuracy of our models considering our mapping into three classes (column *1-3 Accuracy*). This is particularly interesting because these results are similar to the results of Borys et al. [21]. More specifically, our models reached accuracy values ranging between 0.60 and 0.66, while the models proposed by Borys et al. [21] present accuracy values ranging from 0.50 to 0.73, which means that our models have a performance equivalent to the models based on the work of Borys et al. [21]. These findings are even more interesting when we consider that one of their classes (i.e., no task) refers to a resting task and, therefore, to a very distinct behavior when compared to the other two classes (i.e., low

and high cognitive load). As a matter of fact, this difference can be seen in the performance of their binary classification models (i.e., no task vs cognitive load) that have accuracy values ranging from 0.73 to 0.90. Furthermore, as previously mentioned, their classes rely on the time spent on tasks, which can be considered less reliable than the subjective ratings.

From these results, the most interesting finding might be that the model using only interaction features is one of the four models with the best mean recall (0.32) and also one of the most accurate models (0.36). This finding suggests that it might be possible to measure students' effort in virtual learning environments while they are studying by exploiting activity loggers, which are unobtrusive, easy to implement, reliable (the system automatically captures the data), and do not require any equipment from the student. The exploitation of such data to measure students' effort would completely respect C2 (Implicit data) and C3 (Real life application).

To get more information about the behavior of these models, we took a closer look at their confusion matrix. Our main objective when analyzing this confusion matrix is to assess how harmful its misclassifications can be to the educational context. We consider that the harmful recommendations are those in which exercises that demand more effort are recommended for students who should solve less demanding exercises. The reasoning is that, as seen in Chapter 3, a high level of effort may indicate that the students are struggling, which can eventually demotivate them. On the other hand, when recommending an exercise that requires less effort to a student who can solve the most demanding ones, it is not harmful because this student will probably be able to solve it correctly. Therefore, we consider that the ideal behavior of our measurement model is to not classify high effort ratings as low.

For this analysis, we extracted the confusion matrix from the fold with the largest accuracy from the model that was trained with all features (row *all*) and present it in Table 5.2. As can be seen, despite the good performance of the model in not classifying the high effort ratings as low effort ratings (i.e., only 2 classifications fall into this case), there are several low and high effort ratings classified as medium and vice-versa (31%). Given the subjective nature of the ratings, this is not unexpected. However, we cannot tell whether they were misclassified or if in relationship to the other samples they should be classified as such. Following the previous assumption we made, the low effort ratings classified as medium do not pose a problem. On the other hand, the high effort ratings might (although in a smaller intensity than when classified as low). For instance, in this scenario, ratings equal to 6 and classified as 5 are still fine (small difference), but when they are classified as 3 they might be problematic (greater difference). Fortunately, the latter cases are smaller than the former, which suggests that the model does not make harmful misclassifications.

The results shown here indicate that the resulting models have a good performance when classifying the data into different levels of perceived effort. More specifically, these models were capable of measuring the teenage students' perceived effort during English exercises and to outperform a few models found in the cognitive load literature. Our results have also shown that the interaction features can be exploited in order to measure the students' effort in the task-level, which makes our models easy to use in online educational scenarios. Moreover, these models respect all of the constraints imposed in this work, namely, C1 (Privacy and ethics), C2 (Implicit data), and C3 (Real life application).

5.1.3 Added-value of combining the session data

In the previous section, we estimated the effort rating a student would give to an exercise using data collected during said exercise. However, given that some studies show that students tend to present an uniform behavior during a session (c.f., Chapter 4), we hypothesized that we

	Estimated ratings						
	Low		Medium			High	
	1	2	3	4	5	6	7
1	10	0	0	0	0	1	0
2	0	1	4	1	1	1	1
3	5	1	1	5	7	4	0
4	0	1	5	8	14	1	4
5	1	0	1	6	32	6	6
6	0	0	1	2	13	12	2
7	2	0	0	3	4	4	17

Table 5.2: Confusion matrix of the best effort measurement model using effort features

could build an effort profile using the data collected during the entire session and that it could be used to measure the students' effort. In other words, given a student A who has solved exercises 1, 2 and 3, we can extract new features using all the data collected during these three exercises to estimate the effort student A exerted during exercise 3. In a similar way, we can extract new features from the data collected while student B solved exercises 2, 3, 5 and 7 to estimate the effort he exerted during exercise 7. These new features correspond to our definition of engagement, i.e., a long-term construct that encompasses effort. Therefore, from now on, we will refer to them as *engagement features*. These features were used to train new effort measurement models following the same methodology adopted in the previous section.

In this section, we present the engagement features and review the adopted methodology, highlighting the differences between the current and the previous approach, and then present our results comparing them with the results obtained in the previous approach.

Methodology

In this new approach, we followed the same methodology used in the previous approach. As highlighted in Figure 5.2, the main difference is the input features that are now the engagement features. The proposed engagement features were extracted from the data in two different ways, according to how the data is captured:

1. **Continuously captured data:** Some measures were continuously captured while a student solved an exercise (e.g., the heart rate, fixations durations), which means that this measure is a list/collection. Therefore, to extract the corresponding engagement features, the lists/collections of all of the exercises already solved were merged and used to compute the mean, median, standard deviation, sum, maximum value, minimal value, and range. These features were computed just like the related effort features, with the difference that we now consider the data captured in several exercises and not only in one.
2. **Discrete data:** The other measures can be considered discrete, as each exercise solved by a student yields only a scalar value (e.g., number of clicks and keystrokes). However, when we consider all of the exercises already solved by a student, we can also have a list/collection. This allows us to compute their mean, median, standard deviation, sum, maximum value, minimal value, and range.

Overall, 411 engagement features were extracted following this procedure. All of these features can be seen in Appendix H.

The other differences in our methodology arise as a result of the features used. The first difference is related to the different input feature sets we experimented with. The procedure we just described allowed us to create a new feature set that provides a summary of the previously solved exercises, and another feature set that provides a summary of the perceived effort ratings provided in the previous exercises, which will only be used for comparative purposes because they disrespect constraint C2 (Implicit data). At the same time, we no longer have features that correspond to the features described in the works of Borys et al. [21] and Herbig et al. [73] because their features were extracted from the data captured during a given task. However, we kept the feature subsets composed of the same type of features they used in their work to allow at least some form of comparison.

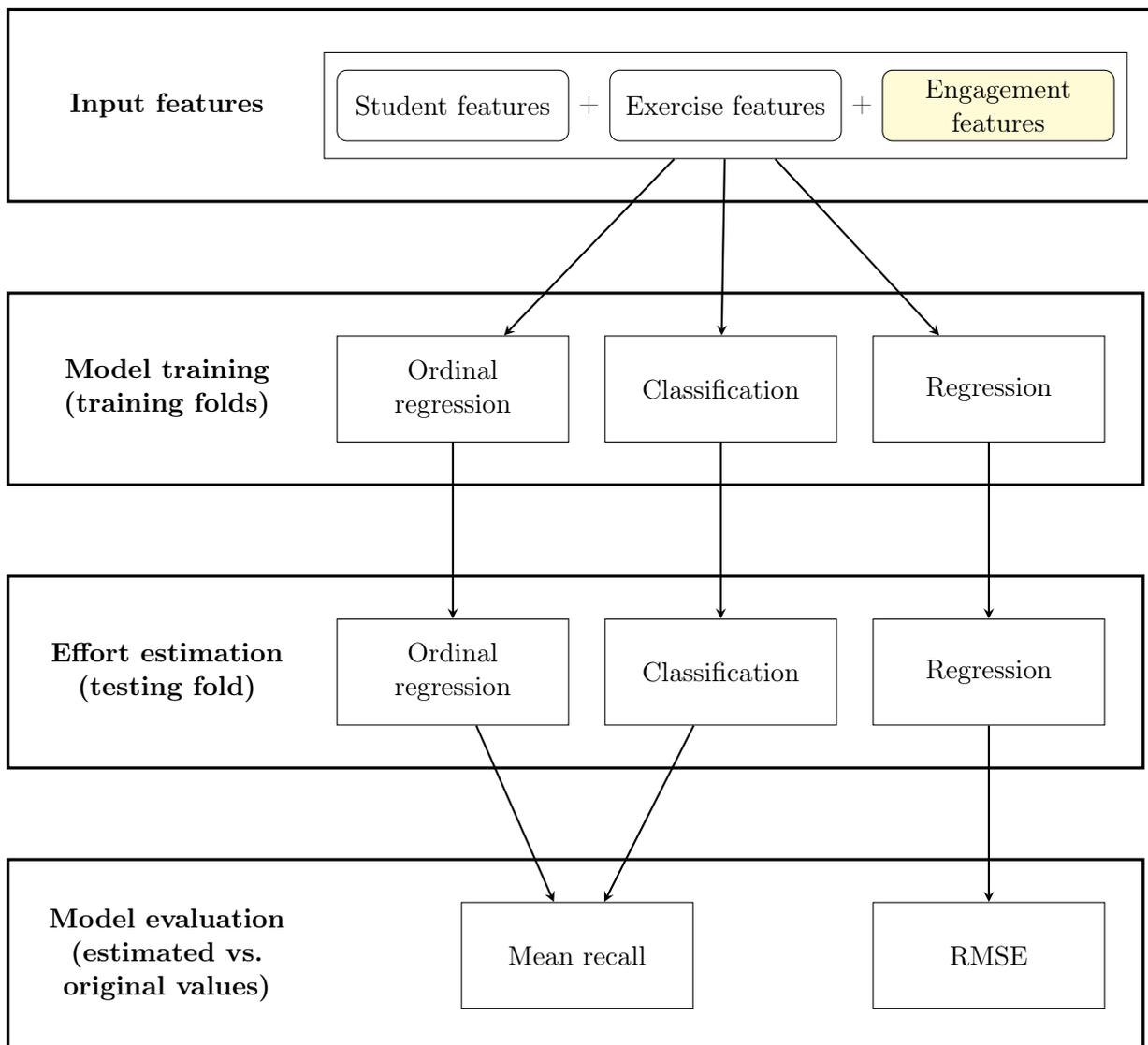


Figure 5.3: Machine learning approach adopted in our approach to measure the students' effort with engagement features

Results

Following the same methodology used to train the previous effort measurement models, we trained new models using the extracted engagement features. Once again, the classification models had a better performance than the ordinal regressions and the regression models, and are the models discussed here. Their mean accuracy values is shown in Table 5.3, and compared to a dummy classifier – with a statistically significant difference denoted by an asterisk (*) – and to the correspondent effort measurement model obtained with the previous approach – with a statistically significant difference denoted by an octothorp (#). The previous color coding scheme is also applied here: **red** means that the model in question has a lower value than its baseline, and **blue** means that the model in question has a larger value.

Subset	#Features	1–7 Recall	1–7 Accuracy	1–3 Accuracy
Dummy	10	0.13	0.17	0.42
Empty	10	0.26 *	0.35 *	0.6 *
Effort	15	0.34 *	0.42 *	0.67 *
Grade	15	0.33 *	0.36 *	0.61 *
Time	16	0.3 *	0.36 *	0.62 *
Heart rate	21	0.36 * #	0.41 * #	0.68 * #
Pupil diameter	34	0.34 * #	0.37 * #	0.63 * #
Fixations	26	0.35 * #	0.41 * #	0.66 * #
Fixations duration	14	0.34 * #	0.37 * #	0.61 * #
Time between fixations	15	0.32 * #	0.36 * #	0.62 * #
Distance between fixation points	15	0.34 * #	0.41 * #	0.67 * #
Fixation points change speed	13	0.32 * #	0.36 * #	0.62 * #
Clicks	19	0.33 * #	0.37 * #	0.62 * #
Time between clicks	14	0.3 * #	0.37 * #	0.63 * #
Keystrokes	21	0.34 * #	0.4 * #	0.67 * #
Time between keystrokes	14	0.33 * #	0.38 * #	0.62 * #
Visible keystrokes	11	0.28 * #	0.36 * #	0.6 * #
Time between visible keystrokes	10	0.26 * #	0.35 * #	0.61 * #
Backspaces	25	0.35 * #	0.41 * #	0.67 * #
Time between backspaces	15	0.33 * #	0.37 * #	0.61 * #
Actions	26	0.33 * #	0.39 * #	0.65 * #
Time between actions	15	0.32 * #	0.37 * #	0.61 * #
Attempts	10	0.26 * #	0.34 * #	0.59 * #
Previous exercises	29	0.28 * #	0.35 * #	0.59 * #
Borys et al. [21] (types)	67	0.36 * #	0.41 * #	0.67 * #
Herbig et al. [73] (types)	78	0.37 * #	0.42 * #	0.68 * #
Eye activity	43	0.36 * #	0.42 * #	0.68 * #
Interactions	110	0.37 * #	0.42 * #	0.68 * #
Performance	21	0.36 * #	0.41 * #	0.68 * #
Behavioral	99	0.36 * #	0.42 * #	0.67 * #
Physiological	78	0.37 * #	0.42 * #	0.68 * #
All	183	0.36 * #	0.41 * #	0.67 * #
All but effort	178	0.37 * #	0.42 * #	0.68 * #

Table 5.3: Classification's accuracy while measuring students' effort with engagement features

As seen in Table 5.3, most of the mean recall values (column *1–7 Recall*) are slightly higher than those obtained with the model trained with effort features, and 14 of them even present

a statistically significant difference when compared to their equivalent model from the previous section: heart rate, fixations, fixations duration, time between fixations, distance between fixation points, clicks, keystrokes, time between keystrokes, backspaces, actions, time between actions, Borys et al. [21] (types), Herbig et al. [73] (types), performance, and behavioral. When considering the general accuracy of the models (columns *1-7 Accuracy* and *1-3 Accuracy*), the accuracy values are also slightly higher, and a few models also present a statistically significant difference.

Interestingly, the feature subset containing the effort ratings (row *effort*) has one of the best accuracy (0.42 in the range 0.34 – 0.42) and mean recall values (0.34 in the range 0.26 – 0.37). However, several models achieve the same recall as the model trained with the effort features (e.g., pupil diameter, fixations duration, etc.) or even better (e.g., heart rate, backspaces). Furthermore, the model trained with all of the features but the effort ones (row *all but effort*) has a mean recall of 0.37, which is higher than the mean recall of the model trained with all of the features (row *all*) (0.36), and this trend is also seen when considering the accuracy of the models. These results suggest that the subjective effort ratings are not a requirement in order to measure the students' effort during a session as better models can be obtained using physiological, behavioral and performance data. These results also show the feasibility of relying only on implicit data to measure the students' effort and allows the resulting recommendation system to be applied in real-life scenarios, which respects constraints C2 (Implicit data) and C3 (Real life application).

Following the same methodology of the previous approach, we look at the confusion matrix of the fold with the highest balanced accuracy extracted from the model trained using all of the previously presented features (row *all but effort*) to assess how harmful the misclassifications can be to our recommendation system. As can be seen in the confusion matrix (Table 5.4), in general, the model did not classify high effort ratings as low effort ratings (i.e., only 1 classification falls into this case), but several low and high effort ratings were classified as medium and vice-versa (28%). These behavior is slightly better than the behavior of the confusion matrix presented in the previous section, which corroborates our hypothesis that the we can build an effort profile to measure the students' effort.

	Estimated ratings						
	Low		Medium			High	
	1	2	3	4	5	6	7
1	9	0	1	0	1	0	0
2	1	1	2	3	1	0	1
3	0	2	5	7	9	1	0
4	1	2	3	13	11	2	0
5	1	1	5	4	29	9	3
6	1	0	0	4	11	11	3
7	1	0	1	1	5	9	14

Table 5.4: Confusion matrix of the best effort measurement model using engagement features

The obtained results suggest that the students' engagement (represented by the engagement features) has a good predictive power when it comes to the students' perceived effort, which is not surprising given that effort is one of its components. This means that the students' engagement combined with the students' personal characteristics can be considered as an effort profile that allows to estimate the effort students exerted while solving several exercises (some

features describing each exercise, as previously described, are part of the input feature set). Moreover, the proposed engagement features respect all of the constraints imposed to this work: C1 (Privacy and ethics), C2 (Implicit data), and C3 (Real life application).

However, as already mentioned, our study and hypothesis relies on data captured during a single session. Given that over long time spans students' engagement can change, using their engagement over the entire duration of a course might not be as representative of their effort than their engagement in the last session. Further studies are required to account for this limitation and understand how much students' engagement changes from one session to another.

5.2 Predicting students' perceived effort

In the previous sections, we adopted two different approaches to measure the students' effort. The first approach used effort features and the second approach used engagement features. Although both approaches yielded good results and respected our constraints, they estimated the students' perceived effort on past tasks and, in order to provide effort-based recommendations, we need to estimate the students' perceived effort on future tasks.

To achieve this goal, we further hypothesized that if the engagement features have a good prediction power to estimate the students' perceived effort in a past task, they might also be able to estimate the students' perceived effort in future tasks. Therefore, we adapted our second approach in order to predict the students' perceived effort in future tasks.

We start this section by presenting two related works that aimed in predicting the students' effort in future tasks, and show how our approach differs from theirs (Section 5.2.1). We then show how we adapted our second approach to predict the students' perceived effort (Section 5.2.2), and the results we achieved (Section 5.2.3).

5.2.1 Related Work

In the literature, we can find two recent related studies who seek to predict the students' effort in future tasks. One of these studies, as shown in Chapter 3, is issued from the cognitive load literature, while the other is issued from the effort/engagement literature. Unfortunately, as will be shown next, these studies cannot be used as baselines to our models.

The study of Kelleher et al. [89]

Kelleher et al. [89] aimed to build a model to predict the cognitive load in a future problem. They ran a data collection study to gather data from 5 pseudo-camps, involving 76 participants for 6 hours a day. However, it is important to note that the study occupied only 1 hour and 30 minutes per day, divided in two sessions of 45 minutes each. Seventy-six teenagers participated, but the data of one of them had to be excluded. From the remaining 75 participants, most of them were girls (69 girls and 6 boys) because 4 out of 5 camps were girls only.

During the sessions, participants were asked to solve a few code puzzles in a platform called Looking Glass, a blocks-based programming environment to create 3-dimensional (3D) animations and games. They were also invited to assess their cognitive load through a 9-point Likert Scale. While they were solving the puzzles, their interaction data with the system was being recorded and a few experts observed their behaviors. The resulting dataset contains more than 5000 examples of code puzzle solutions and the extracted features were designed to reflect four different aspects of the cognitive load construct. For instance some of the intrinsic load features are number of code structures present in the puzzle, the greatest nesting depth of code structures,

and the number of locked code structures. The extraneous load features are, for instance, the number of unique statements in a puzzle, the number of locked statements, and the number of statements users must place. Some of the germane load features are the ratio of participants playing their own animation and the correct one, the number of consecutive plays of their solution, and the number of consecutive plays of the correct solution. Finally, they also considered a few features describing the participants' personal characteristics.

These features were then used to train a random forest classifier for each code structure in order to predict the cognitive load in a future code puzzle. Their models, with mean absolute errors (MAE) between 1.508 and 2.710, outperform two baselines: the most frequent value and the average value rounded to the nearest integer.

The study from Kelleher et al. [89] was carried out with teenagers, in a learning task, using mostly implicit data, and has shown promising results towards the cognitive load prediction for future tasks. However, it suffers from two main limitations. The first one is related to the participants of their study, whose majority are girls. As already discussed, this gender distribution might have added some bias to the resulting model. The second limitation is related to the features themselves. As they were based on code puzzles, they might not be transferable to other types of learning tasks.

Unfortunately, we cannot replicate their study because their data is deeply related to the type of tasks proposed to the participants (i.e., code puzzles). In our study, data is not dependent on the behaviors students present while solving code puzzles, but rather on what actions they can perform on an online platform while solving an exercise, which are easily transferable to other environments and tasks, with the exception of the exercise characteristics, which will need to be adapted to the context. Furthermore, as will be shown in the following sections, we were able to build a single model that predicts the students' perceived effort.

The study of Sharma et al. [156]

In their work, Sharma et al. [156] proposed an approach to classify the type of effort (try to solve or try to guess the answer) students would exhibit in a future task using clickstream, eye tracking, heart rate, electrodermal activity, body temperature, blood volume pulse, EEG, and facial expressions data collected during an online adaptive self-assessment activity. Thirty-two undergraduate students aged between 18 – 21 years old (15 women and 17 men) participated.

In order to classify the data into the two types of effort defined, they used a K-means algorithm to cluster the participants' data into 5 clusters, defining different groups of behaviors. These clusters were then used as the observed states of a Hidden Markov Model (HMM), while the effort types were considered as the hidden states. The HMM probabilities values were initially developed using an uniform distribution and then trained with an expectation-maximization algorithm using the first 10 answers of each participant.

In order to predict the participants' next behavior, they exploited the Viterbi algorithm, which considers a given sequence of states in order to choose a hidden state (i.e., the effort type). The resulting prediction model achieves a mean weighted recall of 0.84.

The study from Sharma et al. [156] shows promising results when identifying whether or not a student will try to solve an exercise or guess the answer. However, their study was carried out with adult students and using equipment that are not widely available to the public, which does not respect the constraint C3 (Real life application) of this thesis. Another limitation of their work, is that their model only identifies whether or not the student will continue to exert effort in the next exercise without taking into consideration its characteristics. In other words, their model does not identify that the student will exert effort in exercise *A*, but will not exert

effort on exercises *B* and *C*. Therefore, their model is not a suitable baseline because it would require lots of adaptations, and maybe a bigger dataset to train the model, in order to consider the task characteristics, which is one of the goals of our effort prediction model. Furthermore, it is important to note that this work was published after we had already finished our experiments.

5.2.2 Methodology

In this new approach, we followed the same methodology used to measure the students' perceived effort with the engagement features. The only change made was to replace the exercise features with the features of a future exercise. For instance, if we want to predict the effort a student *A* who has solved exercises 1 – 5 will exert in exercise 7, we use as input the engagement features computed with the data collected during the exercises 1 – 5 and the exercise features that describe exercise 7.

The other difference in our methodology arises as a result of a new approach. Instead of comparing the resulting models with the effort measurement models trained with the effort features, we compare them with the effort models trained with the engagement features. The goal is to compare the accuracy of the prediction models with the measurement models to see if we lose a significant amount of accuracy or not.

5.2.3 Results

We trained several effort prediction models and their accuracy results can be seen in Table 5.5. Once again, we use the same color coding for the statistically significant differences; and we compare the accuracy, and the mean recall to a dummy classifier (marked by an asterisk). However, this time we compare the resulting effort prediction models with the effort measurement models that use the engagement features as input (marked by an octothorp).

From the presented results, we can see that all of the models performed consistently better than chance as all of the mean recall values, ranging from 0.26 to 0.37, have a statistically significant difference when compared to the dummy classifier (0.13). Furthermore, we can see that the mean recall of the effort predictions made do not have any statistically significant difference in comparison to the effort measurement models trained with engagement features, which means that the effort prediction models are as good as the measurement models. These results further corroborates the predictive power of the engagement features (even though the absolute accuracy is significantly smaller in a few cases).

From these results, we can still see that the model trained with interactions features is one of the most performing models with a mean recall of 0.36, and an accuracy of 0.42. Another observation that still holds is the comparison between models trained with the effort features (rows *effort* and *all*), and those without (e.g., rows *all but effort*, *interactions*, etc.). More specifically, the model trained with all of the features (row *all*), achieved one of the highest mean recall and accuracy values (0.36 and 0.42 respectively), while the models that do not rely on the effort ratings present a better performance. For instance, the model that uses all of the features but the effort ones (row *all but effort* has a mean recall of 0.37 and an accuracy of 0.43, indicating that we do not need the subjective effort features in order to predict the students' effort. This finding has the same implications of the effort measurement models, that is, they show the feasibility of relying only on implicit data to predict the students' effort and allows the resulting recommendation system to respect constraints C1 (Privacy and ethics), C2 (Implicit data), and C3 (Real life application).

Finally, in Table 5.6, we present the confusion matrix of the fold with the highest accuracy

of the model trained using most of the previously presented features (row *all but effort*). By looking at this confusion matrix with the assumption we did in the previous sections, we can see that the prediction model performs well, since it has not classified high effort ratings classified as low. On the other hand, we can still see low and high effort ratings classified as medium and vice-versa (0.29).

In general, the results shown here indicate that the prediction models perform well to infer different levels of students' effort in a future task. More specifically, they provide results similar to the ones yielded by the previously presented effort measurement models, and confirms our hypothesis that the engagement features can also be used to predict the students' effort. Furthermore, we have also shown that the previous effort ratings are not required to achieve good accuracy values. This corroborates the previous results (c.f., Section 5.1) and further suggests that the proposed features that can be captured implicitly are also good enough to predict the students' perceived effort. As already mentioned, this results show the feasibility of relying only on implicit data to measure the students' effort and allows the resulting recommendation system to be applied in real-life scenarios, which respects constraints C2 (Implicit data) and C3 (Real life application).

5.3 Chapter conclusion

In this chapter we sought to validate generalizability of the effort measures found in the learning analytics and cognitive load literature to teenage students solving English exercises. To validate our collected data, we followed three different approaches to train several machine learning models.

The first approach, measuring the students' effort using effort features, consisted in using the data collected during an exercise t to estimate the effort in said exercise, which is commonly seen in the cognitive load literature. These models outperformed the model proposed by Borys et al. [21] and, therefore, we can consider that not only we validated the use of our models in the context of this thesis, but we also made a few improvements vis-a-vis the current state-of-the-art.

The second approach, measuring the students' effort using engagement features, consisted in using the data collected during exercises $1 - t$ to estimate the effort on exercise t , following the hypothesis that the students' engagement would be an uniform representation of the students' effort and would allow its measurement. Our results confirms this hypothesis and show that we can leverage the data collected during a session (i.e., engagement features) to measure and to predict the students' effort. Furthermore, since these models outperformed a few of the previously cited models, we can consider this approach as a further advancement of the current state-of-the-art.

The third approach, predicting the students' effort using engagement features, consisted in using the data collected during exercises $1 - t$ to estimate the effort in a future task $t + 1$, following the hypothesis that the students' engagement would be an uniform representation of the students' effort and capable of predicting it on future tasks. We consider our effort prediction model as one of the first effort prediction models, together with the ones proposed by Kelleher et al. [89] and Sharma et al. [156]. However, our model has the advantage of predicting different levels of students' effort (as opposed to the work from Sharma et al. [156]) and of being easily transferred to other virtual learning environments as its features do not heavily depend on the type of task proposed (as opposed to the work of Kelleher et al. [89]).

We have also shown in this chapter that we do not need any subjective effort ratings to train models with such performances, which is an important requirement of this thesis, as imposed

by constraint C2 (Implicit data). Instead of such data, we could use all of the other proposed physiological, behavioral and performance data that include the interaction data that has not been largely explored to measure the cognitive load. This type of data has the advantage of being easily captured because it does not require any specific equipment from the students, all they have to do is solve exercises.

In summary, all of the aforementioned approaches lead to models that perform consistently better than chance; that, in general, do not classify high perceived effort ratings as low; and that improve the current-state-of-the-art. This indicates that our models can reliably measure and predict teenage students' perceived effort during English exercises. Moreover, all of our models respect the three constraints imposed: C1 (Privacy and ethics), C2 (Implicit data), and C3 (Real life application). This means that our work stands out from other related works because it does not require any cumbersome equipment and can definitely be exploited in real-life scenarios. Therefore, in the next chapter, we exploit these effort models to formalize the foot-in-the-door technique into a recommendation system.

In this chapter, we have shown that the data we collected can successfully estimate the students' perceived effort following three different approaches.

The first approach is widely adopted in the cognitive load literature and uses the data collected during the task t to estimate the students' perceived effort in the task t .

The second approach uses the data collected during all of the exercises already solved, including the task t , to estimate the students' perceived effort in the task t .

The third approach uses the data collected during all of the exercises already solved to predict the students' perceived effort in a future task $t + 1$.

Therefore, we validated the feasibility of our effort indicators to measure and predict the perceived effort of teenage students' solving English exercises.

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Dummy	10	0.13	0.17	0.43
Empty	10	0.26 *	0.28 * #	0.53 * #
Effort	15	0.34 *	0.41 *	0.63 *
Grade	16	0.32 *	0.4 *	0.65 *
Time	16	0.3 *	0.36 *	0.58 *
Heart rate	21	0.33 *	0.39 *	0.63 * #
Pupil diameter	34	0.33 *	0.36 *	0.61 *
Fixations	26	0.36 *	0.41 *	0.63 *
Fixations duration	14	0.34 *	0.37 *	0.6 *
Time between fixations	15	0.33 *	0.38 *	0.6 *
Distance between fixation points	15	0.34 *	0.39 *	0.63 *
Fixation points change speed	13	0.33 *	0.37 *	0.61 *
Clicks	19	0.32 *	0.39 *	0.61 *
Time between clicks	14	0.3 *	0.33 *	0.58 *
Keystrokes	20	0.32 *	0.39 *	0.64 *
Time between keystrokes	14	0.31 *	0.37 *	0.61 *
Visible keystrokes	11	0.28 *	0.35 *	0.58 *
Time between visible keystrokes	10	0.26 *	0.28 * #	0.53 * #
Backspaces	25	0.36 *	0.41 *	0.65 *
Time between backspaces	15	0.32 *	0.37 *	0.59 *
Actions	25	0.32 *	0.4 *	0.64 *
Time between actions	15	0.31 *	0.37 *	0.58 *
Attempts	10	0.26 *	0.29 * #	0.54 *
Previous exercises	29	0.26 *	0.28 * #	0.55 *
Borys et al. [21] (types)	67	0.35 *	0.41 *	0.65 *
Herbig et al. [73] (types)	78	0.36 *	0.41 *	0.64 *
Eye activity	43	0.37 *	0.42 *	0.66 *
Interactions	109	0.36 *	0.42 *	0.68 *
Performance	22	0.34 *	0.41 *	0.66 *
Behavioral	97	0.34 *	0.4 *	0.64 *
Physiological	78	0.35 *	0.41 *	0.63 *
All	182	0.36 *	0.42 *	0.67 *
All but effort	177	0.37 *	0.43 *	0.68 *

Table 5.5: Classification's accuracy while predicting students' effort with engagement features

	Estimated ratings						
	Low		Medium			High	
	1	2	3	4	5	6	7
1	4	2	0	3	0	0	1
2	0	2	1	2	2	0	1
3	0	1	6	3	7	0	2
4	1	0	2	6	11	5	1
5	1	0	4	6	28	4	3
6	0	0	1	4	10	9	5
7	0	0	1	3	3	4	19

Table 5.6: Confusion matrix of the best effort prediction model using engagement features

Chapter 6

Towards engaging recommendations: formalizing the foot-in-the-door technique

In the previous chapter, we described our effort measurement and prediction models. In this chapter, we explore these models to estimate the cost of each task and formalize the foot-in-the-door technique into a recommendation system.

As discussed in Chapter 3, the foot-in-the-door technique promises to increase the compliance rate. For this, we need to make consecutive requests with an increasing cost. Therefore, we hypothesize that the formalization of the foot-in-the-door technique into a recommendation system can influence students to accept more recommendations and, as a consequence, solve more exercises while increasing their effort. This hypothesis not only meets our definition of engagement (i.e., solving multiple exercises while exerting the appropriate levels of effort), but also meets the theory of commitment [93] which suggests that the act of solving exercises may eventually engage students.

As also discussed in Chapter 3, students' effort and engagement are linked to better learning outcomes. That is, students who exert effort and/or are engaged will, in general, succeed in the school year. Since our formalization of the foot-in-the-door increases the students' effort from one task to another, we hypothesize that this can also lead to an increase in the students' performance.

Therefore, our hypothesis is: H_1 – Formalizing the foot-in-the-door into a recommendation system can improve the students' effort, compliance, performance and engagement. To accept (or reject) it, we compared different recommendation models combined or not with the foot-in-the-door technique across these four aspects.

We begin this chapter in Section 6.1, where we present two different ways to formalize the foot-in-the-door technique, the selection criteria, and the algorithms for each recommendation model. In Section 6.2, we present the evaluation protocol we adopted to accept (or reject) our hypothesis. In Section 6.3, we present the results obtained. In Section 6.4, we present the limitations of our study. Finally, in Section 6.5, we conclude this chapter.

6.1 Recommendation models

To evaluate the influences of the foot-in-the-door technique on the recommendation process, we implemented different recommendation models that exploit (or not) the technique. The proposed

recommendation models have the structure presented in Algorithm 1.

Algorithm 1: Proposed recommendation model

```
1 newItems  $\leftarrow$  allItems – userItems;  
2 fitdItems  $\leftarrow$  FITD(newItems);  
3 recItems  $\leftarrow$  recAlg(fitdItems);  
4 return recItems;
```

The first step is to select which new exercises can be recommended for the student (*newItems*). For this, in line 1, we subtract the set of exercises already solved by the student (*userItems*) from the set of all exercises available for recommendation (*allItems*).

Then (line 2), we apply the foot-in-the-door function on these items to assess which ones can be recommended in order to exploit the technique and we get a new set of items (*fitdItems*).

Finally, the *recAlg* recommendation algorithm selects which items contained in *fitdItems* will be recommended (line 3). This step is important in cases where we have many exercises that can apply the foot-in-the-door technique as it serves as an additional selection criterion.

In the next sections, we propose some functions that perform the item selection according to the foot-in-the-door technique and describe the recommendation algorithms we use, as well as the different selection criteria adopted.

6.1.1 Foot-in-the-door functions

As explained in Chapter 3, to apply the foot-in-the-door technique, we must make consecutive requests with an increasing cost. This means that we need to know the cost of each exercise to choose the order in which they should be recommended.

In this thesis, we defined the cost of tasks as the effort they require from the students and we trained several models to measure and predict this effort as each student perceives it (c.f., Chapter 5). The functions we propose to explore the effort models and to apply the foot-in-the-door technique in the recommendation process are:

- **NoFITD**: This function only simulates the situation where the foot-in-the-door technique has not been applied. That is, according to Algorithm 2, this function does not perform any action on the set of new items *newItems* and returns them without any changes. Therefore, this function is only used to perform comparisons.
- **ZpdFITD**: As discussed in Chapter 3, the foot-in-the-door technique is compatible with the zone of proximal development [177], that states that the difficulty/challenge of the tasks should increase gradually. Therefore, our implementation of the foot-in-the-door technique, presented in Algorithm 3, selects all of the items whose predicted effort is higher than the user's previous effort (or all of them if the student has not solved any exercises yet) (lines 2–8). Then, from these items, the algorithm selects and returns the ones with the smaller effort (line 9).

6.1.2 Recommendation algorithms

After applying the foot-in-the-door technique, the recommendation model chooses which ones will be recommended through some recommendation algorithm. In this thesis, we choose the

Algorithm 2: NoFITD function

```

1 Function NoFITD(newItems):
2   | fitdItems  $\leftarrow$  newItems;
3   | return fitdItems;

```

Algorithm 3: ZpdFITD function

```

1 Function ZpdFITD(newItems):
2   | userLastEffort  $\leftarrow$  measureEffort(user, lastItem(user));
3   | for item in newItems do
4     |   | effortitem  $\leftarrow$  predictEffort(item);
5     |   | if effortitem > userLastEffort then
6     |   |   | fitdItems  $\leftarrow$  fitdItems  $\cup$  item;
7     |   |   | end
8     |   | end
9     |   | fitdItems  $\leftarrow$  {item  $\in$  fitdItems | effortitem = min(effort)};
10  | return fitdItems;

```

Top N and Q-learning algorithms. The Top N algorithm was chosen because it is a popular baseline in the recommendation systems literature and often provides competitive results. The Q-Learning algorithm was chosen mainly because it allow us to have a dynamic model. In other words, this algorithm can continuously update the recommendation strategy according to the user interactions, contrary to the models based on machine learning methods that require previous training and are static. Furthermore, this algorithm considers the sequence in which the items are used and, as a consequence, they can create an optimized long-term learning sequence, which, as discussed in Chapter 2, can be interesting for educational recommendation systems like those proposed in this thesis. Both algorithms are presented below.

Top N

The Top N algorithm consists in setting a score for each item and then recommending the N items with the highest (or lowest) scores. In our implementation of the Top N algorithm, we rely on a scoring function that we apply only on the items selected by the foot-in-the-door function. Furthermore, the scoring function (function *score* in line 2) can take different forms that are presented in Section 6.1.3.

Algorithm 4: Top N recommendation algorithm

```

1 for item in items do
2   | scoreitem = score(item);
3   | sort items by score;
4 end
5 return top N items

```

Q-learning

Q-learning is a model-free reinforced learning algorithm. This means that this algorithm does not need any previous information about the problem at hand because it learns from each new information received after performing an action. This algorithm can act as an intelligent agent that is, for example, figuring out how to navigate a new environment (e.g., a vacuum robot that maps a house). This environment is defined by a set of states S , a set of actions A and by the set Q representing the quality of each action a in a given state s [166]. The agent's objective when discovering a new environment is to optimize the reward (or minimize the loss) of a certain sequence of actions. For this, the algorithm can explore the actions or exploit the best-known actions (i.e., find a trade-off between exploration and exploitation).

We can identify three essential functions of the algorithm:

1. **Explore:** This function returns a random action to test it so that the agent can explore its environment and get more information about it.
2. **Exploit:** This function returns the best action a^* to be performed from state s so that the agent can benefit from the knowledge it already has. The action a^* is chosen because its value $q(s, a)$ is the highest value found in table Q from state s (or the lowest in the case of minimizing losses).
3. **Update:** As mentioned, the agent learns after each action performed. This learning takes place by assigning a new reward r to the value $q(s, a)$. This update is done through the Bellman equation [17], presented in Algorithm 5 (line 7). In addition to the s , a , and r parameters, this equation also uses the following parameters:
 - **Learning rate α :** This parameter determines the extent to which the new information (represented by the reward r) overwrites the old information. This parameter is a value between 0 and 1, where $\alpha = 0$ makes the agent completely ignore the new information, while $\alpha = 1$ makes the agent consider only the new information. According to Sutton and Barton [165] the value normally used is $\alpha = 0.1$.
 - **The discount factor γ :** This parameter defines how important future rewards are. Like the α parameter, the γ parameter accepts values between 0 and 1. A value $\gamma = 0$ makes the agent consider only the most recent information and perform short-term estimates, while $\gamma = 1$ makes the agent perform better long-term estimates.
 - **The next state s' :** This parameter represents the next state and is obtained by observing what the next state is after performing action a . It is used to obtain information about future rewards and to help the agent define the best long-term path.

When the Q-learning algorithm is used as a recommendation agent, the states S represent the last item the user interacted with and the actions A represent the possible items to recommend. In our implementation of the functions of this algorithm, we made the following choices:

- **Pre-defined states:** As our dataset contains only 15 exercises and these exercises represent our actions and our states, we decided that our recommendation agent knows all of its 16 states (i.e., 15 exercises and an empty state representing the start of a session) beforehand.

- **Solving the exploration-exploitation dilemma:** As already discussed, in its classic form, the Q-learning algorithm makes a trade-off between exploration and exploitation through approaches such as ϵ -greedy [166]. By using this algorithm as a recommendation agent, however, we assume that users' spontaneous actions (e.g., exercises that the student solved without the influence of the recommendation system) can be considered as an act of exploration. We believe this approach can have the benefit of limiting the recommendation of inappropriate items at a time t . For example, in general, at the end of a school year, students do not access the content studied at the beginning of the school year and vice-versa. Given our choice of only exploiting the best items, the items studied at the beginning of the school year will not be recommended at the end of the school year because, from a state s reached only at the end of the school year, these items will have a low value $q(s, a)$. Therefore, we do not solve the exploration-exploitation dilemma and simply explore the best actions from a given state s .
- **Exploration:** Although we don't solve the exploration-exploration dilemma, we still use an exploration function. However, we use this function only when we have more than one action a^* in our set of best actions A^* . In this case, we recommend a random a^* action (Algorithm 5, line 10).
- **Exploitation:** In the classic algorithm, the best action a^* is one of the actions A . In our implementation, as already described, we chose to exploit the best action a^* from the set of actions A' which contains only the exercises that the student has not yet solved and which have been pre-selected by the foot-in-the-door function. It is important to note that we consider as the best action a^* the one whose value $q(s, a)$ is the highest, or $q(s, a) = q_s^*$.
- **Reward:** In our implementation (Algorithm 5), this reward is obtained by the reward function (line 2). As in our implementation of the Top N algorithm, this function can take the different forms that are presented in Section 6.1.3.

Algorithm 5: Q-Learning recommendation functions

```

1 Function Update( $s, a$ ):
2    $r = \text{reward}(s, a)$  ;
3    $\alpha \leftarrow 0.1$  ;
4    $\gamma \leftarrow 0.8$  ;
5    $q_{s,a} \leftarrow q_{s,a} + \alpha(r + \gamma q_{s'}^* - q_{s,a})$ 

6 Function Explore( $A^*$ ):
7   return random  $a$  in  $A^*$ ;

8 Function Exploit( $s, A'$ ):
9    $A^* \leftarrow \{a \in A' \mid q_{s,a} = q_s^*\}$  ;
10   $a^* \leftarrow \text{Explore}(A^*)$  ;
11  return  $a^*$ ;

```

6.1.3 Selection criteria

When different items are suitable for applying the foot-in-the-door technique, we use the recommendation algorithms to choose which exercises should be recommended. In turn, these algorithms need either a function to score items (Top N algorithm) or a function to calculate the reward of each action (Q-learning). These functions were implemented according to several criteria and take different forms when used with the Top N algorithm (scoring functions) or with the Q-learning algorithm (reward functions).

Criterion	Scoring function	Reward function
Compliance	#students who solved the exercise	1
Difficulty	$\frac{\#students\ with\ a\ grade > 50}{\#students\ who\ solved\ the\ exercise}$	$\begin{cases} 1 & \text{if grade} > 50 \\ 0 & \text{else} \end{cases}$
Grade	predicted grade	real grade
Actions	$\sum_{s=1}^{\#students} \#actions_s$	$\#actions_s$
Item similarity	sim(last, new)	sim(penultimate, last)

Table 6.1: Scoring and rewards functions applied on each exercise

The criteria we have chosen and their respective implementations such as scoring and reward functions are shown in Table 6.1. These criteria are:

- **Compliance:** We chose this selection criteria because it is one of our main interests when applying the foot-in-the-door technique: influencing students' to accept to solve an exercise. As a scoring function, this criterion is applied by counting how many students have solved the exercise in question. As a reward function, this criterion is implemented by always returning 1, signaling to the Q-learning algorithm that the recommended exercise was accepted and that the student's state has changed.
- **Difficulty:** This criterion was inspired by the work of Sharma et al. [156], and it was included in our experiments because, as shown in Chapter 3, it is intrinsically linked to the students' effort and to the zone of proximal development [177]. When it is calculated by the scoring function, it calculates the probability that the student has a good grade (> 50). When it is calculated by the reward function, it simply returns whether the student had a good grade or not. Therefore, lower values indicate exercises that are more difficult and higher values indicate easier exercises.
- **Grade:** This criterion was chosen because it represents the students' learning outcomes, a criterion that several researchers seek to optimize via recommendation systems [139, 38] or exploit in order to make recommendations [126]. Furthermore, as discussed in Chapter 3, we believe that there should be a trade-off between the students' effort and grade. By

combining the foot-in-the-door technique considering effort as the cost of tasks, we assume that effort captures the students' personal characteristics (e.g., students notice less effort when they like English, students notice less effort when they are girls, etc.) and that the grade captures their skills as well as the quality and suitability of the exercises (e.g., exercises where all students have a grade equal to 0 are considered inadequate because they probably have not studied the content yet, easier exercises will lead to better grades, poorly structured exercises will lead to bad grades, etc.). Therefore, we believe that, after applying the foot-in-the-door technique, this criterion can allow the recommendation system to choose exercises at the appropriate level and still encourage learning.

In the reward function this item is the grade given to the student and in the scoring function this criterion is the grade predicted by a machine learning model trained with the engagement features. After following a process similar to the process described in Chapter 5, the most performing model to predict the students' grade was the AdaBoost⁸ regressor with 75 estimators, a learning rate equal to 0.1 and exponential loss.

- **Actions:** Huptych et al. [79] propose a recommendation system where each resource's score is defined by its relevance and the effort that the students exerted. They defined relevance as the proportion of clicks that all students performed in a given week in a given resource. Effort was defined in a similar way, that is, the proportion of clicks a student performed in a given week in the same resource. As the students whose data we collected participated in a single session, we can calculate neither the relevance nor the effort as defined by Huptych et al. [79]. For this reason, we only consider the total number of actions (clicks and keystrokes) performed by them. It is important to note that we also consider key presses because our exercises can be solved using the keyboard and mouse.
- **Item similarity:** This criterion was chosen because it is a popular metric in recommendation systems (c.f., Chapter 2). Furthermore, it might be able to reduce the cognitive dissonance (the mental discomfort that results from holding two conflicting beliefs, values, or attitudes) because it would allow the recommended exercises to have a higher similarity with the exercises students usually solve. Furthermore, this criterion might also help the recommendation model to recommend exercises with the same content.

For both the scoring function and the reward function, this function returns the similarity of two exercises. This similarity is defined as the complement of the normalized Euclidean distance. The Euclidean distance is calculated considering the following exercise attributes: difficulty (as defined in the BRNE system), type (drag n' drop or fill in the boxes), whether the exercise statement contains the possible answers or not, number of inputs, size of the answer, and complexity (as defined in Chapter 4 and in Annex G). The formula used is shown in Equation 6.1, where e and x are two exercises.

When using the scoring function, the difference is calculated between the last item the student solved and the exercise being assessed during the recommendation process. When using this criterion as a reward, this difference is calculated in relation to the item that represents the student's state and the item that represents the action performed.

$$\text{sim}(e, x) = 1 - \frac{\sqrt{\sum_{a=1}^{\#\text{attr}} \left(\frac{e_a}{\max(a)} - \frac{x_a}{\max(a)} \right)^2}}{\#\text{attr}} \quad (6.1)$$

⁸<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html>

6.2 Evaluation protocol

As already described, we intend to test a hypothesis regarding the influence that formalizing the foot-in-the-door technique in a recommendation system can have on the students' effort, compliance, grade and engagement. While we recognize the importance of performing a real-life assessment, we only perform an offline evaluation. The main reason behind this choice is the fact that, at the time an online evaluation could be carried out, we were in the middle of the Covid-19 pandemic. This means that meetings and classes were being conducted mostly through videoconferences and that contact with teachers, students and other stakeholders was difficult and even hazardous. Another reason is that the recommendation system has not yet been integrated into the teachers' dashboard and therefore we cannot test any of the proposed models in real life yet.

To carry out the offline evaluation of the proposed models, we performed a simulation with several different recommendation models (i.e., all possible combinations of the foot-in-the-door functions, algorithms and selection criteria). Therefore, we randomly divided the students into 5 groups that were used to perform a cross-validation.

Data from the students who were assigned to the training folds (4 folds out of 5) were then used to train the effort measurement and prediction models, as well as the recommendation models. To train the recommendation models, we needed to respect two constraints. First, we needed to have a certain level of randomness in our data because the Q-learning algorithm depends on the sequence in which the exercises are solved and, as described in the Chapter 4, in general, the students followed the proposed sequence. Second, we needed to respect the order on which the students solved the proposed exercises because, even though we did not directly apply the foot-in-the-door technique during our data collection sessions, we can see several cases where the foot-in-the-door technique was indirectly applied (i.e., cases where the students exerted a given level of effort in one exercise and in the next exercise they exerted more effort) (c.f., Chapter 4) and, as discussed in Chapter 3, the students' engagement in one exercise has an influence over their engagement on the next exercises. Therefore, we used one subsequence of each student to train the recommendation models because they introduce a certain level of randomness in our data, allowing us to train the machine learning models; and still respect the order on which the exercises were originally solved, allowing us to demonstrate the effects of the foot-in-the-door technique.

As shown in Algorithm 6, to define the students' subsequence, we randomly assigned a value between 0 and 1 to each exercise and used only the ones whose values were greater than 0.4⁹ (lines 1–4). The remaining exercises were ignored, unless their order was different from what was originally proposed (lines 5–13). For example, if the student solved exercise 5 and then exercise 7, it implies the inclusion of both exercises on the training data; on the other hand, if a student solved exercise 1 and then exercise 2, these exercises were only considered during the training phase if their random values are greater than 0.4.

In turn, data from students assigned to the test fold (1 fold out of 5) were used to evaluate the recommendation models (Algorithm 7). For this, we performed consecutive recommendations until this one was not accepted (i.e., the student did not solve the exercise) or until the exercise recommended did not respect the sequence of exercises originally solved by the student (e.g., if the student solved 5 exercises in the order 1-2-3-4-5 and the exercise 1 was recommended after exercise 3, it would not respect the sequence followed by the student). Due to the cold start problems of some Top N algorithm implementations (e.g., the implementation that uses

⁹This value was empirically defined and was not optimized in any way.

Algorithm 6: Recommendation models training phase

```

1 for record in records do
2   | randomrecord  $\leftarrow$  random number between 0 and 1;
3   | trainrecord  $\leftarrow$  randomrecord > 0.4;
4 end

5 for user in records do
6   | for record in records | userrecord = user do
7     | if exerciserecord = (exerciserecord-1 + 1) then
8       |   trainrecord-1  $\leftarrow$  true;
9       |   trainrecord  $\leftarrow$  true;
10    | end
11  | end
12 end

13 for record in records | trainrecord is true do
14 |   updateModel(record);
15 end

```

the students' grade as a criterion), we added an exception to our evaluation process: if no recommendation could be provided due to the lack of data, we updated the user model with the data collected during the first exercises of his subsequence.

Algorithm 7: Recommendation models testing phase

```

1 for user in records do
2   | repeat
3     |   exercise  $\leftarrow$  recommend(user);
4     |   record  $\leftarrow$  {record  $\in$  records | userrecord = user and exerciserecord = exercise};
5     |   updateModel(record);
6   | until recommendation is not accepted;
7 end

```

We chose to create a sequence for each student using the next item approach in order to avoid basing our recommendations and evaluation on the sequences they solved during the data collection sessions. In this way, we believe that it is possible to avoid including biases in the recommendation model that use the Q-learning algorithm (e.g., always recommending the same transitions because most students performed them) during the testing phase. Furthermore, we believe that this approach reflects the application of the foot-in-the-door in real-world scenarios, where several exercises must be recommended in a given sequence to correctly apply the technique. It is important to emphasize that, although we used the next item approach and updated the recommendation models after each recommendation, we could easily adapt the algorithms to generate longer sequences.

After performing this simulation, we computed several metrics to evaluate the proposed recommendation models regarding the four aspects of our hypothesis, namely, the students' effort, compliance, grade and engagement. In relation to the first aspect of our hypothesis (students' effort) we used the metric **Proportion of accepted exercises in which students**

increased their effort. This metric counts how many times the perceived effort ratings provided by students in the accepted exercise were greater than the perceived effort rating in the previous exercise, and then normalizes it by the total of accepted recommendations. Therefore, it can be used to assess how well each recommendation model applies the foot-in-the-door technique.

As previously described, we expect to influence students to accept more exercises and, as a consequence, solve more exercises. Therefore, to assess the second aspect of our hypothesis (students' compliance) we use the following metrics:

- **Number of recommendations:** This metric counts how many recommendations were made, regardless of whether they were accepted or not. This metric is important in order to check whether or not the recommendation models are providing enough recommendations and which ones are not.
- **Average number of recommendations per student:** This metric counts how many recommendations were made per student, regardless of whether they were accepted or not. This metric complements the previous metric as it indicates whether or not the recommendation system is capable of making recommendations for all students or only to a few of them.
- **Proportion of accepted recommendations:** This metric counts how many exercises were accepted and normalizes it by the total number of recommendations. This metric is the main metric directly measures the students' compliance.

To assess the third aspect of our hypothesis (students' performance) we use the following metrics:

- **Proportion of accepted exercises that increased students' grade:** This metric counts how many times the grade of the accepted exercise was greater than the grade of the previous exercise, and then normalizes it by the number of accepted exercises. Therefore, this metric can help us to assess the tendency each recommendation system has in improving the students' grade over time.
- **Average Grades:** This metric shows the average grades students achieved in the accepted exercises. Therefore, this metric can help us to assess the general learning outcomes of the accepted exercises.
- **Proportion of accepted exercises in which students had good grades:** This metric counts in how many exercises students have a grade higher than 50, and then normalizes it by the number of accepted recommendations. Therefore, this metric can help us to assess the learning outcomes of the recommended exercises.

According to the discussion presented in Chapter 3, the task engagement can be identified and/or classified by exploiting the relationship between the students' performance and effort. Therefore, to assess the students' engagement, we propose the following metrics:

- **Proportion of accepted exercises that led to engagement:** This metric counts how many accepted exercises had a grade greater than 50 or an effort greater than 3.5, and then normalizes it by the number of accepted recommendations. This definition matches the engagement definition we discussed/adopted in Chapter 3. Therefore, this metric helps to assess the students' engagement on the accepted exercises.

- **Proportion of accepted exercises that led to learning:** This metric counts how many accepted exercises students had a grade greater than 50 and an effort greater than 3.5, and then normalizes it by the number of accepted recommendations. This definition matches the ideal students’ task engagement discussed in Chapter 3. Since the foot-in-the-door technique imposes an effort increase, which further implies in a grade increase, it is only fair to use this metric to assess the quality of the recommended exercises.
- **Proportion of accepted exercises in which effort and grade increased:** This metric counts in how many exercises the students’ effort and grade was higher than in the previous exercise, and then normalizes it by the number of accepted exercises. This metric helps us to assess the recommendation models regarding their tendency to recommend exercises that will lead students to learn. Since the foot-in-the-door technique can start with exercises that fall into the disengagement category, we use this metric to assess the models’ tendency to recommend exercises within the ideal scenario (i.e., learning) over time.

Each one of these metrics was calculated for each of the folds of each of the recommendation models to allow a statistical comparison¹⁰ of the different recommendation models that formalize the foot-in-the-door technique. Our baselines are:

- **Original model:** The original models are recommendation models that do not apply the foot-in-the-door technique. Therefore, they use the same algorithms, the same selection criteria, and the *NoFITD* function. They are used to assess if applying the foot-in-the-door technique can improve the recommendation process.
- **Original sequence:** The original sequence is the one that students followed during the data collection session. This is an interesting baseline because it helps to assess whether or not the recommendation models can improve a given exercise list.
- **Foot-in-the-door applied to the original sequence:** To find out if our recommended sequences present a good performance, we compare them with the best sequence that could have been obtained with our data. That is, we take each one of the sequences followed by the students and build a subsequence that respects our foot-in-the-door formalization (c.f., Section 6.1, Algorithm 3).
- **Random recommendations:** To check if our results are due to the characteristics of our dataset, we compare our models to a random recommendation.

6.3 Results

Our results are divided into four parts. The first part (Section 6.3.1) deals with the main requirement of our application of the foot-in-the-door technique: increasing the students’ effort. The second part (Section 6.3.2) deals with the students’ compliance with the recommendations made. The third part (Section 6.3.3) deals with the possible influences that the formalization of the foot-in-the-door technique in a recommendation system may have on the students’ performance. Finally, in Section 6.3.4, we look at the students’ engagement.

¹⁰All the statistical tests mentioned in this section were done using the statistical language R with the unpaired Student’s T-test (for the parametric data) or the unpaired Mann-Whitney U test (for the non-parametric data) after checking for normality with the Shapiro-Wilk test. The null hypothesis was rejected when $p\text{-value} \leq 0.05$.

6.3.1 Students' effort

As described in Chapter 3, the foot-in-the-door technique is a social influence technique in which requests are made consecutively and with an increasing cost. To evaluate the recommendation models in this regard, we considered the proportion of accepted exercises in which the effort was higher.

In Table 6.2, we see the results obtained through the proportion of accepted exercises in which the effort has increased. In this table, we have grouped the recommendations models according to their algorithm (i.e., Q-learning or Top N) and each row corresponds to a selection criterion previously described in Section 6.1.3. The columns indicate if the model applied the *ZpdFITD* function to make recommendations (column *Z*) or not (column *I*). For instance, the model that makes recommendations based on the Q-learning algorithm, the difficulty selection criterion and the *ZpdFITD* function (referred to as *QLearning+Difficulty+ZpdFITD*), has its performance shown in group *Q-learning*, row *Difficulty*, column *Z*.

This table also shows three types of comparisons:

1. Comparison between the models that formalize the foot-in-the-door (column *Z*) and their respective models without the foot-in-the-door (same row, column *N*). The statistically significant differences are marked by an asterisk (*).
2. Comparison between the original sequence and the recommendation models: The statistically significant differences are indicated by a diamond suit (\diamond). The original sequence value is shown in the row *Original sequence*, column *N*.
3. Comparison between the recommendation models and random recommendations: The resulting value of the random recommendation is shown in the row *Random*, column *N*. The statistically significant differences are indicated by a heart suit (\heartsuit).

In all comparisons, statistically significant differences are highlighted in red when the value obtained with the model is smaller than its baseline value, and in blue if they are larger.

Model	N	\diamond	\heartsuit	Z	*	\diamond	\heartsuit
Q-learning							
– Compliance	0.5	\diamond		0.79	*	\diamond	\heartsuit
– Difficulty	0.53	\diamond		0.76	*	\diamond	\heartsuit
– Grade	0.5	\diamond		0.74	*	\diamond	\heartsuit
– Actions	0.53	\diamond		0.71	*	\diamond	\heartsuit
– Similarity	0.5	\diamond		0.79	*	\diamond	\heartsuit
Top N							
– Compliance	0.86	\diamond	\heartsuit	0.88		\diamond	\heartsuit
– Difficulty	0.93	\diamond	\heartsuit	0.96		\diamond	\heartsuit
– Grade	0.58			0.47			
– Actions	0.73	\diamond	\heartsuit	0.79		\diamond	\heartsuit
– Similarity	0.67	\diamond	\heartsuit	0.77		\diamond	\heartsuit
Original sequence	0.45		\heartsuit	1			
Random	0.52						

Table 6.2: Proportion of accepted exercises on which students presented an effort increase

As we can see in Table 6.2, all of the recommendation models that formalize the foot-in-the-door technique (column Z) using the Q-learning algorithm present a statistically significant difference in relation to the respective original model (*), the original sequence (\diamond) and also to the random recommendations model (\heartsuit).

However, the recommendation model *TopN+Difficulty+ZpdFITD* (group *Top N*, row *Difficulty* and column Z) is the one with the highest effort increase rate (0.96), but as this model does not present statistically significant differences in relation to the original model, we cannot confirm if the performance of this model was caused by the foot-in-the-door function. Furthermore, the performance of this model does not show statistically significant differences compared to a few other models that also use the ZpdFITD function: *QLearning+ComplianceSfun+ZpdFITD* (0.79), *QLearning+SimilaritySfun+ZpdFITD* (0.79), and *TopN+ComplianceSfun+ZpdFITD* (0.88). Therefore, we can consider that, given that increasing the students' effort is an essential requirement to apply the foot-in-the-door technique, all these models are able to apply foot-in-the-door in a satisfactory manner.

As discussed in this section, we have shown that the foot-in-the-door can be integrated with different educational recommendation systems and, as expected, apply it up to a certain extent. However, our main goal in applying the foot-in-the-door technique is to increase the students' compliance, and as a consequence, to increase their engagement and performance. Thus, in the following sections, we also assess the presented models according to other evaluation criteria.

6.3.2 Students' compliance

We hypothesized that formalizing the foot-in-the-door technique into a recommendation system increases the students' compliance and, in the long run, would increase the number of solved exercises. For this, we evaluated the following metrics: proportion of accepted exercises, number of recommendations, and the average number of recommendations per student.

By analyzing the proportion of accepted exercises, presented in Table 6.3 using the same format and color and symbols system used in the previous section, we see that this metric has statistically significant differences and the highest compliance rates for all recommendation models based on the foot-in-the-door use the Q-learning algorithm. Regarding the Top N algorithm, only the models that use the compliance and the item similarity criteria have shown a significant improvement in relation to their respective original models, but their compliance rate is still as good as chance.

To find out which models have the highest proportion of accepted exercises, we compare the models that best apply the foot-in-the-door with the other models. These results are presented in Table 6.4. In this table, each column represents one of the models that best applies the foot-in-the-door technique and the lines represent the models with which they were compared. The asterisks indicate which comparisons resulted in a statistically significant difference. The symbol is highlighted in blue if the baseline value is larger and in red if the baseline value is lower.

According to the results presented in Table 6.4, the recommendation models based on the Q-learning algorithm present the best results when compared to the results obtained with the Top N algorithm in relation to the proportion of accepted exercises. This gives us a first indication that the recommendation models based on the Top N algorithm may not be the best option.

We continue our analysis by analyzing the number of recommendations made by each recommendation model. Given that we carried out an offline evaluation with a dataset containing data from single sessions, we expected that our recommendation model formalizing the foot-in-the-door technique would present a smaller number of recommendations than the models that did not formalize the technique; because out of 15 proposed exercises, we could recommend a

Model	N	◇	♡	Z	*	◇	♡
Q-learning							
– Compliance	0.83			0.92	*		♡
– Difficulty	0.8			0.93	*		♡
– Grade	0.83			0.92	*		♡
– Actions	0.79		♡	0.89	*		♡
– Similarity	0.83			0.92	*		♡
Top N							
– Compliance	0.58		♡	0.85	*		
– Difficulty	0.52		♡	0.61			♡
– Grade	0.57		♡	0.6			
– Actions	0.56		♡	0.6			♡
– Similarity	0.69		♡	0.79	*		
Original sequence	1			1			
Random	0.83						

Table 6.3: Proportion of recommendations accepted

	Q-learning		Top N	
	Compliance	Similarity	Compliance	Difficulty
Q-learning				
– Compliance				*
– Difficulty				*
– Grade				*
– Actions				*
– Similarity				*
Top N				
– Compliance				*
– Difficulty	*	*	*	
– Grade	*	*		
– Actions	*	*	*	
– Similarity	*	*		*

Table 6.4: Indication of a statistically significant difference between the recommendation models that formalize the foot-in-the-door regarding the metric proportion of recommendations accepted

maximum of 7 exercises, one for each Likert point used to rate the students’ effort. However, despite this limitation, we want to see if the recommendations that formalize the foot-in-the-door technique present different results regarding this evaluation criterion.

In Table 6.5, we see that this metric presents statistically significant differences for almost all implementations of the foot-in-the-door. These differences show that, as expected, the recommendation systems based on the foot-in-the-door perform fewer recommendations than their respective original models and the original sequence. Interestingly, the *TopN+Difficulty+ZpdFITD* model is an exception as it does not present a statistically significant difference when compared to its respective original model. A possible explanation is that, since the difficulty is often linked to the exerted effort (c.f., Chapter 3), applying the foot-in-the-door with this criteria might be somewhat redundant. As a matter of fact, in the previous section we have already seen that

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	122	◇		37	*	◇	♣	♡
– Difficulty	104.2	◇		41.6	*	◇		♡
– Grade	119.6	◇		41.6	*	◇		♡
– Actions	97	◇	♡	49.8	*	◇		♡
– Similarity	119	◇		37	*	◇	♣	♡
Top N								
– Compliance	49.6	◇	♡	31.8	*	◇	♣	♡
– Difficulty	44.6	◇	♡	35.2		◇	♣	♡
– Grade	53	◇	♡	10.8	*	◇	♣	♡
– Actions	46.8	◇	♡	36.6	*	◇	♣	♡
– Similarity	66.6	◇	♡	45.6	*	◇		♡
Original sequence	165.2		♡	51.6	*	◇		♡
Random	118.6							

Table 6.5: Total number of recommendations

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	5.98	◇		1.82	*	◇	♣	♡
– Difficulty	5.11	◇		2.04	*	◇		♡
– Grade	5.86	◇		2.04	*	◇		♡
– Actions	4.75	◇	♡	2.45	*	◇		♡
– Similarity	5.83	◇		1.82	*	◇	♣	♡
Top N								
– Compliance	2.43	◇	♡	1.56	*	◇	♣	♡
– Difficulty	2.18	◇	♡	1.72		◇	♣	♡
– Grade	2.59	◇	♡	1.05	*	◇	♣	♡
– Actions	2.3	◇	♡	1.79	*	◇	♣	♡
– Similarity	3.26	◇	♡	2.23	*	◇		♡
Original sequence	8.09		♡	2.52	*	◇		♡
Random	5.82							

Table 6.6: Average number of recommendations per student

this model did not present a statistically significant difference in relation to the proportion of accepted exercises in which students exerted more effort (Table 6.2), which indicates that the original model applies the foot-in-the-door (probably) as well as the model that uses the function *ZpdFITD*.

When we compare the best four foot-in-the-door models with the others, we notice that the number of recommendations made by them is equivalent to the number of recommendations made by all of the other models, which suggests that the coverage of all the models is somewhat equivalent. The only exception to this rule is the model *TopN+Grade+ZpdFITD* that presents a smaller number of recommendations.

On the other hand, when we compare these models with the application of the *ZpdFITD* function over the original sequence (row *Original sequence*, columns *Z*), we find a statistically significant difference for several models, most of them using the Top N algorithm (marked by the club suit ♣). In other words, even though the number of recommendations is smaller than the recommendations made without the foot-in-the-door formalization, it is likely that the other proposed recommendation models have the best performance possible if we consider the characteristics of our dataset. The results also suggest that the models based on the Q-learning algorithm have a slightly greater coverage, even though, as already mentioned, we have not found statistically significant differences between the different models proposed.

As expected, the results obtained when analyzing the average number of recommendations per student (Table 6.6) mirror the results obtained when analyzing the number of recommendations and indicate that the recommendations are being uniformly made (c.f., Annex L).

We have seen that the recommendation models based on the Q-learning algorithm presented the highest proportion of recommendations accepted and also the highest number of recommendations made. These results can be considered a first evidence that, although the model *TopN+Difficulty+ZpdFITD* has the highest proportion of effort increase, it might not be the most beneficial model for the learning process.

6.3.3 Students' performance

In this section, we analyze the effects of the foot-in-the-door vis-a-vis the students' performance. In Chapter 3, we showed that higher students' effort and engagement are related to a better performance. Therefore, we hypothesized that by increasing the effort inherent in the foot-in-the-door technique, we might also improve the students' performance. This hypothesis is evaluated through three metrics: proportion of accepted exercises in which the students' grade increases, proportion of accepted exercises in which students had a good grade, and average grades.

When analyzing the proportion of accepted exercises that led to a grade increase (Table 6.7), we see that the recommendation models based on the foot-in-the-door improved their respective original model (with the exception of the model that uses the *actions* selection criterion). Furthermore, when compared to the original sequence these models also present a statistically significant difference that indicates an improvement in their grades. In turn, when we compare the models that best apply the foot-in-the-door technique with the other models based on the foot-in-the-door (Table 6.8), the results show that, in general, these models do not stand out. The only exception to this rule is the model *TopN+Difficulty+ZpdFITD* whose criterion, as defined in Section 6.1.3, relies on probability of the students having good grades and, therefore, does not surprise us. Overall, these results suggest that by applying the foot-in-the-door technique we can increase the students' grade and, over time, achieve a good final grade.

When observing the average grades (Table 6.9), the models that showed an improvement

Model	N	◇	♡	Z	*	◇	♡
Q-learning							
– Compliance	0.45			0.66	*	◇	♡
– Difficulty	0.48	◇		0.67	*	◇	♡
– Grade	0.46			0.66	*	◇	♡
– Actions	0.5	◇		0.65		◇	♡
– Similarity	0.46			0.67	*	◇	♡
Top N							
– Compliance	0.7	◇	♡	0.74		◇	♡
– Difficulty	0.93	◇	♡	0.95		◇	♡
– Grade	0.29	◇	♡	0.14		◇	♡
– Actions	0.55	◇		0.58		◇	
– Similarity	0.61	◇	♡	0.72		◇	♡
Original sequence	0.41		♡	0.45	*		
Random	0.47						

Table 6.7: Proportion of accepted exercises on which students presented a grade increase

	Q-learning		Top N	
	Compliance	Similarity	Compliance	Difficulty
Q-learning				
– Compliance				*
– Difficulty				*
– Grade				*
– Actions				*
– Similarity				*
Top N				
– Compliance				*
– Difficulty	*	*	*	
– Grade	*	*	*	*
– Actions				*
– Similarity				*

Table 6.8: Indication of a statistically significant difference between the recommendation models that formalize the foot-in-the-door regarding the metric proportion of accepted exercises on which students presented a grade increase

when compared to the original models and sequence are those based on the Q-learning algorithm and the *TopN+Difficulty+ZpdFITD* model, the latter being the one with the best performance in this metric. In fact, as we can see in Table 6.10, when comparing the models that best applied the foot-in-the-door with the other models, the *TopN+Difficulty+ZpdFITD* model stands out. However, as already mentioned, this does not surprise us because its selection criterion is based on the probability of students' achieving a good grade.

Finally, when analyzing the proportion of accepted exercises in which students had a good grade, we can see that only two models that use the Q-learning algorithm showed improvements over the original model, and that the *TopN+Difficulty+ZpdFITD* model is still the one with the highest proportion. However, it is important to note that it is possible that this proportion is

Model	N	◇	♡	Z	*	◇	♡
Q-learning							
– Compliance	49.43	◇		62.22	*	◇	♡
– Difficulty	50.49	◇		64.26	*	◇	♡
– Grade	49.38	◇		64.81	*	◇	♡
– Actions	50.22	◇		58.93	*	◇	♡
– Similarity	49.5	◇		62.2	*	◇	♡
Top N							
– Compliance	49.98			50.46			
– Difficulty	72.31	◇	♡	72.47		◇	♡
– Grade	42.03		♡	46.51			
– Actions	30.83	◇	♡	30.09		◇	♡
– Similarity	63.65	◇	♡	65.56		◇	♡
Original sequence	41.51		♡	45.8	*		
Random	50.4						

Table 6.9: Average grade on accepted exercises

	Q-learning		Top N	
	Compliance	Similarity	Compliance	Difficulty
Q-learning				
– Compliance				*
– Difficulty				*
– Grade			*	*
– Actions				*
– Similarity				*
Top N				
– Compliance				*
– Difficulty	*	*	*	
– Grade				*
– Actions	*	*	*	*
– Similarity				*

Table 6.10: Indication of a statistically significant difference between the recommendation models that formalize the foot-in-the-door regarding the metric average grade on accepted exercises

related to the selection criterion (difficulty) and not to the foot-in-the-door, as the value obtained with this model is the same when we apply and when we do not apply the technique (columns N and Z). Furthermore, as shown in Table 6.11, this model only stands out from the models that, for the most part, have the lowest proportions: $QLearning+Actions+ZpdFITD$ (0.66), $TopN+Grade+ZpdFITD$ (0.29) and $TopN+Actions+ZpdFITD$ (0.3).

Overall, the results presented in this section show that several of the proposed models showed an improvement in relation to the original model and also in relation to the original sequence. Therefore, we found evidence that formalizing the foot-in-the-door technique in a recommendation system can lead to a better student performance.

Model	N	◇	♡	Z	*	◇	♡
Q-learning							
– Compliance	0.47	◇		0.64		◇	
– Difficulty	0.53	◇		0.73	*	◇	♡
– Grade	0.48	◇		0.69	*	◇	♡
– Actions	0.53	◇		0.66		◇	♡
– Similarity	0.47	◇		0.64		◇	
Top N							
– Compliance	0.46			0.46			
– Difficulty	0.83	◇	♡	0.83		◇	♡
– Grade	0.32		♡	0.29			
– Actions	0.3		♡	0.3			♡
– Similarity	0.7	◇	♡	0.73		◇	♡
Original sequence	0.38		♡	0.42	*		
Random	0.47						

Table 6.11: Proportion of accepted exercises on which the students' had good grades

6.3.4 Students' engagement

Model	N	◇	♡	Z	*	◇	♡
Q-learning							
– Compliance	0.25	◇		0.62	*	◇	♡
– Difficulty	0.28	◇		0.56	*	◇	♡
– Grade	0.26	◇		0.57	*	◇	♡
– Actions	0.3	◇		0.52	*	◇	♡
– Similarity	0.26	◇		0.62	*	◇	♡
Top N							
– Compliance	0.67	◇	♡	0.72		◇	♡
– Difficulty	0.89	◇	♡	0.93		◇	♡
– Grade	0.12	◇	♡	0.04		◇	♡
– Actions	0.4	◇	♡	0.46		◇	♡
– Similarity	0.46	◇	♡	0.61		◇	♡
Original sequence	0.18		♡	0.45		◇	♡
Random	0.25						

Table 6.12: Proportion of accepted exercises on which students presented an increase on effort and grade

The last aspect to be evaluated is the students' engagement, which, as shown in Chapter 3, depends on two factors: the students' effort and performance. Therefore, we analyzed three metrics: proportion of accepted exercises in which students increased their effort and grade, proportion of accepted exercises in which students were engaged, and proportion of accepted exercises in which students learned.

The results obtained with the first metric, proportion of accepted exercises in which students increased their effort and grade (Table 6.12), show that all the models based on the Q-learning algorithm improve the model and the original sequence. However, the model that best applied

	Q-learning		Top N	
	Compliance	Similarity	Compliance	Difficulty
Q-learning				
– Compliance				*
– Difficulty				*
– Grade				*
– Actions				*
– Similarity				*
Top N				
– Compliance				*
– Difficulty	*	*	*	
– Grade	*	*	*	*
– Actions			*	*
– Similarity				*

Table 6.13: Indication of a statistically significant difference between the recommendation models that formalize the foot-in-the-door regarding the metric proportion of accepted exercises on which students presented an increase on effort and grade

the foot-in-the-door technique is again the model that presents the best result and, according to the results presented in Table 6.13, it stands out from the rest.

Model	N	◇	♡	Z	*	◇	♡
Q-learning							
– Compliance	0.31	◇		0.42		◇	
– Difficulty	0.34	◇		0.47	*	◇	♡
– Grade	0.3	◇		0.43	*	◇	♡
– Actions	0.34	◇		0.43		◇	♡
– Similarity	0.31	◇		0.41		◇	
Top N							
– Compliance	0.32			0.33			
– Difficulty	0.56	◇	♡	0.55		◇	♡
– Grade	0.25		♡	0.29			
– Actions	0.18		♡	0.19			♡
– Similarity	0.45	◇	♡	0.46		◇	♡
Original sequence	0.25		♡	0.22	*		♡
Random	0.32						

Table 6.14: Proportion of accepted exercises that lead to learning

Regarding the proportion of accepted exercises in which students learned (Table 6.14), only two models improve the original model, but several improved the original sequence. Furthermore, none of the models that best apply foot-in-the-door stand out positively from the rest.

Finally, in relation to the students’ engagement, none of the models presents improvements in relation to the original model nor the original sequence. Since none of the models shows improvements over the random recommendation, we believe that we did not find any statistically significant difference between the models due to the characteristics of our dataset.

The results presented in this section show that several of the proposed models have improved both the original models and sequence regarding the proportion of accepted exercises in which students exerted more effort and had better grades, as well as in relation to the proportion of accepted exercises in which students learned. Therefore, we found evidence that formalizing the foot-in-the-door technique in a recommendation system can improve the students' engagement over time.

6.4 Limitations

As already discussed, we performed an offline assessment based on the dataset described in Chapter 4. This means and/or implies that:

- As the main purpose of our data collection was to measure the students' effort, we did not explicitly apply the foot-in-the-door during our data collection. This means that our data does not fully reflect the behavior of students under the direct influence of the foot-in-the-door. However, as shown in Chapter 4, we observed several cases where it was indirectly applied and, therefore, the results discussed in this thesis can demonstrate the possible benefits of the technique in recommendation systems.
- Our dataset contains mostly exercises in which students were engaged (i.e., either a good grade or a high effort) and exercises in which they got bad grades. As already discussed, it is possible that due to these characteristics we cannot find statistically significant differences in some metrics.
- The dataset contains data for a single session per student (c.f., Chapter 4) and therefore might not reflect the students' behavior during the entire academic year and/or during a course (long time periods).
- Data were collected from students who voluntarily agreed to participate in our study. This means that students were relatively motivated to participate, and perhaps our dataset contains less data on students who should benefit the most from the proposed recommendation system (i.e., students who are not engaged).
- As we carried out our data collection in sessions organized in schools, it is possible that the data does not perfectly reflect the reality because our presence may have been considered as a form of external pressure/motivation. Another related limitation is the session time that was limited by the schools, which did not allow the participants to solve all of the proposed exercises even if they wanted to. However, as shown in Chapter 4, the session time does not seem to have influenced on their engagement while solving the exercises they could.

Given that the aforementioned limitations are the same for all of the students who participated in our data collection sessions, we consider that they are somewhat mitigated by the fact that we only make comparisons between the recommendations made by different models. Furthermore, we carefully interpreted our results because they can be interpreted as the effects of the technique in real life only up to a certain extent. Therefore, we believe we have carried out a reasonable evaluation of the proposed recommendation models.

6.5 Chapter conclusion

In this chapter, we have proposed a formalization of the foot-in-the-door technique into recommendation systems. This formalization takes into account the zone of proximal development [177] and, therefore, gradually increases the cost of the recommended exercises. To assess our proposal, we compared these formalizations with the same recommendation model in its original form (i.e., the same algorithms and selection criteria, but without the foot-in-the-door formalization) and with the sequences followed by the students during the collection sessions.

Our results show that by formalizing the foot-in-the-door technique into a recommendation systems we can have a positive effect on the four aspects we evaluated: students' effort, compliance, performance, and engagement. More specifically, we were able to increase the students' effort more often than the recommendations models that do not apply the foot-in-the-door technique and, as a result of this approach, we have also increased the students' compliance, grades and engagement, specially in tasks where they exert a proper level of effort and achieve good grades.

However, despite having obtained positive results, we cannot accept nor reject our hypothesis due to the limitations present in this work, the main one being our evaluation method (i.e., an offline evaluation rather than a real-life evaluation). Therefore, we conclude that foot-in-the-door technique has the potential to improve the students' effort, compliance, performance and engagement leading them to succeed during the academic year.

In this chapter, we formalize the foot-in-the-door technique in order to use it in a recommendation system. We also performed different comparisons between recommendation models that use or not the formalization of the foot-in-the-door technique, as well as with real data. Our results suggest that the foot-in-the-door can improve the students' effort, compliance, performance and engagement.

Chapter 7

Conclusions and perspectives

In this thesis we proposed the formalization of the foot-in-the-door social influence tactic into a recommendation system in order to influence students to solve more exercises, which according to the theory of commitment has the potential to engage them into learning.

To achieve this goal, we carried out two literature reviews. The first sought to present the recommendation systems and the techniques used, as well as to identify the main challenges of developing an educational recommendation system. Among these challenges, we have considered four of them:

1. **Personalization:** The personalization aspect of our recommendation system comes mainly from our effort models that predicts the effort according to the effort presented by each one of the students. This means that our foot-in-the-door formalizations consider the level of effort each student has exerted and will exert.
2. **Foster learning:** We seek to engage students by applying the foot-in-the-door technique and the theory of commitment. As discussed in Chapter 3, students' engagement is related to better student performance and encompasses the students' effort in a set of tasks.
3. **System acceptance:** Through the foot-in-the-door technique, we seek to influence students to solve more exercises. In other words, in a way, we are influencing students to accept the recommendations made.
4. **Assessment:** As discussed in Chapter 2, this is an inherent challenge to recommendation systems. In this thesis, this challenge was addressed in Chapter 6, where different criteria related to the learning process were analyzed to assess the influences of the foot-in-the-door technique for educational scenarios.

The second literature review we carried out allowed us to distinguish effort, engagement and cognitive load. Effort was defined as a factor related to the actions taken to overcome difficulties. Engagement was defined as a multidimensional construct that encompasses effort. Finally, the cognitive load was defined as the amount of cognitive resources required to execute a task. These definitions can further imply that the effort and the cognitive load happen at the task level, and that the engagement happens during a set of tasks (e.g., session or course level).

By adopting a popular assumption in the cognitive load theory, the one that the mental effort is equivalent to the cognitive load, this literature review also allowed us to identify different data that can be used to measure and predict the students' effort. However, these measures were

mostly proposed based on experiments carried out with adults in non-educational settings, which prompted us to validate them.

However, we could not find any datasets that fully respected the context of our study and the constraints we established – C1 (Privacy and ethics), C2 (Implicit data), and C3 (Real life application). Therefore, we carried out several data collection sessions. The resulting dataset contains subjective, performance, physiological and behavioral data from 102 seventh grade students from five French schools. We intend to make this dataset available to the research community in the near future.

We then exploited this dataset to train several machine learning models to measure and predict the students' effort. We followed three different approaches:

1. **Effort measurement using effort features:** These models were trained using features extracted from data collected during an exercise e to measure the effort on the same exercise e . These models consistently outperformed chance and the state-of-the-art models and, therefore, we consider that our data can reliably measure the effort exerted by teenage students during English exercises.
2. **Effort measurement using engagement features:** We hypothesized that we could build an effort profile that would be representative enough to measure the students effort in an exercise e . Therefore, these models were trained using features extracted from the data collected during exercises $1 - e$ to measure the effort exerted while solving the exercise e . These models consistently outperform chance and the results obtained with the previous approach, which suggest that the engagement features are representative enough to measure the students' effort in a past exercise.
3. **Effort prediction using engagement features:** Since the previous approach showed that we can build an effort profile that is uniform enough to measure the students' effort, we further hypothesized that this profile would also allow the students' effort prediction. Therefore, these models were trained using features extracted from the data collected during exercises $1 - e$ to predict the effort in a future exercise $e + 1$. The resulting models perform consistently better than chance and are as good as the previous effort measurement models, which corroborates our assumption that the engagement features can also be exploited to predict the students' effort.

In all of these three approaches, the interaction features (i.e., a combination of behavioral and performance data easily captured in a virtual learning environments through activity loggers) have as much predictive power than other types of data when it comes to estimating the students' effort in past or future tasks. This is particularly interesting because this type of data fully respects all of the three constraints we established in this thesis and allows our models to be fully exploited in real-life applications. Another interesting finding related to our effort models is that we were able to achieve high accuracy levels without the need of training the effort models using the previous effort ratings as input, suggesting that our measures are reliable enough to estimate the students' effort using only implicit data.

As already mentioned, all of the of these models performed consistently better than chance and the models trained with engagement features performed slightly better than the model trained with effort features, which, in their turn, performed slightly better than the state-of-the-art models. Therefore, we consider that the data collected and the models presented are valid and reliable to be used by middle school students in the context of learning a foreign language.

Consequently, these models were used to formalize the foot-in-the-door technique into an recommendation system. This formalization was then combined with different recommendation algorithms and different scoring functions to allow a comparison between the recommendation models that apply the technique and those that do not, and also with the path that was actually followed by the students. These comparisons were made to assess the following hypothesis: H_1 – Formalizing the foot-in-the-door into a recommendation system can improve the students’ effort, compliance, performance and engagement.

Our results show that the formalizations of the foot-in-the-door technique into a recommendation system can have a positive influence over the students’ effort, compliance, performance, and engagement. More specifically, the recommendations models that formalize the foot-in-the-door technique chose more exercises that were accepted and presented an increase in the students’ effort, performance and engagement.

However, several questions are still open and constitute possible future works that are discussed in the following section.

7.1 Research perspectives

The main open question of this thesis is directly related to its major limitation: as we were unable to carry out a controlled experiment to validate our hypothesis, it was neither confirmed nor rejected. Therefore, given that the offline assessment showed encouraging results, the first proposed future work is to carry out such experiment to validate the hypotheses.

Another research perspectives are related to the fact that we only have session-related data. This means that we did not studied what is the best approach to apply the foot-in-the-door technique more than once, and neither the influence of its use on recommendation systems in the long term. In the long-term, we could also investigate the role different task engagement types play during the recommendation process. For instance, for disengaged students it might be better to recommend exercises they already mastered the required content to motivate them and show that they can solve the proposed exercises, while for students who are struggling it might be better to recommend exercises that will allow them to learn.

We could also extend the recommendation system to other types of learning resources (e.g., articles, videos, etc.), to different learning goals (e.g., review content students are struggling with and/or learn something new), to other types of virtual learning environments (e.g., Massive Open Online Courses, mobile applications, etc.), etc. This would call for an investigation of the relationship between these different types of learning resources and how to exploit them to engage students using the foot-in-the-door technique.

In their turn, the results of these studies could not only be used to guide the improvement of the recommendation models proposed in this thesis, but also to guide research related to other types of educational technologies, such as intelligent tutoring systems. In this scenario, an intelligent agent could apply the foot-in-the-door technique to influence students to solve exercises, read articles, watch videos, etc. This possibility is particularly interesting because, as suggested by Sharma et al. [156], such agents could be the necessary incentive to change the cognitive state of students from, for example, non-engaged (i.e., low effort and low grade) to learning (i.e., high effort and high grade).

Other research questions – that arise from the project in which this thesis is inserted – are related to the inclusion of the professor in the recommendation loop. For example: Will this recommendation system be truly adopted by teachers on a daily basis or is it just another tool available? Can the recommendation system meet the needs and preferences of the student and

teacher? What are the differences that the inclusion of the teacher in the recommendation cycle brings to the learning process?

Chapter 8

Thesis summary in French / Résumé de thèse

Cette thèse fait partie du projet PIA e-Fran METAL, un projet d'analyse de l'apprentissage pour enseigner langues étrangères aux collégiens via plusieurs outils numériques personnalisés. Plus précisément, cette thèse fait partie de l'axe 1.3 du projet dont le but est de réduire le taux d'échec des apprenants en fournissant un tableau de bord pour les enseignants. Ce tableau de bord affiche :

- L'activité globale des apprenants, en mettant l'accent sur leur effort lors de l'exécution des tâches éducatives proposées; et
- Une vue personnalisée des profils des apprenants, qui comprend des recommandations engageantes de tâches pédagogiques dans le but d'aider l'enseignant dans son nouveau rôle de coach pour les apprenants.

8.1 Contexte général

La collecte et l'exploitation de données est un phénomène croissant dans de nombreux domaines, dont l'éducation. Dans ce domaine, l'utilisation des données est prometteuse car elle peut contribuer à faire progresser notre compréhension du processus d'apprentissage et à l'améliorer [61]. De cette promesse a émergé le domaine de la recherche sur l'analyse des données d'apprentissage (Learning Analytics), défini comme la mesure, la collecte, l'analyse et la communication de données sur les élèves et leurs contextes, afin de comprendre et d'optimiser le processus d'apprentissage [158].

Selon Suthers et al. [164] et plusieurs autres chercheurs [159, 11, 10, 176, 85, 164, 67, 61], l'analyse des données d'apprentissage devrait idéalement tirer parti de la technologie pour collecter et traiter les données collectées, mais aussi prendre en compte autant que possible les théories pédagogiques.

Coté technologique

D'un point de vue technologique, nous nous intéressons à la collecte et à l'exploitation de données afin de fournir des informations utiles aux élèves et aux enseignants. Parmi la grande variété de méthodes et d'applications de la littérature, les systèmes de recommandation ont acquis une grande importance [148]. Ces systèmes sont définis comme étant des outils qui suggèrent de manière personnalisée les ressources les plus susceptibles d'intéresser un utilisateur spécifique [144].

Dans le contexte éducatif, l'objectif principal de ces systèmes est de soutenir le processus d'apprentissage. Afin d'atteindre cet objectif, les recommandations doivent être adaptées à chaque élève et/ou groupe afin de répondre à leurs besoins. Par exemple, Fotopoulou et al. [55] proposent un système qui recommande des activités pour aider l'enseignant à améliorer les compétences sociales et émotionnelles de ses élèves, tandis que Pineda et al. [12, 13] proposent un système qui recommande des ressources d'apprentissage afin d'aider les étudiants à surmonter leurs difficultés dans un sujet donné.

Les ressources recommandées, le public cible et les objectifs des systèmes de recommandation pédagogique sont divers, ainsi que leurs contextes d'application [172]. Cependant, à notre connaissance, les systèmes de recommandations pédagogiques n'ont encore jamais explicitement ciblé l'engagement des élèves. Pour cette raison, en respectant le contexte du projet PIA e-Fran METAL, nous proposons un système de recommandation pédagogique dont l'objectif est d'engager les élèves.

Coté éducatif

D'un point de vue pédagogique, l'engagement des élèves est essentiel pour profiter de ce que l'école a à offrir et ainsi acquérir les compétences nécessaires à leur succès [56]. En d'autres termes, l'engagement est considéré comme un facteur qui influence positivement l'apprentissage car il peut augmenter les performances d'apprentissage des élèves, augmenter leur maturité et réduire la tendance à l'abandon [160, 57].

En raison de ces avantages, l'engagement a pris une grande importance dans l'étude du processus d'apprentissage. Dans la psychologie sociale, l'engagement est défini comme « la connexion de l'individu à des actes comportementaux » [94]. La théorie associée à ce concept – la théorie de l'engagement [93] – soutient que ce ne sont pas nos croyances et convictions qui nous engagent, mais nos actions.

L'une des diverses raisons pour lesquelles une personne peut s'engager est appelée « soumission librement consentie ». De nombreuses études menées sous ce paradigme montrent qu'il est possible d'influencer un sujet pour qu'il change ses croyances, ses choix et ses comportements sans recourir à l'argumentation, à la récompense ou à la punition. Cette caractéristique rend le paradigme intéressant pour une utilisation dans le contexte éducatif [87]. Plus précisément, Joule et Almeida [87] proposent cinq techniques à utiliser comme outil pédagogique. Parmi elles, la technique du pied-dans-la-porte, qui consiste à faire des requêtes consécutives avec un coût croissant [59]. Cette technique semble particulièrement pertinente dans un contexte pédagogique car elle est compatible avec la théorie de la zone de développement proximal [177], selon laquelle la difficulté des activités éducatives devrait augmenter progressivement.

Dans cette thèse, nous proposons d'utiliser la technique du pied-dans-la-porte pour influencer les étudiants à réaliser leurs activités. En considérant simultanément la technique du pied-dans-la-porte et la théorie de l'engagement, nous nous attendons à ce que l'utilisation de la technique du pied-dans-la-porte influence les étudiants à réaliser des activités (actions) et, par conséquent,

à s'engager.

La technologie au service de l'éducation

D'un point de vue technologique, nous nous intéressons aux systèmes de recommandation car ils recommandent des activités susceptibles d'être pertinentes pour les étudiants et peuvent être utilisés pour améliorer de nombreux aspects du processus d'apprentissage, dont l'engagement des étudiants.

D'un point de vue pédagogique, nous nous intéressons à la technique du pied-dans-la-porte car nous pensons qu'elle peut être utilisée pour engager les étudiants et, par conséquent, améliorer leurs résultats d'apprentissage.

Ainsi, en unissant les deux points de vue, nous proposons un système de recommandation qui exploite la technique du pied-dans-la-porte pour choisir les meilleures activités à recommander. Pour ce faire, nous définissons le coût de chaque activité comme l'effort des élèves car chaque tâche nécessite un niveau d'effort différent pour être accomplie. De plus, l'effort a également été cité comme un facteur clé de la réussite des apprenants [28, 64, 169, 115] et par conséquent, il semble approprié pour identifier quelle tâche d'apprentissage est la plus exigeante et aussi pour favoriser l'apprentissage.

8.2 Définition du problème

Notre question de recherche peut ainsi être définie comme suit : **Dans quelle mesure la formalisation de la technique du pied-dans-la-porte dans un système de recommandation peut influencer l'effort, la conformité, la performance et l'engagement des étudiants ?**

Pour répondre à cette question, nous cherchons à accepter (ou rejeter) l'hypothèse suivante : *H1 – Formaliser le pied-dans-la-porte par un système de recommandation peut améliorer l'effort, la conformité, la performance et l'engagement des étudiants.*

L'évaluation de cette hypothèse nécessite l'utilisation d'un jeu de données adapté à notre objectif et au contexte du projet PIA e-Fran METAL, et la mise en place d'un système de recommandation appliquant la technique du pied-dans-la-porte. Ce scénario soulève également des questions spécifiques :

- *Qu'est-ce que l'effort ? Quelles sont les différences entre l'effort, l'engagement et la charge cognitive ?*

L'effort des élèves et leur engagement sont des concepts associés aux comportements souhaités. Les deux concepts sont souvent utilisés sans être définis [39, 72, 140] et, lorsque des définitions sont proposées, elles se chevauchent souvent [57, 72, 28, 152]. De plus, on peut trouver dans la littérature un autre concept, la charge cognitive, qui est souvent considéré comme équivalente à l'effort mental [129, 105].

Par conséquent, définir et distinguer ces trois concepts les uns des autres est une étape importante pour répondre à notre problématique.

- *Comment estimer l'effort que les élèves ont exercé dans leurs activités d'apprentissage précédentes ?*

L'un des premiers prérequis pour la mise en œuvre de la technique du pied-dans-la-porte est de définir le coût des tâches pour les proposer dans le bon ordre. Comme défini précédemment, dans cette thèse, nous considérons la quantité d'effort que les étudiants devaient

exercer pour réaliser une tâche donnée comme le coût d'une tâche. Cela signifie que nous devons être en mesure de quantifier cet effort, ce qui présente un certain nombre de difficultés.

La mesure de l'effort la plus courante, les (auto-)évaluations effectuées par les étudiants et/ou les enseignants [72, 58], sont chronophages et nécessitent une saisie manuelle. Dans certains scénarios, tels que l'apprentissage en ligne, demander aux enseignants d'évaluer les efforts de leurs élèves est impossible car leur interaction avec les élèves est pratiquement inexistante. De plus, ce type de données n'est pas idéal pour être utilisé dans un système de recommandation car il peut compromettre la qualité des recommandations en raison de problèmes tels que le démarrage à froid et la rareté des données [144].

D'un autre côté, les mesures objectives posent également des problèmes. Par exemple, certaines métriques, telles que le temps passé sur une tâche, peuvent conduire à des résultats contradictoires [74, 155, 78] ; tandis que d'autres métriques, telles que le nombre de documents consultés [152], sont capturées sur une période de temps plus longue que celle d'une tâche individuelle, et ne conviennent donc pas à notre objectif.

- *Comment estimer l'effort que les élèves vont déployer dans les futures activités d'apprentissage ?*

Dans le cadre des systèmes de recommandation, l'une des exigences pour formaliser la technique du pied-dans-la-porte est de définir le coût d'une activité a dans le futur en utilisant uniquement des données passées afin de recommander les meilleures activités. En d'autres termes, comment pourrions-nous utiliser les mesures d'effort pour prédire l'effort qu'une activité exigera d'un élève ?

Dans la littérature, nous n'avons trouvé que deux travaux qui cherchent à prédire une certaine forme d'effort dans des tâches futures. Le premier travail cherche à identifier si les étudiants continueront ou non à faire des efforts dans leur prochaine tâche [156]. La deuxième étude tente quant à elle de prédire l'effort qu'un étudiant exercera dans une tâche future un modèle différent pour chaque type de tâche [89]. Ces deux travaux présentent quelques limites par rapport à notre objectif principal. La première étude ne permet pas de distinguer entre différents niveaux d'effort, tandis que la seconde est profondément liée aux tâches de programmation et ne peut être facilement transposée dans le contexte de notre étude.

Enfin, afin d'atteindre notre objectif dans le cadre de l'axe 1.3 du projet PIA e-Fran METAL tout en respectant les lois et principes éthiques en vigueur, nous avons établi les contraintes suivantes :

- **C1 (Confidentialité et éthique)** : Tel qu'imposé par le Règlement Général sur la Protection des Données (RGPD) appliqué en France, pays où cette thèse a été réalisée, les aspects vie privée et éthique doivent être respectés tout au long de la thèse. Une autre motivation de cette contrainte est la prise de conscience de la population mondiale sur ces questions, notamment après le scandale Cambridge Analytics en 2018 [116].
- **C2 (Données implicites)** : Comme mentionné précédemment, demander aux élèves et/ou aux enseignants d'évaluer l'effort fourni prend beaucoup de temps et peut compromettre la qualité des recommandations. Ainsi, afin de garantir la satisfaction des utilisateurs et la fiabilité des modèles proposés, nous fixons comme contrainte de nous appuyer sur des données implicites.

- **C3 (Application dans la vie réelle)** : En raison du contexte imposé par le projet PIA e-Fran METAL, nous devons nous assurer que notre système de recommandation peut être utilisé par les élèves du secondaire et que les recommandations leur sont adaptées. Cela implique de considérer les caractéristiques spécifiques de ces étudiants et de s'assurer que les sources de données choisies sont cohérentes avec les scénarios de la vie réelle.

8.3 Contributions

Nous avons proposé la formalisation de la technique d'influence sociale du pied-dans-la-porte dans un système de recommandation afin d'inciter les étudiants à résoudre plus d'exercices, ce qui, selon la théorie de l'engagement, a le potentiel de les engager. Pour atteindre cet objectif, nous avons réalisé deux travaux bibliographiques. Le premier visait à présenter les systèmes de recommandation et les techniques utilisées, ainsi qu'à identifier les principaux défis du développement d'un système de recommandation pédagogique. Parmi ces défis, nous en avons retenu quatre :

1. **Personnalisation** : La personnalisation de notre système de recommandation provient principalement de nos modèles d'effort qui prédisent l'effort en fonction de l'effort présenté par chacun des étudiants. Cela signifie que nos formalisations du pied-dans-la-porte tiennent compte du niveau d'effort que chaque élève a exercé et exercera.
2. **Favoriser l'apprentissage** : Nous cherchons à engager les étudiants en appliquant la technique du pied-dans-la-porte et la théorie de l'engagement.
3. **Acceptation du système** : Grâce à la technique du pied-dans-la-porte, nous cherchons à inciter les étudiants à résoudre plus d'exercices. Autrement dit, nous incitons les étudiants à accepter les recommandations.
4. **Évaluation** : Il s'agit d'un défi inhérent aux systèmes de recommandation. Dans cette thèse, nous avons utilisé différents critères liés au processus d'apprentissage pour évaluer les influences de la technique du pied-dans-la-porte pour les scénarios éducatifs.

Le second travail bibliographique que nous avons réalisé nous a permis d'établir une distinction claire entre l'effort, l'engagement et la charge cognitive. L'effort a été défini comme un facteur lié aux actions prises pour surmonter les difficultés. L'engagement a été défini comme une construction multidimensionnelle qui englobe l'effort. Enfin, la charge cognitive a été définie comme la quantité de ressources cognitives nécessaires pour effectuer la tâche. Ces définitions peuvent en outre impliquer que l'effort et la charge cognitive se produisent au niveau de la tâche, et que l'engagement se produit sur un ensemble de tâches (par exemple, au niveau de la session ou du cours).

En faisant l'hypothèse que l'effort mental est équivalent à la charge cognitive, ce travail nous a également permis d'identifier différentes données qui peuvent être utilisées pour mesurer et prédire l'effort des élèves. Cependant, la plupart de ces mesures ont été proposées à partir d'expérimentations menées auprès d'adultes et non en milieu scolaire, ce qui nous a amenés à les valider dans le cadre de cette thèse.

Cependant, nous n'avons pu trouver aucun jeu de données respectant pleinement le contexte de notre étude et les contraintes que nous avons établies : C1 (Confidentialité et éthique), C2 (Données implicites) et C3 (Application dans la vie réelle). Par conséquent, nous avons effectué plusieurs sessions de collecte de données. Le jeu de données résultant contient des données subjectives, de performance, physiologiques et comportementales de 102 élèves de cinquième

de cinq collèges français. Nous prévoyons de mettre ce jeu de données à la disposition de la communauté des chercheurs dans un avenir proche.

Ensuite, nous avons exploité ce jeu de données pour former plusieurs modèles d'apprentissage automatique afin de mesurer et de prédire l'effort des étudiants. Nous avons suivi trois approches différentes :

1. **Mesure de l'effort à l'aide de motifs d'effort** : L'apprentissage de ces modèles a été effectué à l'aide de motifs extraits de données recueillies lors d'un exercice e pour mesurer l'effort sur ce même exercice e . Ces modèles ont systématiquement dépassé les modèles de l'état-de-l'art et, par conséquent, nous considérons que nos données peuvent mesurer de manière fiable l'effort exercé par les étudiants adolescents lors d'exercices d'anglais.
2. **Mesure de l'effort à l'aide de motifs d'engagement** : Nous avons émis l'hypothèse que nous pourrions construire un profil d'effort qui serait suffisamment uniforme pour mesurer l'effort des élèves dans un exercice e . Par conséquent, ces modèles ont été entraînés à l'aide de caractéristiques extraites des données collectées au cours des exercices 1 à e pour mesurer l'exercice lors de la résolution de l'exercice e . Ces modèles surpassent systématiquement les résultats obtenus avec l'approche précédente, ce qui suggère que les motifs d'engagement sont suffisamment uniformes pour mesurer l'effort des étudiants dans un exercice passé.
3. **Prédiction de l'effort à l'aide de motifs d'engagement** : Étant donné que l'approche précédente a montré que nous pouvons créer un profil d'effort suffisamment uniforme pour mesurer l'effort des étudiants, nous émettons l'hypothèse que ce profil permettrait également la prédiction de l'effort. Par conséquent, l'apprentissage de ces modèles a été effectué à l'aide de caractéristiques extraites des données recueillies lors des exercices 1 à e pour prédire l'effort d'un futur exercice $e+1$. Les modèles résultants sont aussi précis que les modèles de mesure de l'effort précédents, ce qui corrobore notre hypothèse selon laquelle les caractéristiques d'engagement peuvent également être exploitées pour prédire l'effort des étudiants.

Dans chacune de ces trois approches, les motifs d'interaction (c'est-à-dire la combinaison de données comportementales et de performance facilement capturées dans des environnements d'apprentissage virtuels) ont autant de pouvoir prédictif que d'autres types de données lorsqu'il s'agit d'estimer l'effort des élèves dans des tâches passées ou futures. Ceci est particulièrement intéressant car ce type de données respecte pleinement les trois contraintes que nous avons établies dans cette thèse et permet à nos modèles d'être pleinement exploités dans des applications réelles. Une autre découverte intéressante liée à nos modèles d'effort est que nous sommes capables d'atteindre des niveaux de précision élevés sans avoir besoin d'évaluations subjectives, ce qui suggère que nos mesures sont suffisamment fiables pour estimer l'effort des élèves en utilisant uniquement des données implicites.

Comme déjà mentionné, tous les modèles ont obtenu de meilleurs résultats que le hasard et les modèles entraînés avec des motifs d'engagement ont eu des résultats légèrement meilleurs que celui du modèle entraîné avec des motifs d'effort, qui ont eu des résultats légèrement meilleurs que les modèles de l'état-de-l'art. Par conséquent, nous considérons que les données recueillies et les modèles présentés sont valides et suffisamment fiables pour être utilisés par les élèves de collège dans le cadre de l'apprentissage d'une langue étrangère.

Ces modèles ont alors été utilisés pour formaliser la technique du pied-dans-la-porte par un système de recommandation. Cette formalisation a ensuite été combinée avec différents algorithmes de recommandation et différentes fonctions de notation pour permettre une comparaison

entre les modèles de recommandation qui appliquent la technique et ceux qui ne l'appliquent pas, ainsi qu'avec le chemin effectivement suivi par les étudiants.

Nos résultats montrent que la formalisation de la technique du pied-dans-la-porte par un système de recommandation peut avoir une influence positive sur l'effort, la conformité, la performance et l'engagement des étudiants. Plus précisément, les modèles de recommandations qui formalisent la technique du pied-dans-la-porte ont sélectionné plus d'exercices qui ont été acceptés et ont présenté une augmentation de l'effort, de la performance et de l'engagement des étudiants.

8.4 Publications scientifiques

Ce travail a donné lieu à cinq publications scientifiques :

- Moissa, B., Bonnin, G. and Boyer, A., 2019. Exploiting Wearable Technologies to Measure and Predict Students' Effort. In *Perspectives on Wearable Enhanced Learning* (pp. 411-431). Springer, Cham.
- Moissa, B., Bonnin, G., Castagnos, S. and Boyer, A., 2019, March. Modelling students' effort using behavioral data. In *Technology-enhanced & Evidence-based Education & Learning Workshop at LAK (TeEL'19)*.
- Moissa, B., Bonnin, G. and Boyer, A., 2019. Building a student effort dataset: what can we learn from behavioral and physiological data. In *Learning & Student Analytics Conference (LSAC'19)*.
- Moissa, B., Bonnin, G., and Boyer, A., 2020, Towards the exploitation of multimodal data to measure students' mental effort. In *Proceedings of the 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT'20)*.
- Moissa, B., Bonnin, G., and Boyer, A., 2021, Measuring and predicting students' effort: a study on the feasibility of cognitive load measures to real-life scenarios. In *Proceedings of the 16th European Conference on Technology Enhanced Learning (EC-TEL'21)*.

8.5 Note importante

Nous tenons à souligner que durant les deux dernières années de cette thèse, nous faisons face à la pandémie de Covid-19. Cette situation a eu un impact sur ce travail de recherche. La fermeture des écoles, universités et laboratoires de recherche ont rendu difficile le contact avec les parties prenantes du projet; ont compromis la collecte prévue de données supplémentaires; et ont empêché l'évaluation des systèmes de recommandation avec les enseignants et/ou les étudiants.

Bibliography

- [1] ADOMAVICIUS, G., AND TUZHILIN, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17, 6 (2005), 734–749.
- [2] AGGARWAL, C. C. *Recommender Systems: The Textbook*. Springer, Cham, Switzerland, 2016.
- [3] AL-GHURIBI, S. M., AND NOAH, S. A. M. Multi-criteria review-based recommender system – the state of the art. *IEEE Access* 7 (2019), 169446–169468.
- [4] ANTONENKO, P. D., AND KEIL, A. Assessing working memory dynamics with electroencephalography: Implications for research on cognitive load. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 7, pp. 93–111.
- [5] APPLETON, J. J., CHRISTENSON, S. L., AND FURLONG, M. J. Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools* 45, 5 (2008), 369–386.
- [6] ARSHAD, S. Z., WANG, Y., AND CHEN, F. Analysing mouse activity for cognitive load detection. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration* (New York, United States, 2013), H. Shen, R. Smith, J. Paay, P. Calder, and T. Wyeld, Eds., ACM, pp. 115–118.
- [7] AYRES, P. Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction* 16, 5 (2006), 389–400.
- [8] AYRES, P. Subjective measures of cognitive load: What can they reliably measure? In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 2, pp. 9–28.
- [9] AZEVEDO, R. Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist* 50, 1 (2015), 84–94.
- [10] BAKER, R. S., AND INVENTADO, P. S. Educational data mining and learning analytics. In *Learning Analytics: From Research to Practice*, J. A. Larusson and B. White, Eds., 1 ed. Springer, New York, United States, 2014, pp. 61–75.

- [11] BAKER, R. S. J. D., DUVAL, E., STAMPER, J., WILEY, D. A., AND SHUM, S. B. Educational data mining meets learning analytics. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (New York, United States, 2012), ACM.
- [12] BARRIA-PINEDA, J., AKHUSEYINOGLU, K., AND BRUSILOVSKY, P. L. Explaining need-based educational recommendations using interactive open learner models. In *UMAP'19 Adjunct: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (New York, United States, 2019), ACM, pp. 273–277.
- [13] BARRIA-PINEDA, J., AND BRUSILOVSKY, P. Explaining educational recommendations through a concept-level knowledge visualization. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (2019), ACM, New York, United States, pp. 103–104.
- [14] BEATTY, J., AND LUCERO-WAGNER, B. The pupillary system. In *Handbook of Psychophysiology*, J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, Eds., 2 ed. Cambridge University Press, 2000, ch. 6, pp. 142–162.
- [15] BEDENLIER, S., BOND, M., BUNTINS, K., ZAWACKI-RICHTER, O., AND KERRES, M. Facilitating student engagement through educational technology in higher education: A systematic review in the field of arts and humanities. *Australasian Journal of Educational Technology* 36, 4 (2020), 126–150.
- [16] BEHESHTI, A., YAKHCHI, S., MOUSAEIRAD, S., GHAFARI, S. M., GOLUGURI, S. R., AND EDRISI, M. A. Towards cognitive recommender system. *Algorithms* 13, 8 (2020), 176.
- [17] BELLMAN, R. The theory of dynamic programming. *Bulletin of the American Mathematical Society* 60, 6 (1954), 503–515.
- [18] BHATT, B., PATEL, P. J., AND GAUDANI, H. A review paper on machine learning based recommendation system. *International Journal of Engineering Development and Research* 2, 4 (2014), 3955–3961.
- [19] BLUMENFELD, P., MODELL, J., BARTKO, W. T., SECADA, W. G., FREDRICKS, J. A., FRIEDEL, J., AND PARIS, A. School engagement of inner-city students during middle childhood. In *Developmental Pathways Through Middle Childhood: Rethinking Contexts and Diversity as Resources*, C. R. Cooper, C. T. G. Coll, W. T. Bartko, H. M. Davis, and C. Chatman, Eds., 1 ed. Psychology Press, New York, United States, 2005, ch. 7, pp. 145–170.
- [20] BOBADILLA, J., HERNANDO, A., ORTEGA, F., AND BERNAL, J. A framework for collaborative filtering recommender systems. *Expert Systems with Applications* 38, 12 (2011), 14609–14623.
- [21] BORYS, M., PLECHAWSKA-WÓJCIK, M., WAWRZYK, M., AND WESOŁOWSKA, K. Classifying cognitive workload using eye activity and eeg features in arithmetic tasks. In *Information and Software Technologies: ICIST 2017* (Cham, Switzerland, 2017), R. Damaševičius and V. Mikašytė, Eds., vol. 756 of *Communications in Computer and Information Science*, Springer, pp. 90–105.

-
- [22] BOURAGA, S., JURETA, I., FAULKNER, S., AND HERSSENS, C. Knowledge-based recommendation systems: A survey. *International Journal of Intelligent Information Technologies 10*, 2 (2014), 1–19.
- [23] BURKE, R. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction 12* (User Modeling and User-Adapted Interaction), 331–370.
- [24] BURKE, R., AND ABDOLLAHPOURI, H. Educational recommendation with multiple stakeholders. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)* (2016), IEEE, pp. 62–63.
- [25] BURKE, R., ABDOLLAHPOURI, H., MOBASHER, B., AND GUPTA, T. Towards multi-stakeholder utility evaluation of recommender systems. In *SOAP 2016: Workshop on Surprise, Opposition, and Obstruction in Adaptive and Personalized Systems* (2016).
- [26] CAMP, G., PAAS, F., RIKERS, R., AND VAN MERRIENBOER, J. Dynamic problem selection in air traffic control training: a comparison between performance, mental effort and mental efficiency. *Computers in Human Behavior 17*, 5–6 (2001), 575–595.
- [27] CAÑIZALES, Y. A. M., DA SILVA, V. A. A., AO GOMES JR., J., VITOR, M. A., MARTINS, A. F., AND DE SOUZA, J. F. Evaluating educational recommendation systems: a systematic mapping. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação* (2020), S. C. C. da Silva Pinto and A. F. de Andrade, Eds., pp. 912–921.
- [28] CARONARO, W. Tracking, students’ effort, and academic achievement. *Sociology of Education 78*, 1 (2005), 27–49.
- [29] CHANDLER, P. Foreword. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, pp. vii–x.
- [30] CHEN, F., ZHOU, J., WANG, Y., YU, K., ARSHAD, S. Z., KHAWAJI, A., AND CONWAY, D. *Robust Multimodal Cognitive Load Measurement*. Springer, Cham, Switzerland, 2016.
- [31] CHEN, F., ZHOU, J., AND YU, K. Multimodal and data-driven cognitive load measurement. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 10, pp. 147–164.
- [32] CIERNIAK, G., SCHEITER, K., AND GERJETS, P. Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior 25*, 2 (2009), 315–324.
- [33] COGNITIVE ENGAGEMENT WITH SELF-REPORT SCALES: REFLECTIONS FROM OVER 20 YEARS OF RESEARCH, M. Greene, b a. *Educational Psychologist 50*, 1 (2015), 14–30.
- [34] COOK, A. E., WEI, W., AND PREZIOSI, M. A. The use of ocular-motor measures in a convergent approach to studying cognitive load. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 8, pp. 112–128.
- [35] DALE, R., ROCHE, J., SNYDER, K., AND MCCALL, R. Exploring action dynamics as an index of paired-associate learning. *PLOS ONE 3*, 3 (2008), e1728.

- [36] DARROW, C. W. The rationale for treating the change in galvanic skin response as a change in conductance. *Psychophysiology* 1, 1 (1964), 31–38.
- [37] DE GREEF, T., LAFEVER, H., VAN OOSTENDORP, H., AND LINDENBERG, J. Eye movement as indicators of mental workload to trigger adaptive automation. In *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience* (Berlin, Germany, 2009), D. D. Schmorrow, I. V. Estabrooke, and M. Grootjen, Eds., vol. 5638 of *Lecture Notes in Computer Science*, Springer, pp. 219–228.
- [38] DE PAULA, L. C., DE OLIVEIRA FASSBINDER, A. G., AND BARBOSA, E. F. A recommendation system to support the students performance in programming contests. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings* (2014), IEEE.
- [39] D’MELLO, S., DIETERLE, E., AND DUCKWORTH, A. Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist* 52, 2 (2017), 104–123.
- [40] DOLINSKI, D. A rock or a hard place: The foot-in-the-face technique for inducing compliance without pressure. *Journal of Applied Social Psychology* 41, 6 (2011), 1514–1537.
- [41] DRACHSLER, H., HUMMEL, H., AND KOPER, R. Recommendations for learners are different: Applying memory-based recommender system techniques to lifelong learning. In *Proceedings of the 1st Workshop on Social Information Retrieval for Technology-Enhanced Learning & Exchange* (2007), pp. 18–26.
- [42] DRACHSLER, H., VERBERT, K., SANTOS, O. C., AND MANOUSELIS, N. Panorama of recommender systems to support learning. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds., 2 ed. Springer, Boston, United States, 2016, pp. 421–451.
- [43] DRUIN, A., FOSS, E., HUTCHINSON, H., GOLUB, E., AND HATLEY, L. Children’s roles using keyword search interfaces at home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, United States, 2010), ACM, pp. 413–422.
- [44] DURALL, E., AND GROS, B. Learning analytics as a metacognitive tool. In *Proceedings of the 6th International Conference on Computer Supported Education (CSEDU 2014)* (2014), pp. 380–384.
- [45] EKSTRAND, M. Challenges in evaluating recommendations for children. In *International Workshop on Children & Recommender Systems* (2017).
- [46] EKSTRAND, M. D., AZPIAZU, I. M., WRIGHT, K. L., AND PERA, M. S. Retrieving and recommending for the classroom: Stakeholders, objectives, resources, and users. In *Proceedings of ComplexRec 2018: Second Workshop on Recommendation in Complex Scenarios* (2018).
- [47] EKSTRAND, M. D., AND KONSTAN, J. A. Recommender systems notation. *Computer Science Faculty Publications and Presentations* (2019), 1–7.
- [48] EYSINK, T. H. S., DE JONG, T., BERTHOLD, K., KOLLOFFEL, B., OPFERMANN, M., AND WOUTERS, P. Learner performance in multimedia learning arrangements: An analysis across instructional approaches. *American Educational Research Journal* 46, 4 (2009), 1107–1149.

-
- [49] FADEL, C., BIALIK, M., AND TRILLING, B. *Four-Dimensional Education: The Competencies Learners Need to Succeed*. Center for Curriculum Redesign, 2015.
- [50] FAZELI, S., DRACHSLER, H., AND SLOEP, P. Applying recommender systems for learning analytics: A tutorial. In *The Handbook of Learning Analytics*, C. Lang, G. Siemens, A. Wise, and D. Gašević, Eds., 1 ed. Society for Learning Analytics Research, 2017, ch. 20, pp. 235–240.
- [51] FELFERNIG, A., FRIEDRICH, G., JANNACH, D., AND ZANKER, M. Constraint-based recommender systems. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds., 2 ed. Springer, Boston, United States, 2015, pp. 161–190.
- [52] FINCHAM, E., WHITELOCK-WAINWRIGHT, A., KOVANOVIĆ, V., JOKSIMOVIĆ, S., VAN STAALDUINEN, J.-P., AND GAŠEVIĆ, D. Counting clicks is not enough: Validating a theorized model of engagement in learning analytics. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (New York, United States, 2019), ACM, pp. 501–510.
- [53] FINK, A., AND NEUBAUER, A. C. Individual differences in time estimation related to cognitive ability, speed of information processing and working memory. *Intelligence* 33, 1 (2005), 5–26.
- [54] FINN, J. D., AND ZIMMER, K. S. Student engagement: What is it? why does it matter? In *Handbook of Research on Student Engagement*, S. L. Christenson, A. L. Reschly, and C. Wylie, Eds. Springer, Boston, United States, 2012, pp. 97–131.
- [55] FOTOPOULOU, E., ZAFEIROPOULOS, A., FEIDAKIS, M., METAFAS, D., AND PAPAVALASSIOU, S. An interactive recommender system based on reinforcement learning for improving emotional competences in educational groups. In *Intelligent Tutoring Systems: ITS 2020* (Cham, Switzerland, 2020), V. Kumar and C. Troussas, Eds., vol. 12149 of *Lecture Notes in Computer Science*, Springer, pp. 248–258.
- [56] FREDRICKS, J. A., BLUMENFELD, P. C., AND PARIS, A. H. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research* 74, 1 (2004), 59–109.
- [57] FREDRICKS, J. A., FILSECKER, M., AND LAWSON, M. A. Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction* 43 (2016), 1–4.
- [58] FREDRICKS, J. A., AND MCCOLSKEY, W. The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of Research on Student Engagement*, S. L. Christenson, A. L. Reschly, and C. Wylie, Eds. Springer, Boston, United States, 2012, pp. 763–782.
- [59] FREEDMAN, J. L., AND FRASER, S. C. Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology* 4, 2 (1966), 195–202.
- [60] GARCIA-MARTINEZ, S., AND HAMOU-LHADJ, A. Educational recommender systems: A pedagogical-focused perspective. In *Multimedia Services in Intelligent Environments: Recommendation Services*, G. A. Tsihrantzis, M. Virvou, and L. C. Jain, Eds. Springer, Heidelberg, Germany, 2013, pp. 113–124.

- [61] GAŠEVIĆ, D., DAWSON, S., AND SIEMENS, G. Let's not forget: Learning analytics are about learning. *TechTrends* 59, 1 (2015), 64–71.
- [62] GHAZANFAR, M. A., AND PRÜGEL-BENNETT, A. Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. *Expert Systems with Applications* 41, 7 (2014), 3261–3275.
- [63] GIBBONS, H., AND STAHL, J. Cognitive load reduces visual identity negative priming by disabling the retrieval of task-inappropriate prime information: An ERP study. *Brain Research* 1330 (2010), 101–113.
- [64] GIPPS, C., AND TUNSTALL, P. Effort, ability and the teacher: Young children's explanations for success and failure. *Oxford Review of Education* 24, 2 (1998), 149–165.
- [65] GOLDMAN, M., SEEVER, M., AND SEEVER, M. Social labeling and the foot-in-the-door effect. *The Journal of Social Psychology* 117, 1 (1982), 19–23.
- [66] GREEN, M., ANUYAH, O., KARSANN, D., AND PERA, M. S. Evaluating prediction-based recommenders for kids. In *3rd KidRec Workshop co-located with ACM IDC* (2019).
- [67] GRELLER, W., AND DRACHSLER, H. Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society* 15, 3 (2012), 42–57.
- [68] GUEGUEN, N., AND PASCUAL, A. Evocation of freedom and compliance: The "but you are free of..." technique. *Current Research in Social Psychology* 5, 18 (2000), 264–270.
- [69] GUHE, M., GRAY, W. D., SCHOELLES, M. J., LIAO, W., ZHU, Z., AND JI, Q. Non-intrusive measurement of workload in real-time. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2005), pp. 1157–1161.
- [70] GUTIÉRREZ, P. A., PÉREZ-ORTIZ, M., SÁNCHEZ-MONEDERO, J., FERNÁNDEZ-NAVARRO, F., AND HERVÁS-MARTÍNEZ, C. Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering* 28, 1 (2016), 127–146.
- [71] HART, S. G., AND STAVELAND, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology* 52 (1988), 139–183.
- [72] HENRIE, C. R., HALVERSON, L. R., AND GRAHAM, C. R. Measuring student engagement in technology-mediated learning: A review. *Computers & Education* 90 (2015), 36–53.
- [73] HERBIG, N., DÜWEL, T., HELALI, M., ECKHART, L., SCHUCK, P., CHOUDHURY, S., AND KRÜGER, A. Investigating multi-modal measures for cognitive load detection in e-learning. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (New York, United States, 2020), ACM, pp. 88–97.
- [74] HILL, L. Effort and reward in college: A replication of some puzzling findings. In *Handbook of Replication Research in the Behavioral and Social Sciences* (1990), J. W. Neuliep, Ed., Select Press, pp. 139–149.
- [75] HOSSEINI, R., AND BRUSILOVSKY, P. A study of concept-based similarity approaches for recommending program examples. *New Review of Hypermedia and Multimedia* 23, 3 (2017), 161–188.

-
- [76] HOSSEINI, R., BRUSILOVSKY, P., AND GUERRA, J. Knowledge maximizer: Concept-based adaptive problem sequencing for exam preparation. In *Artificial Intelligence in Education: AIED 2013* (Heidelberg, Germany, 2013), H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds., vol. 7926 of *Lecture Notes in Computer Science*, Springer, pp. 848–851.
- [77] HUIBERS, T., FAILS, J. A., KUCIRKOVA, N., LANDONI, M., MURGIA, E., AND PERA, M. S. 3rd kidrec workshop: What does good look like? In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (New York, United States, 2019), ACM, pp. 681–688.
- [78] HUIBERS, T., AND WESTERVELD, T. Relevance and utility in an educational search environment. In *3rd International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems* (2019).
- [79] HUPTYCH, M., BOHUSLAVEK, M., HLOSTA, M., AND ZDRAHAL, Z. Measures for recommendations based on past students’ activity. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (New York, United States, 2017), ACM, pp. 404–408.
- [80] IBRAHIM, M. S., AND SAIDU, C. I. Recommender systems: Algorithms, evaluation and limitations. *Journal of Advances in Mathematics and Computer Science* 35, 2 (2020), 121–137.
- [81] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. Iso 9241-11:2018(en): Ergonomics of human-system interaction — part 11: Usability: Definitions and concepts. <https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en>, 2018. Accessed 19 April 2021.
- [82] JANNACH, D., MOBASHER, B., AND BERKOVSKY, S. Research directions in session-based and sequential recommendation. *User Modeling and User-Adapted Interaction* 30 (2020), 609–616.
- [83] JAUŠOVEC, N., AND JAUŠOVEC, K. Differences in induced brain activity during the performance of learning and working-memory tasks related to intelligence. *Brain and Cognition* 54, 1 (2004), 65–74.
- [84] J.BOBADILLA, F.ORTEGA, A.HERNANDO, AND A.GUTIÉRREZ. Recommender systems survey. *Knowledge-Based Systems* 46 (2013), 109–132.
- [85] JIVET, I., SCHEFFEL, M., SPECHT, M., AND DRACHSLER, H. License to evaluate: preparing learning analytics dashboards for educational practice. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (New York, United States, 2018), ACM, pp. 31–40.
- [86] JOCHMANN-MANNAK, H., HUIBERS, T., AND SANDERS, T. Children’s information retrieval: beyond examining search strategies and interfaces. In *2nd BCS IRSG Symposium: Future Directions in Information Access 2008* (2008), pp. 64–72.
- [87] JOULE, R.-V., AND DE OLIVEIRA ALMEIDA, A. M. Por uma pedagogia do compromisso. *Psicologia: Teoria e pesquisa* 22, 1 (2006), 35–42.

- [88] KALYUGA, S., AND PLASS, J. L. Cognitive load as a local characteristic of cognitive processes: Implications for measurement approaches. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 5, pp. 59–74.
- [89] KELLEHER, C., AND HNIN, W. Predicting cognitive load in future code puzzles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, United States, 2019), ACM, pp. 1–12.
- [90] KENNEDY, D. O., AND SCHOLEY, A. B. Glucose administration, heart rate and cognitive performance: effects of increasing mental effort. *Psychopharmacology* 149 (2000), 63–71.
- [91] KHAWAJA, M. A., CHEN, F., AND MARCUS, N. Measuring cognitive load using linguistic features: implications for usability evaluation and adaptive interaction design. *International Journal of Human–Computer Interaction* 30, 5 (2014), 343–368.
- [92] KHAWAJA, M. A., RUIZ, N., AND CHEN, F. Potential speech features for cognitive load measurement. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces* (New York, United States, 2007), ACM, pp. 57–60.
- [93] KIESLER, C. A. *The psychology of commitment: Experiments linking behavior to belief*. Academic Press, 1971.
- [94] KIESLER, C. A., AND SAKUMURA, J. A test of a model for commitment. *Journal of Personality and Social Psychology* 3, 3 (1966), 349–353.
- [95] KLAŠNJA-MILIĆEVIĆ, A., VESIN, B., IVANOVIĆ, M., BUDIMAC, Z., AND JAIN, L. C. Recommender systems in e-learning environments. In *E-Learning Systems: Intelligent Techniques for Personalization*, A. Klačnja-Milićević, B. Vesin, M. Ivanović, Z. Budimac, and L. C. Jain, Eds., vol. 112 of *Intelligent Systems Reference Library*. Springer, Cham, Switzerland, 2017, pp. 51–75.
- [96] KLEINKE, C. L. Compliance to requests made by gazing and touching experimenters in field settings. *Journal of Experimental Social Psychology* 13, 3 (1977), 218–223.
- [97] KORBACH, A., BRÜNKEN, R., AND PARK, B. Differentiating different types of cognitive load: a comparison of different measures. *Educational Psychology Review* 30 (2018), 503–529.
- [98] KOREN, Y., BELL, R., AND VOLINSKY, C. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
- [99] KRAMER, A. C. Physiological metrics of mental workload: A review of recent progress, 1990.
- [100] KUMAR, B., AND SHARMA, N. Approaches, issues and challenges in recommender systems: a systematic review. *Indian Journal of Science and Technology* 9, 47 (2016), 1–12.
- [101] LANDONI, M., MATTERI, D., MURGIA, E., HUIBERS, T., AND PERA, M. S. Sonny, cerca! evaluating the impact of using a vocal assistant to search at school. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: CLEF 2019*, F. Crestani,

-
- M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. H. Bürki, L. Cappellato, and N. Ferro, Eds., vol. 11696 of *Lecture Notes in Computer Science*. Springer, Cham, Switzerland, 2019, pp. 101–113.
- [102] LANDONI, M., MURGIA, E., HUIBERS, T., AND PERA, M. S. My name is Sonny, how may I help you searching for information? In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (2019).
- [103] LARMUSEAU, C., VANNESTE, P., DESMET, P., AND DEPAEPE, F. Multichannel data for understanding cognitive affordances during complex problem solving. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (New York, United States, 2019), ACM, pp. 61–70.
- [104] LEDGER, H. The effect cognitive load has on eye blinking. *The Plymouth Student Scientist* 6, 1 (2013), 206–223.
- [105] LEPPINK, J. Cognitive load theory: Practical implications and an important challenge. *Journal of Taibah University Medical Sciences* 12, 5 (2017), 385–391.
- [106] LEPPINK, J., PAAS, F., DER VLEUTEN, C. P. M. V., GOG, T. V., AND MERRIËNBOER, J. J. G. V. Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods* 45 (2013), 1058–1072.
- [107] LIN, Y., LIU, Y., LIN, F., WU, P., ZENG, W., AND MIAO, C. A survey on reinforcement learning for recommender systems, 2021.
- [108] LINNENBRINK, E. A., AND AUL R. PINTRICH. The role of self-efficacy beliefs in student engagement and learning in the classroom. *Reading & Writing Quarterly* 19, 2 (2003), 119–137.
- [109] LINNENBRINK-GARCIA, L., ROGAT, T. K., AND KOSKEY, K. L. K. Affect and engagement during small group instruction. *Contemporary Educational Psychology* 36, 1 (2011), 13–24.
- [110] LIU, M.-C., YU, C.-H., WU, J., LIU, A.-C., AND CHEN, H.-M. Applying learning analytics to deconstruct user engagement by using log data of MOOCs. *Journal of Information Science & Engineering* 34, 5 (2018), 1175–1186.
- [111] MANOUSELIS, N., DRACHSLER, H., VERBERT, K., AND DUVAL, E. Preface. In *Recommender Systems for Learning*, N. Manouselis, H. Drachsler, K. Verbert, and E. Duval, Eds., 1 ed. Springer, New York, United States, 2013.
- [112] MARCHAL, F., CASTAGNOS, S., AND BOYER, A. A first step toward recommendations based on the memory of users. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)* (2016), IEEE, pp. 54–61.
- [113] MARTIN, S. A critical analysis of the theoretical construction and empirical measurement of cognitive load. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 3, pp. 29–44.
- [114] MCKINSTRY, C., DALE, R., AND SPIVEY, M. J. Action dynamics reveal parallel competition in decision making. *Psychological Science* 19, 1 (2008), 22–24.

- [115] MELTZER, L., KATZIR-COHEN, T., MILLER, L., AND RODITI, B. The impact of effort and strategy use on academic performance: Student and teacher perceptions. *Learning Disability Quarterly* 24, 2 (2001), 85–98.
- [116] MILANO, S., TADDEO, M., AND FLORIDI, L. Recommender systems and their ethical challenges. *AI & Society* 35 (2020), 957–967.
- [117] MOCK, P., GERJETS, P., TIBUS, M., TRAUTWEIN, U., MÖLLER, K., AND ROSENSTIEL, W. Using touchscreen interaction data to predict cognitive workload. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (New York, United States, 2016), ACM, pp. 349–356.
- [118] MOISSA, B., BONNIN, G., AND BOYER, A. Exploiting wearable technologies to measure and predict students’ effort. In *Perspectives on Wearable Enhanced Learning (WELL): Current Trends, Research, and Practice*, I. Buchem, R. Klamma, and F. Wild, Eds. Springer, Cham, Switzerland, 2019, pp. 411–431.
- [119] MOTIVATION, I., AND ACADEMIC ACHIEVEMENT: WHAT DOES THEIR RELATIONSHIP IMPLY FOR THE CLASSROOM TEACHER? Poonam c. dev. *Remedial and Special Education* 18, 1 (1997), 12–19.
- [120] MULDER, L. J. M. Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology* 34, 2-3 (1992), 205–236.
- [121] MURGIA, E., LANDONI, M., HUIBERS, T., FAILS, J. A., AND PERA, M. S. The seven layers of complexity of recommender systems for children in educational contexts. In *CEUR Workshop Proceedings* (2019), pp. 5–9.
- [122] NABIZADEH, A. H., LEAL, J. P., RAFSANJANI, H. N., AND SHAH, R. R. Learning path personalization and recommendation methods: A survey of the state-of-the-art. *Expert Systems with Applications* 159 (2020), 1–20.
- [123] NAGY, R. P. Tracking and visualising student effort: Evolution of a practical analytics tool for staff and student engagement. *Journal of Learning Analytics* 3, 2 (2016), 164–192.
- [124] NAISMITH, L. M., AND CAVALCANTI, R. B. Validity of cognitive load measures in simulation-based training: a systematic review. *Academic Medicine* 90, 11 (2015), S24–S35.
- [125] NAISMITH, L. M., CHEUNG, J. J., RINGSTED, C., AND CAVALCANTI, R. B. Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education* 49, 8 (2015), 805–814.
- [126] NGUYEN, V. A., NGUYEN, H.-H., NGUYEN, D.-L., AND LE, M.-D. A course recommendation model for students based on learning outcome. *Education and Information Technologies* 26 (2021), 5389–5415.
- [127] NICKEL, P., AND NACHREINER, F. Psychometric properties of the 0.1 Hz component of HRV as an indicator of mental strain. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44, 12 (2000), 2–747–2–750.
- [128] OR, C. K., AND DUFFY, V. G. Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics* 7, 2 (2007), 83–94.

-
- [129] PAAS, F., TUOVINEN, J. E., TABBERS, H., AND GERVEN, P. W. M. V. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist* 38, 1 (2003), 63–71.
- [130] PAAS, F. G. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *1992* 84, 4 (Journal of Educational Psychology), 429–434.
- [131] PAAS, F. G. W. C., AND VAN MERRIËNBOER, J. J. G. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors* 35, 4 (1993), 737–743.
- [132] PAAS, F. G. W. C., AND VAN MERRIËNBOER, J. J. G. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review* 6, 4 (1994), 351–371.
- [133] PARK, B., AND BRÜNKEN, R. Secondary task as a measure of cognitive load. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 6, pp. 75–92.
- [134] PATRON, H., AND LOPEZ, S. Student effort, consistency, and online performance. *Journal of Educators Online* 8, 2 (2011).
- [135] PERA, M. S., WRIGHT, K., AND EKSTRAND, M. D. Recommending texts to children with an expert in the loop. In *Proceedings of the 2nd International Workshop on Children & Recommender Systems* (2018).
- [136] PHAN, D., SIONG, L. Y., PATHIRANA, P. N., AND SENEVIRATNE, A. Smartwatch: Performance evaluation for long-term heart rate monitoring. In *2015 International Symposium on Bioelectronics and Bioinformatics (ISBB)* (2015), pp. 144–147.
- [137] PRATKANIS, A. R. *The Science of Social Influence: Advances and Future Progress*. Psychology Press, New York, United States, 2007.
- [138] QUADRANA, M., CREMONESI, P., AND JANNACH, D. Sequence-aware recommender systems. *ACM Computing Surveys* 51, 4 (2018), 1–36.
- [139] RAMYA, P., BALAKRISHNAN, S. G., AND KANNAN, M. Recommendation system to improve students performance using machine learning. In *Second International Conference on Materials Science and Manufacturing Technology* (2020).
- [140] REDMOND, P., ABAWI, L.-A., BROWN, A., HENDERSON, R., AND HEFFERNAN, A. An online engagement framework for higher education. *Online Learning* 22, 1 (2018), 183–204.
- [141] REEVE, J., AND CHING-MEITSENG. Agency as a fourth aspect of students’ engagement during learning activities. *Contemporary Educational Psychology* 36, 4 (2011), 257–267.
- [142] RESCHLY, A. L., AND CHRISTENSON, S. L. Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In *Handbook of Research on Student Engagement*, S. L. Christenson, A. L. Reschly, and C. Wylie, Eds. Springer, Boston, United States, 2012, pp. 3–19.

- [143] RICCI, F., ROKACH, L., AND SHAPIRA, B. Introduction to recommender systems handbook. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer, Boston, United States, 2011, pp. 1–35.
- [144] RICCI, F., ROKACH, L., AND SHAPIRA, B. Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. Springer, Boston, United States, 2015, pp. 1–34.
- [145] RIVERA, A. C., TAPIA-LEON, M., AND LUJAN-MORA, S. Recommendation systems in education: A systematic mapping study. In *International Conference on Information Theoretic Security (2018)*, Springer, pp. 937–947.
- [146] ROTHSCHILD, M., HORIUCHI, T., AND MAXEY, M. Evaluating “just right” in edtech recommendation systems. In *3rd International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems (2019)*.
- [147] ROZO, R. C. *Suivi de l’engagement des apprenants lors de la construction de cartes mentales à partir de traces d’interaction*. PhD thesis, Université de Lyon, 2019.
- [148] RUIPÉREZ-VALIENTE, J. A., MUÑOZ-MERINO, P. J., LEONY, D., AND DELGADO KLOOS, C. ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan academy platform. *Computers in Human Behavior* 47 (2015), 139–148.
- [149] SALDEN, R. J., PAAS, F., BROERS, N. J., AND VAN MERRIËNBOER, J. J. Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training. *Instructional Science* 32 (2004), 153–172.
- [150] SALOMON, G. Television is “easy” and print is “tough”: The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology* 76, 4 (1984), 647–658.
- [151] SARWAR, B., KARYPIS, G., KONSTAN, J., AND RIEDL, J. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web (2001)*, pp. 285–295.
- [152] SCARIOT, A. P., ANDRADE, F. G., DA SILVA, J. M. C., AND IMRAN, H. Students effort vs. outcome: Analysis through Moodle logs. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT) (2016)*, IEEE, pp. 371–372.
- [153] SCHULTHEIS, H., AND JAMESON, A. Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In *Adaptive Hypermedia and Adaptive Web-Based Systems: AH 2004* (Heidelberg, Germany, 2004), P. M. E. De Bra and W. Nejdl, Eds., vol. 3137 of *Lecture Notes in Computer Science*, Springer, pp. 225–234.
- [154] SCHUMAN, H. Comment: Students’ effort and reward in college settings. *Sociology of Education* 74, 1 (2001), 73–74.
- [155] SCHUMAN, H., WALSH, E., OLSON, C., AND ETHERIDGE, B. Effort and reward: The assumption that college grades are affected by quantity of study*. *Social Forces* 63, 4 (1985), 945–966.

-
- [156] SHARMA, K., PAPAMITSIOU, Z., OLSEN, J. K., AND GIANNAKOS, M. Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (New York, United States, 2020), ACM, pp. 480–489.
- [157] SHI, Y., RUIZ, N., TAIB, R., CHOI, E., AND CHEN, F. Galvanic skin response (GSR) as an index of cognitive load. In *CHI '07 Extended Abstracts on Human Factors in Computing System* (New York, United States, 2007), ACM, pp. 2651–2656.
- [158] SIEMENS, G. Learning Analytics & Knowledge: February 27-march 1, 2011 in banff, alberta. <https://tekri.athabascau.ca/analytics/>, 2010. Accessed on 2021-12-11.
- [159] SIEMENS, G., AND BAKER, R. S. J. D. Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (New York, United States, 2012), ACM, pp. 252–254.
- [160] SINATRA, G. M., HEDDY, B. C., AND LOMBARDI, D. The challenges of defining and measuring student engagement in science. *Educational Psychologist* 50, 1 (2015), 1–13.
- [161] SINGH, A. Is big data the new black gold? <https://www.wired.com/insights/2013/02/is-big-data-the-new-black-gold/>, 2013. Accessed on 10-08-2021.
- [162] SPÜLER, M., WALTER, C., ROSENSTIEL, W., GERJETS, P., MOELLER, K., AND KLEIN, E. EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM* 48, 3 (2016), 267–278.
- [163] STABLES, A., MURAKAMI, K., MCINTOSH, S., AND MARTIN, S. Conceptions of effort among students, teachers and parents within an English secondary school. *Research Papers in Education* 29, 5 (2014), 626–648.
- [164] SUTHERS, D., AND VERBERT, K. Learning analytics as a “middle space”. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (New York, United States, 2013), D. Suthers, K. Verbert, E. Duval, and X. Ochoa, Eds., ACM, pp. 1–4.
- [165] SUTTON, R. S., AND BARTO, A. G. *Introduction to reinforcement learning*. MIT Press, London, England, 1998.
- [166] SUTTON, R. S., AND BARTO, A. G. *Reinforcement learning: An introduction*. MIT Press, London, England, 2018.
- [167] SWELLER, J. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12, 2 (1988), 257–285.
- [168] SWELLER, J. The role of independent measures of load in cognitive load theory. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 1, pp. 3–8.
- [169] SWINTON, O. H. The effect of effort grading on learning. *Economics of Education Review* 29, 6 (2010), 1176–1182.

- [170] TOBII PRO. Types of eye movement. <https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/types-of-eye-movements/>, 2021. Accessed on 2021-09-17.
- [171] URANOWITZ, S. W. Helping and self-attributions: A field experiment. *Journal of Personality and Social Psychology* 31, 5 (1975), 852–854.
- [172] URDANETA-PONTE, M. C., MENDEZ-ZORRILLA, A., AND OLEAGORDIA-RUIZ, I. Recommendation systems for education: Systematic review. *Electronics* 10, 14 (2021), 1611.
- [173] VAN ORDEN, K. F., LIMBERT, W., MAKEIG, S., AND JUNG, T.-P. Eye activity correlates of workload during a visuospatial memory task. *Human Factors* 43, 1 (2001), 111–121.
- [174] VELTMAN, J. A., AND JANSEN, C. The role of operator state assessment in adaptive automation. Tech. rep., TNO Defence Security and Safety Soesterberg, 2005.
- [175] VERREL, J., LÖVDÉN, M., SCHELLENBACH, M., SCHAEFER, S., AND LINDENBERGER, U. Interacting effects of cognitive load and adult age on the regularity of whole-body motion during treadmill walking. *Psychology and Aging* 24, 1 (2009), 75–81.
- [176] VIEIRA, C., PARSONS, P., AND BYRD, V. Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education* 122 (2018), 119–135.
- [177] VYGOTSKY, L. S. *Mind in society: The development of higher psychological processes*. Harvard University Press, 1980.
- [178] WALTER, C., ROSENSTIEL, W., BOGDAN, M., GERJETS, P., AND SPÜLER, M. Online EEG-based workload adaptation of an arithmetic learning environment. *Frontiers in human Neuroscience* 11 (2017).
- [179] WALTER, C., WOLTER, P., ROSENSTIEL, W., BOGDAN, M., AND SPÜLER, M. Towards cross-subject workload prediction. In *Proceedings of the 6th International Brain-Computer Interface Conference 2014* (2014).
- [180] WANG, S., HU, L., WANG, Y., CAO, L., SHENG, Q. Z., AND ORGUN, M. Sequential recommender systems: Challenges, progress and prospects. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (2019), pp. 6332–6338.
- [181] XIE, B., AND SALVENDY, G. Prediction of mental workload in single and multiple tasks environments. *International Journal of Vognitive Ergonomics* 4, 3 (2000), 213–242.
- [182] YEN, C.-H. Exploring the choices for an effective method for cognitive load measurement in asynchronous interaction of e-learning. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 12, pp. 183–198.
- [183] YIN, B., RUIZ, N., CHEN, F., AND KHAWAJA, M. A. Automatic cognitive load detection from speech features. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces* (New York, United States, 2007), ACM, pp. 249–255.

-
- [184] YU, K., EPPS, J., AND CHEN, F. Cognitive load evaluation of handwriting using stroke-level features. In *Proceedings of the 16th international conference on Intelligent user interfaces* (New York, United States, 2011), ACM, pp. 423–426.
- [185] YU, K., EPPS, J., AND CHEN, F. Mental workload classification via online writing features. In *2013 12th International Conference on Document Analysis and Recognition* (2013), IEEE, pp. 1110–1114.
- [186] ZARGARI MARANDI, R., MADELEINE, P., VUILLERME, N., AND SAMANI, A. Heart rate monitoring for the detection of changes in mental demands during computer work. In *World Congress on Medical Physics and Biomedical Engineering 2018*, L. Lhotska, L. Sukupova, I. Lacković, and G. S. Ibbott, Eds., vol. 68/2 of *IFMBE Proceedings*. Springer, Singapore, 2018, pp. 367–370.
- [187] ZHANG, Y., AND CHEN, X. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval* 14, 1 (2018), 1–101.
- [188] ZHANG, Z., BRUN, A., AND BOYER, A. New measures for offline evaluation of learning path recommenders. In *Addressing Global Challenges and Quality Education* (Cham, Switzerland, 2020), C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, and S. M. Dennerlein, Eds., Springer, pp. 259–273.
- [189] ZHAO, X., XIA, L., ZHANG, L., DING, Z., YIN, D., AND TANG, J. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (New York, United States, 2018), ACM, pp. 95–103.
- [190] ZHENG, R. Z., AND GREENBERG, K. The boundary of different approaches in cognitive load measurement: Strengths and limitations. In *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*, R. Z. Zheng, Ed., 1 ed. Routledge, New York, United States, 2017, ch. 4, pp. 45–56.
- [191] ZHENG, Y. Preference corrections: capturing student and instructor perceptions in educational recommendations. *Smart Learning Environments* 6, 1 (2019), 1–15.
- [192] ZHENG, Y., GHANE, N., AND SABOURI, M. Personalized educational learning with multi-stakeholder optimizations. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (2019), pp. 283–289.
- [193] ZHONG, J., XIE, H., AND WANG, F. L. The research trends in recommender systems for e-learning: A systematic review of ssci journal articles from 2014 to 2018. *Asian Association of Open Universities Journal* 14, 1 (2019), 12–27.

Appendix A

Proposed exercises

This appendix presents the exercises participants were invited to solve exactly as they were proposed to the students. To the exercises proposed in French, a translation is provided.

A.1 Exercise 1

Here is a flyer about Buckingham Palace, complete it using : a – an – the – 0 (for article ∅).

Voici un dépliant concernant Buckingham Palace, complète-le en utilisant au choix : a - an - the - 0 (pour article ∅).

Buckingham Palace is Queen's home: it has more than 600 rooms, 80 bathrooms, exceptional dining-room, cinema, swimming-pool and even post office.

Most of tourists take pictures outside but you can also visit palace inside.

Of course you can't visit all rooms, but you can see many of them.

Everyone would like to see Queen but usually, people see her on television.

When you see flag flying outside, it means she is inside Buckingham Palace.

A.2 Exercise 2

Anna and Elinor want to visit a famous museum in London. Complete the conversation using the compounds of some.

Anna et Elinor veulent visiter un célèbre musée de Londres. Complète la conversation en employant les composés de *some*.

Anna: Last week, told me that *The Sherlock Holmes Museum* was fantastic!

Elinor: I love Sherlock Holmes, there is fascinating about him!

Anna: And it is very famous in London: in Baker Street.

Elinor: I think said the museum was only £10 for teenagers.

Anna: That's perfect, I don't have to do next Saturday, what about you?

Elinor: I have planned! Let's go together!

someone	anything	somewhere
something	nothing	

A.3 Exercise 3

Do you know the queen Elizabeth II well? Complete the informations using who – which – that.

Connais-tu bien la reine Elizabeth II ? Complète les informations en utilisant au choix *who - which - that*.

Queen Elizabeth II, is pronounced Queen Elizabeth the second, is famous all over the world. This exceptional woman was born on 21 April 1926 became queen in 1952. She lives in Buckingham Palace is situated near Saint James's Park. She has three sons and one daughter live in the United Kingdom. She has a passion is horse-riding, she loves riding in the country-side. When she is at home she spends hours with her corgis, she knows they love her! That is not surprising because she is a pet-lover; a person loves animals!

A.4 Exercise 4

Complete this study regarding the social network use in the United States using the logical connectors.

Complète cette étude sur l'utilisation des réseaux sociaux aux États-Unis en utilisant les connecteurs logiques.

Most of the young Americans are called "Generation Y"

because it is a reference to a generation born between 1982 and 2004. They know a lot about new technologies,

so their nicknames is "tech savvy". They write

emails, **but** they prefer sending text messages

in order to communicate with their family and

friends. They often like to have the latest phone or computer,

although it is very expensive. They use the social

medias **in order** miss anything from the new trends.

but	in order to	in order not to
because	so	although

A.5 Exercise 5

Complete this article about Muhammed Ali. Use emphasis adding an auxiliary or an exclamation to talk about his inspiring life.

Muhammed Ali was **SUCH A** talented boxer.

WHAT A unique symbol for America!

He **DID** push everyone to be better.

WHAT strength he had!

He was **SO** involved in political life as well

and defended **SUCH** important causes!

He **DID** refuse to fight in the Vietnam War.

SUCH	SO	WHAT A	DID
WHAT	SUCH A		

A.6 Exercise 6

Helen, a 12 year old girl traveling aboard the Titanic, writes in her diary the day after departure. Complete the text by conjugating the verbs in parentheses in the present tense in be + V-ing.

Helen, une jeune fille de 12 ans voyageant à bord du Titanic, écrit dans son journal intime le lendemain du départ. Complète le texte en conjugant les verbes entre parenthèses au présent en be + V-ing.

"Dear Journal,

It is a wonderful night! We (travel) **are travelling** aboard a fantastic ship! In our first-class room, Dad (listen) **is listening** to some music and Mamma (read) **is reading** her book. We brought Myrtle, our cat, with us. What (do) **is** she **doing**? She (sleep) **is sleeping** of course! My brothers (play) **are playing** upstairs, they (not read) **are not reading** like Mamma, they (get) **are getting** ready for their new life!"

A.7 Exercise 7

Thomas tells his friend Louise the story of Romeo and Juliet: complete the dialogue by conjugating the verbs in parentheses in the present simple.

Thomas raconte à son amie Louise l'histoire de *Romeo and Juliet* : complète le dialogue en conjuguant les verbes entre parenthèses au présent simple.

Thomas: Romeo Montague (fall) falls in love with Juliet Capulet at a ball. His friends (not like) don't like the Capulet but Romeo (not listen) doesn't listen to them.

Louise: What (decide) does he decide to do?

Thomas: He (meet) meets Juliet in secret: he (climb) climbs on her balcony!

Louise: What (talk) do they talk about?

Thomas: They (talk) talk about love.

A.8 Exercise 8

Complete Paddington's story by conjugating the verbs in parentheses in the past.

Complète l'histoire de Paddington en conjuguant les verbes entre parenthèses au prétérit.

Paddington (live) lived with his aunt and uncle in Peru. When they (become) became too old to look after him, he (go) went to London and (meet) met a nice family at Paddington station.

He (not have) didn't have a name yet, so the family (decide) decided to call him Paddington. They (ask) asked him a lot of questions: what (like) did he like to eat? Where (want) did he want to sleep?

A.9 Exercise 9

The following sentences are about New Year's resolutions. Complete them using the future in will.

Les phrases suivantes portent sur les résolutions du Nouvel An. Complète-les en utilisant le futur en **will**.

1. What resolutions (make) **will** you **make**?
2. I (work) **will work** harder at school.
3. I (not make) **won't make** stupid excuses if I forget my homework.
4. My brother (help) **will help** my parents around the house.
5. My sister and I (recycle) **will recycle** paper.
6. We (brush) **will brush** our teeth after every meal.

A.10 Exercise 10

Complete the story of Little Red Riding Hood by conjugating the verbs in parentheses in the simple past.

Complète le conte du Petit Chaperon rouge en conjuguant les verbes entre parenthèses au prétérit simple.

Little Red Riding Hood (go) **went** to see her grandma. She (take) **took** a basket to carry food. On her way, she (not meet) **didn't meet** any friends but she (meet) **met** a big bad wolf. He (ask) **asked** her:

- What (put) **did** you **put** in your bag? (take) **Did** you **take** some food?

- Yes, I **did**. It's for my grandma!

Little Red Riding Hood (not know) **didn't know** the wolf wanted to eat her so she (tell) **told** him where she (live) **lived**.

A.11 Exercise 11

Use the passive form to complete this biography of Nelson Mandela.

Utilise la forme passive pour compléter cette biographie de Nelson Mandela.

In 1964, he **was separated** from his family. He **was sent** to prison for twenty-
separate send

seven years. In 1990 he **was released** from prison. In 1993, he **was given** the
release give

Nobel Peace Prize. In 1994, he **was elected** President of South Africa.
elect

Today, he **is remembered** as a hero. He **is known** as "Madiba", which is his
remember know

tribal name. For the future generations, he **will be considered** as a model.
consider

A.12 Exercise 12

Fill in the blanks using the correct tense to express future.

1. I can't go to the restaurant tonight, I (meet) 'm meeting a colleague at 8 p.m.!
2. (Learn) Will you learn another language next year?
3. Hurry up! It's 8:00 clock! You (miss) are going to miss your bus!
4. If the weather is nice, we (go) will go to the beach for the day.
5. (call) I will call you as soon as I (get) get home.
6. When he (be) 's older, he (go) 'll go to the US.

A.13 Exercise 13

Fill in the blanks using the passive voice and the correct tense.

The body (discover) has just been discovered by the police. The victim (kill) was killed yesterday. It happened during a robbery. The bank (not attack) wasn't attacked before closing time. The employees were about to go home when the burglars came in. The employees (tell) were told to lie down on the floor. They (threaten) were threatened, but one person tried to take her phone in her bag. She (shoot) was shot by one of the burglars. The other employees shouted and said he (catch) would be caught by the police. The bank hope the culprits (find) will be found very soon.

A.14 Exercise 14

On Pancake Day, you and your mum are talking together. Recap the conversation using reported speech.

Mum: "Do you want to invite friends today?"

Me: "Of course I do!"

Mum: "Make a list of ingredients for the pancakes!"

Me: "We can do it together."

Mum: "Don't forget to tell your friends to bring maple syrup."

Mum asked me if I wanted to invite friends on that day. I replied of course I did. She told me to make a list of the ingredients. I replied we could do it together. She told me not to forget to tell my friends to bring maple syrup.

A.15 Exercise 15

Change these sentences to express your regrets using the correct tense after **"If only"** and **"I wish"**.

1. I haven't booked my holidays in Miami.

→ If only I my holidays in Miami!

2. I didn't take my sunglasses.

→ I wish I my sunglasses!

3. I didn't get the umbrella.

→ If only I the umbrella!

4. I didn't buy a magazine to read on the beach.

→ I wish I a magazine to read on the beach!

5. I haven't brought my camera.

→ If only I my camera!

Appendix B

Informed consent form

B.1 Original version in French



Nous sommes des chercheurs de l'équipe KIWI (Knowledge, Information and Web Intelligence) du LORIA et nous réalisons une expérience pour comprendre comment mesurer et prédire l'effort des élèves.

Cette expérience utilise une barre oculométrique et une smartwatch pour collecter des données sur l'activité de votre enfant. Ces données sont collectées de manière pseudonymisées et sont d'une part comportementales (ce qu'il a regardé, son utilisation de la souris et du clavier et les mouvements de son poignet) et d'autre part physiologiques (sa fréquence cardiaque et la dilatation de ses pupilles) pendant qu'il effectue des exercices d'anglais.

Nous aimerions inviter votre enfant à participer à cette expérience qui a une durée d'environ 30 min. Pour cela, nous avons besoin de votre autorisation. Merci de bien vouloir compléter l'autorisation ci-dessous et de la rapporter à l'école pour le jour de la passation de l'expérience. Cette expérience ne constitue pas une évaluation des connaissances ou compétences de votre enfant. Elle n'entre pas dans le cadre de son cursus scolaire et les résultats ne seront pas communiqués à l'établissement. Il s'agit uniquement d'étudier les interactions avec les outils informatiques permettant aux élèves de mieux apprendre leurs leçons. Si vous avez des questions ou des remarques, n'hésitez pas à nous contacter à l'adresse ci-dessous :

[ommitted@ommitted.com]

Les conditions générales sont décrites au verso de ce document.

Je, soussigné(e) _____

autorise mon enfant _____

à participer à une expérience destinée à mesurer et prédire l'effort des élèves.

à _____ le ____ / ____ / ____

Signature du responsable légal :

Signature du participant :

Conditions Générales

Dans le texte qui suit, votre enfant sera désigné par “participant” et la ou les personnes qui supervisent l’expérience sont nommées “expérimentateur”.

Droits du participant

La participation à cette expérience est volontaire. Aucune rémunération ne sera versée. Les informations collectées ainsi que toutes les données acquises au cours de l’étude sont strictement confidentielles et pseudonymisées, conformément à la Loi Informatique et Libertés. Les données seront utilisées à des fins de recherche uniquement. Les résultats des analyses pourront faire l’objet d’une publication scientifique, en respectant strictement l’anonymat du participant. Aucune donnée ne sera utilisée à des fins commerciales ou industrielles. Le participant peut quitter l’expérience à tout moment sans donner de raison. A la demande écrite du participant (ou du responsable légal), les données peuvent être effacées et/ou lui être communiquées, sans aucun frais de votre part. Pour cela, il suffit de nous contacter à l’adresse [ommitted@ommitted.com] (n’oubliez pas de préciser votre numéro d’identification transmis lors de l’expérience, les données sont pseudonymisées et sans cela nous ne pourrions pas vous retrouver).

Devoirs du participant

Le participant s’engage à respecter les règles et à suivre les instructions qui lui seront présentées. En outre, il ne doit pas divulguer les instructions des expériences afin de ne pas nuire au bon déroulement des expériences à venir.

Droits et devoirs de l’expérimentateur

L’expérimentateur peut exclure un participant de l’étude en lui précisant le motif, par exemple s’il ne répond plus aux exigences prévues par le protocole. L’expérimentateur s’engage à respecter les droits des participants.

B.2 Translation to English



We are researchers from the KIWI (Knowledge, Information and Web Intelligence) group from LORIA and we are carrying an study to understand how to measure and predict students’ effort.

This study uses an eye tracker and a smartwatch to capture data regarding the activity of your child. This data is collected in a pseudonymised manner while he solves English exercises, and is behavioral (what he has looked at, his mouse and keyboard usage, and his hand movements) and physiological (his heart rate and pupil dilation).

We would like to invite your child to participate in this experience which lasts approximately 30 minutes. For this we need your permission. Please complete the authorization below and return it to the school before the day of the experiment. This experience is not an assessment of your child's knowledge or skills. It is not part of his school curriculum and the results will not be communicated to the institution. It is only a question of studying interactions with computer tools allowing students to better learn their lessons. If you have any questions or comments, please do not hesitate to contact us at the address below:

[ommitted@ommitted.com]

The general conditions are described on the back of this document.

I _____

authorize my child _____

to participate in an experiment designed to measure and predict students' effort.

_____ on _____ / _____ / _____

Signature of the legal guardian:

Signature of the participant:

Terms and conditions

In the following text, your child will be referred to as "participant" and the person(s) who supervise the experiment are referred to as "experimenter".

Participant's rights

Participation in this experience is voluntary. No remuneration will be paid. The information collected and all data acquired during the study are strictly confidential and pseudonymized, in accordance with the Data Protection Act. The data will be used for research purposes only. The results of the analyzes may be the subject of a scientific publication, strictly respecting the anonymity of the participant. No data will be used for commercial or industrial purposes. The participant can leave the experience at any time without giving a reason. At the written request of the participant (or the legal guardian), the data may be erased and / or communicated to him, at no cost. To do this, simply contact us at [ommitted@ommitted.com] (do not forget to specify your identification number transmitted during the experience, the data is pseudonymized and without this we could not find it).

Participant's duties

The participant agrees to respect the rules and to follow the instructions presented to him. In addition, he must not divulge the instructions for the experiments so as not to interfere with the smooth running of future experiments.

Experimenter's rights and duties

The experimenter can exclude a participant from the study by specifying the reason, for example if he no longer meets the requirements of the protocol. The experimenter undertakes to respect the rights of the participants.

Appendix C

Questionnaires

During the data collection sessions, we used 3 questionnaires. We later asked teachers to answer a fourth questionnaire. As none of them complied to our request, we asked another English teacher to share his opinions. All of these questionnaires are presented in this appendix. As this thesis is written in English, we present the questionnaires in English. However, it must be noted they were applied in French because it is participants' native language.

C.1 Questionnaire 1: Original version in French

Answering this questionnaire was the first task students had to do.

Quel âge as-tu ?

Es-tu un garçon ou une fille ?

- Garçon
- Fille

Comment te sens-tu (bien, malade, fatigué, ennuyé, etc.) ?

Quel est ton niveau de fatigue ?

- Vraiment très fatigué
- Très fatigué
- Plutôt fatigué
- Ni fatigué ni en forme
- Plutôt forme
- Très en forme
- Vraiment très en forme

Aimes-tu l'anglais ?

- Oui
- Non

Trouves-tu l'anglais facile ?

- Oui
- Non

As-tu de bonnes notes en anglais ?

- Oui
- Non

C.2 Questionnaire 1: English translation

Answering this questionnaire was the first task students had to do.

What is your age?

Are you a boy or a girl?

- Boy
- Girl

How are you (well, sick, tired, bored, etc.)?

How tired are you?

- Very, very tired
- Very tired
- Tired
- Neither tired nor energetic
- Energetic
- Very energetic
- Very, very energetic

Do you like English?

- Yes
- No

Do you think English is easy?

- Yes
- No

Do you have good grades in English?

- Yes
- No

C.3 Questionnaire 2: Original version in French

This questionnaire was applied after each exercise to assess students' effort on it.

As-tu eu besoin de beaucoup réfléchir pour faire cet exercice ?

- Vraiment très peu
- Très peu
- Un peu
- Ni peu ni beaucoup
- Beaucoup
- Vraiment beaucoup
- Enormément

C.4 Questionnaire 2: English translation

This questionnaire was applied after each exercise to assess students' effort on it.

How much did you have to think to solve this exercise?

- Very, very little
- Very little
- Little
- Neither little nor much
- Much
- Very much
- Very, very much

C.5 Questionnaire 3: Original version in French

This questionnaire was answered by the participants at the end of the session (i.e., when they finished all of the exercises or when their time was up).

Est-ce que quelque chose t'a dérangé pendant l'expérience (bruit, lumière, obscurité, la montre) ? Si oui, quoi ?

Quel est ton niveau de fatigue maintenant ?

- Vraiment très fatigué
- Très fatigué
- Plutôt fatigué
- Ni fatigué ni en forme
- Plutôt forme
- Très en forme
- Vraiment très en forme

C.6 Questionnaire 3: English translation

This questionnaire was answered by the participants at the end of the session (i.e., when they finished all of the exercises or when their time was up).

Did something bother you during the session (noise, lightness, darkness, the smartwatch)? If yes, what?

How tired are you now?

- Very, very tired
- Very tired
- Tired
- Neither tired nor energetic
- Energetic
- Very energetic
- Very, very energetic

C.7 Questionnaire 4: Original version in French

This questionnaire was supposed to be answered by the participants' teachers. However, we did not get an answer from them and another English teacher gave us his point-of-view.

We asked the same two questions for each one of the 15 exercises proposed to the students (c.f., Appendix A). Each questions had the structure presented here.

The exercise being assessed

à la fin de l'année scolaire 2018-2019, ont-ils appris le contenu requis pour résoudre l'exercice ci-dessus ?

- Oui
- Non

Comment évaluez-vous les exercices ci-dessus par rapport à la difficulté, considérant qu'ils ont été proposés aux élèves du 5ème ?

- Facile
- Moyen
- Difficile

C.8 Questionnaire 4: English translation

This questionnaire was supposed to be answered by the participants' teachers. However, we did not get an answer from them and another English teacher gave us his point-of-view.

We asked the same two questions for each one of the 15 exercises proposed to the students (c.f., Appendix A). Each questions had the structure presented here.

The exercise being assessed

At the end of the 2018-2019 school year, have students learned the content required to solve the exercise above?

- Yes
- No

How do you assess the above exercise regarding its difficulty considering they were propose to 7th grade students?

- Easy
- Medium
- Hard

Appendix D

Participation Certificate

The participation certificate presented below is just a representation of the certificate that was given to students. We provide a translated version below of the French one.

D.1 Original version in French

Recto

CERTIFICAT DE PARTICIPATION

Nous exprimons notre sincère reconnaissance à

pour avoir participé à notre expérience.

Nous admirons votre contribution et votre dévouement.

_____ / _____ / _____

[signature des expérimentateurs]



**UNIVERSITÉ
DE LORRAINE**



Loria
Laboratoire lorrain de recherche
en informatique et ses applications



KIWI
KNOWLEDGE, INFORMATION
and WEB INTELLIGENCE



metal

Verso

Conditions Générales

Dans le texte qui suit, votre enfant sera désigné par “participant” et la ou les personnes qui supervisent l’expérience sont nommées “expérimentateur”.

Droits du participant

La participation à cette expérience est volontaire. Aucune rémunération ne sera versée. Les informations collectées ainsi que toutes les données acquises au cours de l’étude sont strictement confidentielles et pseudonymisées, conformément à la Loi Informatique et Libertés. Les données seront utilisées à des fins de recherche uniquement. Les résultats des analyses pourront faire l’objet d’une publication scientifique, en respectant strictement l’anonymat du participant. Aucune donnée ne sera utilisée à des fins commerciales ou industrielles. Le participant peut quitter l’expérience à tout moment sans donner de raison. A la demande écrite du participant (ou du responsable légal), les données peuvent être effacées et/ou lui être communiquées, sans aucun frais de votre part. Pour cela, il suffit de nous contacter à l’adresse [ommitted@ommitted.com] en précisant le code _____, les données sont pseudonymisées et sans cela nous ne pourrions pas vous retrouver).

Devoirs du participant

Le participant s’engage à respecter les règles et à suivre les instructions qui lui seront présentées. En outre, il ne doit pas divulguer les instructions des expériences afin de ne pas nuire au bon déroulement des expériences à venir.

Droits et devoirs de l’expérimentateur

L’expérimentateur peut exclure un participant de l’étude en lui précisant le motif, par exemple s’il ne répond plus aux exigences prévues par le protocole. L’expérimentateur s’engage à respecter les droits des participants.

D.2 English translation

Recto

PARTICIPATION CERTIFICATE

We express our sincere appreciation to

for participating in our study.

We admire your contribution and dedication.

_____ / _____ / _____

[Experimenters' signature]



Verso

Terms and conditions

In the following text, your child will be referred to as "participant" and the people who supervise the experiment are referred to as "experimenter".

Participant's rights

Participation in this experience is voluntary. No remuneration will be paid. The information collected and all data acquired during the study are strictly confidential and pseudonymized, in accordance with the Data Protection Act. The data will be used for research purposes only. The results of the analyzes may be the subject of a scientific publication, strictly respecting the anonymity of the participant. No data will be used for commercial or industrial purposes. The participant can leave the experience at any time without giving a reason. At the written request of the participant (or the legal guardian), the data may be erased and / or communicated to him, at no cost. To do this, simply contact us at [ommitted@ommitted.com] with the code _____ (the data is pseudonymized and without this we could not find it).

Participant duties

The participant agrees to respect the rules and follow the instructions presented to him. In addition, he must not divulge the instructions for the experiments so as not to interfere with the smooth running of future experiments.

Experimenter's rights and duties

The experimenter can exclude a participant from the study by specifying the reason, for example if he no longer meets the requirements of the protocol. The experimenter undertakes to respect the rights of the participants.

Appendix E

Effort indicators

From the data collected in our study, several indicators were computed for each exercise e solved by student s . These indicators are presented in Table E.

Category	Measure	Indicator(s)	Observation
Subjective	Perceived effort	Rating	As answered by the student on Questionnaire 2 (c.f., Appendix C)
	Students perceptions	Tiredness at the beginning of the session, tiredness at the end of the session, something bothered the student (or not),	As answered by the student on Questionnaires 1 and 3 (c.f., Appendix C)
	Teacher perceptions	The exercise's content was taught (or not), exercise difficulty	As answered by an English teacher according to Questionnaire 4 (c.f., Appendix C)
Performance	Score	Score	Computed as the percentage of right answers in the exercise.
	Time	Time	
Physiological	Beats per minute (BPM)	Absolute decrease, absolute increase, relative decrease, relative increase, latency to the peak, minimum, maximum, average, standard deviation, range	We excluded all values that were not in the range 50–120.

Table E.1: Effort indicators description (continues in the next page)

(continuation)

Category	Measure	Indicator(s)	Observation
	Pupil diameter	Absolute decrease, absolute increase, relative decrease, relative increase, latency to the peak, minimum, maximum, average, standard deviation, range	In each sample, we averaged the left and right pupils diameters.
	Fixations	Count, area, frequency, time until first, time until last, time between the first and the last	
	Fixations duration	Minimum, maximum, average, standard deviation, range, sum	
	Time between fixations	Minimum, maximum, average, standard deviation, range	
	Saccades length	Minimum, maximum, average, standard deviation, range, sum	
	Saccades speed	Minimum, maximum, average, standard deviation, range	
Behavioral	Clicks	Count, ratio, frequency, time until the first, time until the last, time between the first and the last	
	Time between clicks	Minimum, maximum, average, standard deviation, range	
	Keystrokes	Count, ratio, frequency, time until the first, time until the last, time between the first and the last	
	Time between keystrokes	Minimum, maximum, average, standard deviation, range	
	Visible keystrokes	Count, ratio, frequency, time until the first, time until the last, time between the first and the last	Visible keystrokes are those not related to control keys (e.g., Ctrl, Shift, Alt, Esc)

Table E.1: Effort indicators description (continues in the next page)

(continuation)

Category	Measure	Indicator(s)	Observation
	Time between visible keystrokes	Minimum, maximum, average, standard deviation, range	
	Backspaces	Count, frequency, time until the first, time until the last, time between the first and the last	
	Time between backspaces	Minimum, maximum, average, standard deviation, range	
	Actions	Count, ratio, frequency, time until the first, time until the last, time between the first and the last	Actions are defined as clicks and keystrokes.
	Time between actions	Minimum, maximum, average, standard deviation, range	
	Attempts	Count	The number of attempts was limited by 2 in the system. Furthermore, it was only considered when it was done in the same page visit.
Others	Subject personal information	Age, gender, school, eye tracker, the student believes he has good grades in English (or not), the student likes English (or not), the student thinks English is easy (or not)	As answered by the student on Questionnaire 1 (c.f., Appendix C) or observed by us during the session
	Exercise metadata	Answer length, complexity, number of inputs, difficulty as defined by the system, contains possible answers (or not), is well structured (or not), type (drag n' drop or fill the blanks)	

Table E.1: Complete list of effort indicators

Appendix F

Correlations

The effort subjective ratings are and ordinal type of data and, according to the Shapiro-Wilk normality test, the scores do not present a normal distribution (p-value < 0.05). Thus, all the values shown in Table F.1 were computed using the Spearman's rank correlation coefficient ρ as implemented in the statistical language R (version 3.6.1).

Category	Measure	Indicator	Spearman's ρ				
			Effort	Grades			
Subjective	Perceived effort	Rating	–	-0.32	*		
		Students perceptions	Tiredness at the beginning of the session	0.05	0.01		
	Tiredness at the end of the session		-0.02	0.04			
	Teacher perceptions	Exercise difficulty	0.08	*	-0.16	*	
Performance	Score	Score	0.00	*	–		
	Time	Time	0.10	*	0.39	*	
Physiological	Beats per minute (BPM)	Absolute decrease	0.01		-0.10	*	
		Absolute increase	0.00		0.11	*	
		Relative decrease	0.01		-0.09	*	
		Relative increase	0.00		0.10	*	
		Latency to the peak	-0.08	*	0.01		
		Minimum	-0.09	*	0.06		
		Maximum	-0.08	*	0.11	*	
		Average	-0.09	*	0.09	*	
		Standard deviation	-0.07		0.09	*	
		Range	-0.04		0.17	*	
		Pupil diameter	Absolute decrease	-0.05		-0.17	*
			Absolute increase	0.04		0.07	
			Relative decrease	-0.07		-0.14	*
			Relative increase	0.05		0.03	
			Latency to the peak	0.06		0.20	*
Minimum	-0.09		*	-0.05			
Maximum	-0.01			0.14	*		
Average	-0.05		0.11	*			

Table F.1: Full effort and scores correlations table (continues in the next page)

(continuation)

Category	Measure	Indicator	Spearman's ρ			
			Effort	Grades		
		Standard deviation	0.01	-0.02		
		Range	0.03	0.13	*	
	Fixations	Count	0.11	*	0.40	*
		Area	0.07	*	0.07	*
		Frequency	0.03		0.15	*
		Time until the first	-0.10	*	-0.04	
		Time until the last	0.10	*	0.38	*
		Time between the first and the last	0.11	*	0.38	*
	Fixations duration	Minimum	0.10	*	-0.14	*
		Maximum	0.05		0.11	*
		Average	-0.02		0.06	
		Standard deviation	0.02		-0.02	
		Range	0.00	*	0.00	*
		Sum	0.08	*	0.37	*
	Time between fixations	Minimum	-0.06		-0.07	*
		Maximum	0.00		-0.01	
		Average	-0.08	*	-0.09	*
		Standard deviation	0.00	*	0.00	*
		Range	0.00		-0.01	
	Saccades length	Minimum	0.04		-0.14	*
		Maximum	0.08	*	0.05	
		Average	0.00		-0.14	*
		Standard deviation	0.00	*	0.00	*
		Range	0.08	*	0.06	
		Sum	0.10	*	0.36	*
	Saccades speed	Minimum	-0.11	*	-0.12	*
		Maximum	0.03		0.10	*
		Average	-0.02		-0.09	*
		Standard deviation	-0.02		-0.07	*
		Range	0.03		0.10	*
		Sum	0.00	*	0.00	*
Behavioral	Clicks	Count	0.01		0.36	*
		Ratio	-0.02		0.35	*
		Frequency	-0.14	*	-0.17	*
		Time until the first	0.10	*	0.13	*
		Time until the last	0.10	*	0.39	*
		Time between the first and the last	0.09	*	0.39	*
	Time between clicks	Minimum	-0.01		0.14	*
		Maximum	0.13	*	0.17	*
		Average	0.13	*	0.20	*

Table F.1: Full effort and scores correlations table (continues in the next page)

(continuation)

Category	Measure	Indicator	<i>Spearman's ρ</i>			
			Effort	Grades		
		Standard deviation	0.15	*	0.11	*
		Range	0.13	*	0.16	*
	Keystrokes	Count	-0.01		0.03	
		Ratio	-0.08	*	0.57	*
		Frequency	-0.01		-0.10	*
		Time until the first	0.05		0.10	*
		Time until the last	-0.01		0.39	*
		Time between the first and the last	-0.02		0.40	*
		Time between keystrokes	Minimum	0.07		0.00
	Maximum		-0.06		0.41	*
	Average		-0.09	*	0.45	*
	Standard deviation		-0.10	*	0.46	*
	Range		-0.06		0.41	*
	Visible keystrokes	Count	-0.01		0.02	
		Ratio	-0.08		0.57	*
		Frequency	-0.01		-0.10	*
		Time until the first	0.05		0.10	*
		Time until the last	-0.01		0.39	*
		Time between the first and the last	-0.02		0.40	*
		Time between visible keystrokes	Minimum	0.07		0.02
	Maximum		-0.05		0.41	*
	Average		-0.09	*	0.45	*
	Standard deviation		-0.10	*	0.46	*
	Range		-0.05		0.41	*
	Backspaces	Count	0.01		0.05	
		Frequency	0.01		-0.01	
		Time until the first	0.01		0.18	*
		Time until the last	0.05		0.28	*
		Time between the first and the last	0.06		0.14	*
	Time between backspaces	Minimum	-0.03		0.10	*
		Maximum	0.02		0.18	*
		Average	-0.05		0.19	*
		Standard deviation	-0.04		0.22	*
		Range	0.02		0.15	*
	Actions	Count	-0.01		0.12	*
		Ratio	-0.05		0.50	*
		Frequency	-0.08	*	-0.18	*

Table F.1: Full effort and scores correlations table (continues in the next page)

(continuation)

Category	Measure	Indicator	Spearman's ρ			
			Effort	Grades		
		Time until the first	0.09	*	0.13	*
		Time until the last	0.10	*	0.39	*
		Time between the first and the last	0.09	*	0.39	*
	Time between actions	Minimum	0.04		0.00	
		Maximum	0.12	*	0.23	*
		Average	0.09	*	0.24	*
		Standard deviation	0.12	*	0.23	*
		Range	0.12	*	0.23	*
	Attempts	Count	0.11	*	-0.06	
Others	Subject personal information	Age	-0.03		-0.05	
		Exercise metadata	Answer length	0.02		-0.37
		Complexity	0.08	*	-0.50	*
		Number of inputs	-0.06		-0.03	
		Difficulty (system)	0.03		-0.12	*

Table F.1: Full effort and scores correlations table

Appendix G

Effort features

We could not use all the data we captured to train the machine learning models because some of them are subjective (c.f., Chapter 4) and demands that students and teachers provide the data, which is time consuming and violates one of our constraints (c.f., Introduction). Thus, in Table G.1, we provide a full list of all the indicators used as input for the effort measurement models (c.f., Chapter 5).

It is important to note that we kept only the effort ratings from the subjective data because they are our ground truth and, just like in any recommendation system, they can be provided only if the student wants to rate a given exercise.

Category	Measure	Feature(s)
Subjective	Perceived effort	Rating
Performance	Score	Score
	Time	Time
Physiological	Beats per minute (BPM)	Absolute decrease, absolute increase, relative decrease, relative increase, latency to the peak, minimum, maximum, median, average, standard deviation, range
	Pupil diameter	Absolute decrease, absolute increase, relative decrease, relative increase, latency to the peak, minimum, maximum, median, average, standard deviation, range
	Fixations	Count, area, frequency, time until first, time until last, time between the first and the last
	Fixations duration	minimum, maximum, median, average, standard deviation, range, sum
	Time between fixations	minimum, maximum, median, average, standard deviation, range
	Saccades length	minimum, maximum, median, average, standard deviation, range, sum

Table G.1: Effort features description (continues in the next page)

(continuation)

Category	Measure	Feature(s)
	Saccades speed	minimum, maximum, median, average, standard deviation, range
Behavioral	Clicks	Count, ratio, frequency, time until the first, time until the last, time between the first and the last
	Time between clicks	minimum, maximum, median, average, standard deviation, range
	Keystrokes	Count, ratio, frequency, time until the first, time until the last, time between the first and the last
	Time between keystrokes	minimum, maximum, median, average, standard deviation, range
	Visible keystrokes	Count, ratio, frequency, time until the first, time until the last, time between the first and the last
	Time between visible keystrokes	minimum, maximum, median, average, standard deviation, range
	Backspaces	Count, frequency, time until the first, time until the last, time between the first and the last
	Time between backspaces	minimum, maximum, median, average, standard deviation, range
	Actions	Count, ratio, frequency, time until the first, time until the last, time between the first and the last
	Time between actions	minimum, maximum, median, average, standard deviation, range
	Attempts	Count
Others	Subject personal information	Age, gender, the student likes English (or not), the student thinks English is easy (or not)
	Exercise metadata	Answer length, complexity, number of inputs, difficulty as defined by the system, contains possible answers (or not), is well structured (or not), type (drag n' drop or fill the blanks)

Table G.1: Complete list of effort description

Appendix H

Engagement features

The effort features can be used to estimate the effort a student t exerted in a task t (c.f., Chapter 5). Unfortunately, they do not allow us to estimate the effort a student s exerted in a task $t + 1$ because it requires that the task to which the effort is being estimated is already solved. Thus, we computed new features called engagement features.

This features were computed taking into account all the data computed during the previously solved exercises. This means that if a student solved 10 exercises until now, his engagement features will be computed taking into account the data collected during these 10 exercises. Obviously that for the purposes of this study, we computed the engagement features after each exercise was solved in order to train a model to predict the effort a student t exerted in a task $t + 1$

Table H.1 shows these features highlighting the new features. For instance, the features related to the clicks were just the count, the ratio, the frequency, the time until the first click, the time until the last click, and the time between the first and the last. Now, each one of these features has been derived into 5 or 6 features as all of them now have a mean, a standard deviation, a minimum, a maximum, a range, and some of them also have a sum.

It is important to note that we computed two new types of features: previous effort and previous exercises. These features describe the previously solved exercises and also the effort exerted on them.

Category	Measure	Feature(s)
Subjective	Perceived effort	Rating (on task $t + 1$)
	Effort on previous tasks	minimum, maximum, median, average, standard deviation, range
Performance	Score	minimum, maximum, median, average, standard deviation, range
	Time	minimum, maximum, median, average, standard deviation, range

Table H.1: Engagement features description (continues in the next page)

(continuation)

Category	Measure	Feature(s)
Physiological	Beats per minute (BPM)	Absolute decrease (minimum, maximum, median, average, standard deviation, range), absolute increase (minimum, maximum, median, average, standard deviation, range), relative decrease (minimum, maximum, median, average, standard deviation, range), relative increase (minimum, maximum, median, average, standard deviation, range), latency to the peak (minimum, maximum, median, average, standard deviation, range), minimum, maximum, median, average, standard deviation, range
	Pupil diameter	Absolute decrease (minimum, maximum, median, average, standard deviation, range), absolute increase (minimum, maximum, median, average, standard deviation, range), relative decrease (minimum, maximum, median, average, standard deviation, range), relative (minimum, maximum, median, average, standard deviation, range), latency to the peak (minimum, maximum, median, average, standard deviation, range), minimum, maximum, median, average, standard deviation, range

Table H.1: Engagement features description (continues in the next page)

(continuation)

Category	Measure	Feature(s)
	Fixations	Count (minimum, maximum, median, average, standard deviation, range, sum), area (minimum, maximum, median, average, standard deviation, range), frequency (minimum, maximum, median, average, standard deviation, range), time until first (minimum, maximum, median, average, standard deviation, range), time until last (minimum, maximum, median, average, standard deviation, range), time between the first and the last (minimum, maximum, median, average, standard deviation, range)
	Fixations duration	minimum, maximum, median, average, standard deviation, range, sum
	Time between fixations	minimum, maximum, median, average, standard deviation, range
	Saccades length	minimum, maximum, median, average, standard deviation, range, sum
	Saccades speed	minimum, maximum, median, average, standard deviation, range
Behavioral	Clicks	Count (minimum, maximum, median, average, standard deviation, range, sum), ratio (minimum, maximum, median, average, standard deviation, range), frequency (minimum, maximum, median, average, standard deviation, range), time until the first (minimum, maximum, median, average, standard deviation, range), time until the last (minimum, maximum, median, average, standard deviation, range), time between the first and the last (minimum, maximum, median, average, standard deviation, range)
	Time between clicks	minimum, maximum, median, average, standard deviation, range

Table H.1: Engagement features description (continues in the next page)

(continuation)

Category	Measure	Feature(s)
	Keystrokes	Count (minimum, maximum, median, average, standard deviation, range, sum), ratio (minimum, maximum, median, average, standard deviation, range), frequency (minimum, maximum, median, average, standard deviation, range), time until the first (minimum, maximum, median, average, standard deviation, range), time until the last (minimum, maximum, median, average, standard deviation, range), time between the first and the last (minimum, maximum, median, average, standard deviation, range)
	Time between keystrokes	minimum, maximum, median, average, standard deviation, range
	Visible keystrokes	Count (minimum, maximum, median, average, standard deviation, range, sum), ratio (minimum, maximum, median, average, standard deviation, range), frequency (minimum, maximum, median, average, standard deviation, range), time until the first (minimum, maximum, median, average, standard deviation, range), time until the last (minimum, maximum, median, average, standard deviation, range), time between the first and the last (minimum, maximum, median, average, standard deviation, range)
	Time between visible keystrokes	minimum, maximum, median, average, standard deviation, range

Table H.1: Engagement features description (continues in the next page)

(continuation)

Category	Measure	Feature(s)
	Backspaces	Count (minimum, maximum, median, average, standard deviation, range, sum), frequency (minimum, maximum, median, average, standard deviation, range), time until the first (minimum, maximum, median, average, standard deviation, range), time until the last (minimum, maximum, median, average, standard deviation, range), time between the first and the last (minimum, maximum, median, average, standard deviation, range)
	Time between backspaces	minimum, maximum, median, average, standard deviation, range
	Actions	Count (minimum, maximum, median, average, standard deviation, range, sum), ratio (minimum, maximum, median, average, standard deviation, range), frequency (minimum, maximum, median, average, standard deviation, range), time until the first (minimum, maximum, median, average, standard deviation, range), time until the last (minimum, maximum, median, average, standard deviation, range), time between the first and the last (minimum, maximum, median, average, standard deviation, range)
	Time between actions	minimum, maximum, median, average, standard deviation, range
Others	Subject personal information	Age, gender, the student likes English (or not), the student thinks English is easy (or not)
	Exercise metadata (task $t + 1$)	Answer length, complexity, number of inputs, difficulty as defined by the system, contains possible answers (or not), is well structured (or not), type (drag n' drop or fill the blanks)

Table H.1: Engagement features description (continues in the next page)

(continuation)

Category	Measure	Feature(s)
	Previous exercises metadata	Answer length(minimum, maximum, median, average, standard deviation, range), complexity(minimum, maximum, median, average, standard deviation, range), number of inputs(minimum, maximum, median, average, standard deviation, range), difficulty as defined by the system(minimum, maximum, median, average, standard deviation, range), contains possible answers (or not)(minimum, maximum, median, average, standard deviation, range), is well structured (or not)(minimum, maximum, median, average, standard deviation, range), type (drag n' drop or fill the blanks) (minimum, maximum, median, average, standard deviation, range)

Table H.1: Complete list of engagement features

Appendix I

Full results of the effort measurement using the effort features

I.1 Classification results

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Dummy	10	0.13 #	0.17 #	0.44 #
Borys et al. [21]	22	0.25 *	0.29 *	0.56 *
Empty	10	0.27 *	0.35 *	0.6 *
Score	11	0.29 *	0.33 *	0.61 *
Time	11	0.28 *	0.34 *	0.6 *
Heart rate	15	0.26 *	0.33 *	0.6 *
Pupil diameter	20	0.29 *	0.36 *	0.61 *
Fixations	14	0.28 *	0.35 *	0.61 *
Fixations duration	14	0.29 *	0.36 *	0.62 *
Time between fixations	14	0.28 *	0.35 *	0.61 *
Distance between fixation points	15	0.27 *	0.32 *	0.6 *
Fixation points change speed	13	0.28 *	0.37 *	0.61 *
Clicks	13	0.27 *	0.36 *	0.63 *
Time between clicks	14	0.27 *	0.36 *	0.61 *
Keystrokes	11	0.27 *	0.34 *	0.6 *
Time between keystrokes	14	0.29 *	0.37 *	0.61 *
Visible keystrokes	10	0.27 *	0.34 *	0.59 *
Time between visible keystrokes	10	0.27 *	0.34 *	0.6 *
Backspaces	14	0.28 *	0.35 *	0.6 *
Time between backspaces	14	0.28 *	0.35 *	0.61 *
Actions	13	0.27 *	0.36 *	0.62 *
Time between actions	15	0.27 *	0.36 *	0.61 *
Attempts	11	0.28 *	0.35 *	0.6 *
Borys et al. [21] (types)	40	0.31 * #	0.37 * #	0.63 * #
Herbig et al. [73] (types)	45	0.3 * #	0.37 * #	0.62 * #
Eye activity	30	0.3 * #	0.36 * #	0.61 * #
Interactions	41	0.32 * #	0.36 * #	0.66 * #

Table I.1: Classification’s accuracy while measuring students’ effort with effort features(continues in the next page)

(continuation)

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Performance	12	0.29 *	0.34 *	0.63 *
Behavioral	39	0.28 *	0.35 *	0.63 *
Physiological	45	0.31 * #	0.36 * #	0.63 * #
All	76	0.33 * #	0.39 * #	0.65 * #

Table I.1: Classification’s accuracy while measuring students’ effort with effort features

I.2 Ordinal regression results

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Dummy	10	0.13 #	0.17 #	0.44 #
Borys et al. [21]	22	0.25 *	0.29 *	0.56 *
Empty	10	0.15 #	0.28 *	0.57 *
Score	11	0.17 * #	0.28 *	0.59 *
Time	11	0.15 * #	0.28 *	0.57 *
Heart rate	15	0.16 * #	0.29 *	0.58 *
Pupil diameter	20	0.16 * #	0.29 *	0.58 *
Fixations	14	0.16 * #	0.28 *	0.58 *
Fixations duration	14	0.16 * #	0.29 *	0.58 *
Time between fixations	14	0.15 * #	0.28 *	0.57 *
Distance between fixation points	15	0.15 * #	0.28 *	0.57 *
Fixation points change speed	13	0.15 * #	0.28 *	0.58 *
Clicks	13	0.15 * #	0.28 *	0.57 *
Time between clicks	14	0.16 * #	0.29 *	0.58 *
Keystrokes	11	0.15 #	0.28 *	0.57 *
Time between keystrokes	14	0.15 #	0.28 *	0.57 *
Visible keystrokes	10	0.15 #	0.28 *	0.57 *
Time between visible keystrokes	10	0.15 #	0.28 *	0.57 *
Backspaces	14	0.15 #	0.28 *	0.57 *
Time between backspaces	14	0.15 #	0.28 *	0.57 *
Actions	13	0.15 * #	0.28 *	0.58 *
Time between actions	15	0.17 * #	0.29 *	0.59 *
Attempts	11	0.15 * #	0.28 *	0.57 *
Borys et al. [21] (types)	40	0.17 * #	0.28 *	0.58 *
Herbig et al. [73] (types)	45	0.17 * #	0.29 *	0.58 *
Eye activity	30	0.17 * #	0.29 *	0.58 *
Interactions	41	0.21 *	0.3 *	0.61 * #
Performance	12	0.19 * #	0.3 *	0.61 * #
Behavioral	39	0.18 * #	0.29 *	0.59 *
Physiological	45	0.17 * #	0.29 *	0.58 *
All	76	0.23 * #	0.32 *	0.61 * #

Table I.2: Ordinal regression’s accuracy while measuring students’ effort with effort features

I.3 Regression results

Subset	#Features	RMSE		
Dummy	10	1.66		
Herbig et al. [73]	26	1.87		
Empty	10	1.6	*	
Score	11	1.53	*	#
Time	11	1.59	*	
Heart rate	15	1.6	*	
Pupil diameter	20	1.54	*	#
Fixations	14	1.57	*	
Fixations duration	14	1.56	*	
Time between fixations	14	1.6	*	
Distance between fixation points	15	1.57	*	
Fixation points change speed	13	1.56	*	#
Clicks	13	1.58	*	
Time between clicks	14	1.58	*	
Keystrokes	11	1.59	*	
Time between keystrokes	14	1.57	*	
Visible keystrokes	10	1.6	*	
Time between visible keystrokes	10	1.6	*	
Backspaces	14	1.59	*	
Time between backspaces	14	1.59	*	
Actions	13	1.59	*	
Time between actions	15	1.59	*	
Attempts	11	1.58	*	
Borys et al. [21] (types)	40	1.48	*	#
Herbig et al. [73] (types)	45	1.49	*	#
Eye activity	30	1.51	*	#
Interactions	41	1.5	*	#
Performance	12	1.51	*	#
Behavioral	39	1.55	*	#
Physiological	45	1.49	*	#
All	76	1.44	*	#

Table I.3: Regressions' RMSE while measuring students' effort with effort features

Appendix J

Full results of the effort measurement using the engagement features

J.1 Classification results

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Dummy	10	0.13	0.17	0.42
Empty	10	0.26 *	0.35 *	0.6 *
Effort	15	0.34 *	0.42 *	0.67 *
Score	15	0.33 *	0.36 *	0.61 *
Time	16	0.3 *	0.36 *	0.62 *
Heart rate	21	0.36 * #	0.41 * #	0.68 * #
Pupil diameter	34	0.34 *	0.37 *	0.63 *
Fixations	26	0.35 * #	0.41 * #	0.66 * #
Fixations duration	14	0.34 * #	0.37 *	0.61 *
Time between fixations	15	0.32 *	0.36 *	0.62 *
Distance between fixation points	15	0.34 * #	0.41 * #	0.67 * #
Fixation points change speed	13	0.32 *	0.36 *	0.62 *
Clicks	19	0.33 * #	0.37 *	0.62 *
Time between clicks	14	0.3 *	0.37 *	0.63 *
Keystrokes	21	0.34 * #	0.4 * #	0.67 * #
Time between keystrokes	14	0.33 * #	0.38 *	0.62 *
Visible keystrokes	11	0.28 *	0.36 *	0.6 *
Time between visible keystrokes	10	0.26 *	0.35 *	0.61 *
Backspaces	25	0.35 * #	0.41 * #	0.67 * #
Time between backspaces	15	0.33 *	0.37 *	0.61 *
Actions	26	0.33 * #	0.39 *	0.65 *
Time between actions	15	0.32 * #	0.37 *	0.61 *
Attempts	10	0.26 *	0.34 *	0.59 *
Previous exercises	29	0.28 *	0.35 *	0.59 *
Borys et al. [21] (types)	67	0.36 * #	0.41 * #	0.67 * #
Herbig et al. [73] (types)	78	0.37 * #	0.42 * #	0.68 * #
Eye activity	43	0.36 *	0.42 *	0.68 * #

Table J.1: Classification’s accuracy while measuring students’ effort with engagement features (continues in the next page)

(continuation)

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Interactions	110	0.37 *	0.42 * #	0.68 *
Performance	21	0.36 * #	0.41 * #	0.68 *
Behavioral	99	0.36 * #	0.42 * #	0.67 * #
Physiological	78	0.37 *	0.42 * #	0.68 * #
All	183	0.36 *	0.41 *	0.67 *
All but effort	178	0.37 *	0.42 *	0.68 *

Table J.1: Classification’s accuracy while measuring students’ effort with engagement features

J.2 Ordinal regression results

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Dummy	10	0.13	0.17	0.42
Empty	10	0.15 *	0.28 *	0.58 *
Effort	15	0.25 *	0.35 *	0.65 *
Score	15	0.15 #	0.26 * #	0.55 * #
Time	16	0.16 *	0.28 *	0.57 *
Heart rate	21	0.16 *	0.29 *	0.57 *
Pupil diameter	34	0.18 *	0.29 *	0.58 *
Fixations	26	0.17 *	0.3 *	0.58 *
Fixations duration	14	0.17 * #	0.3 *	0.59 *
Time between fixations	15	0.16 * #	0.29 *	0.58 *
Distance between fixation points	15	0.17 * #	0.29 *	0.57 *
Fixation points change speed	13	0.16 *	0.3 *	0.59 *
Clicks	19	0.16 *	0.29 *	0.58 *
Time between clicks	14	0.15	0.28 *	0.56 * #
Keystrokes	21	0.17 *	0.29 *	0.58 *
Time between keystrokes	14	0.16 *	0.29 *	0.58 *
Visible keystrokes	11	0.15 * #	0.28 *	0.58 *
Time between visible keystrokes	10	0.15 *	0.28 *	0.58 *
Backspaces	25	0.16 * #	0.29 *	0.59 *
Time between backspaces	15	0.15 *	0.28 *	0.58 *
Actions	26	0.17 * #	0.29 * #	0.6 *
Time between actions	15	0.16	0.28 *	0.58 *
Attempts	10	0.15 *	0.28 *	0.58 *
Previous exercises	29	0.16	0.29 *	0.57 *
Borys et al. [21] (types)	67	0.18 *	0.29 *	0.58 *
Herbig et al. [73] (types)	78	0.18 *	0.29 *	0.58 *
Eye activity	43	0.18 *	0.28 *	0.57 *
Interactions	110	0.21 *	0.28 * #	0.58 * #
Performance	21	0.17 *	0.29 *	0.57 * #
Behavioral	99	0.2 *	0.28 *	0.57 *

Table J.2: Ordinal regression’s accuracy while measuring students’ effort with engagement features (continues in the next page)

(continuation)

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Physiological	78	0.18 *	0.29 *	0.58 *
All	183	0.26 *	0.32 *	0.62 *
All but effort	178	0.23 *	0.3 *	0.61 *

Table J.2: Ordinal regression’s accuracy while measuring students’ effort with engagement features

J.3 Regression results

Subset	#Features	RMSE
Dummy	10	1.66
Empty	10	1.6 *
Effort	15	1.5 *
Score	15	1.52 *
Time	16	1.54 *
Heart rate	21	1.51 * #
Pupil diameter	34	1.45 * #
Fixations	26	1.45 * #
Fixations duration	14	1.51 *
Time between fixations	15	1.54 *
Distance between fixation points	15	1.52 *
Fixation points change speed	13	1.52 *
Clicks	19	1.52 *
Time between clicks	14	1.58 *
Keystrokes	21	1.54 * #
Time between keystrokes	14	1.58 *
Visible keystrokes	11	1.6
Time between visible keystrokes	10	1.6 *
Backspaces	25	1.52 * #
Time between backspaces	15	1.6
Actions	26	1.52 * #
Time between actions	15	1.57 *
Attempts	10	1.6 *
Previous exercises	29	1.61
Borys et al. [21] (types)	67	1.35 * #
Herbig et al. [73] (types)	78	1.35 * #
Eye activity	43	1.36 * #
Interactions	110	1.37 * #
Performance	21	1.46 *
Behavioral	99	1.41 * #
Physiological	78	1.35 * #
All	183	1.33 * #
All but effort	178	1.34 *

Table J.3: Regressions’ RMSE while measuring students’ effort with engagement features

Appendix K

Full results of the effort prediction using the engagement features

K.1 Classification results

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Dummy	10	0.13	0.17	0.43
Empty	10	0.26 *	0.28 *	0.53 *
Effort	15	0.34 *	0.41 *	0.63 *
Score	16	0.32 *	0.4 *	0.65 *
Time	16	0.3 *	0.36 *	0.58 *
Heart rate	21	0.33 *	0.39 *	0.63 *
Pupil diameter	34	0.33 *	0.36 *	0.61 *
Fixations	26	0.36 *	0.41 *	0.63 *
Fixations duration	14	0.34 *	0.37 *	0.6 *
Time between fixations	15	0.33 *	0.38 *	0.6 *
Distance between fixation points	15	0.34 *	0.39 *	0.63 *
Fixation points change speed	13	0.33 *	0.37 *	0.61 *
Clicks	19	0.32 *	0.39 *	0.61 *
Time between clicks	14	0.3 *	0.33 *	0.58 *
Keystrokes	20	0.32 *	0.39 *	0.64 *
Time between keystrokes	14	0.31 *	0.37 *	0.61 *
Visible keystrokes	11	0.28 *	0.35 *	0.58 *
Time between visible keystrokes	10	0.26 *	0.28 *	0.53 *
Backspaces	25	0.36 *	0.41 *	0.65 *
Time between backspaces	15	0.32 *	0.37 *	0.59 *
Actions	25	0.32 *	0.4 *	0.64 *
Time between actions	15	0.31 *	0.37 *	0.58 *
Attempts	10	0.26 *	0.29 *	0.54 *
Previous exercises	29	0.26 *	0.28 *	0.55 *
Borys et al. [21] (types)	67	0.35 *	0.41 *	0.65 *
Herbig et al. [73] (types)	78	0.36 *	0.41 *	0.64 *
Eye activity	43	0.37 *	0.42 *	0.66 *

Table K.1: Classification’s accuracy while predicting students’ effort with engagement features (continues in the next page)

(continuation)

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Interactions	109	0.36 *	0.42 *	0.68 *
Performance	22	0.34 *	0.41 *	0.66 *
Behavioral	97	0.34 *	0.4 *	0.64 *
Physiological	78	0.35 *	0.41 *	0.63 *
All	182	0.36 *	0.42 *	0.67 *
All but effort	177	0.37 *	0.43 *	0.68 *

Table K.1: Classification’s accuracy while predicting students’ effort with engagement features

K.2 Ordinal regression results

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Dummy	10	0.13	0.17	0.43
Empty	10	0.15	0.28 *	0.55 * #
Effort	15	0.25 *	0.35 *	0.62 * *
Score	16	0.17 * #	0.28 *	0.57 * *
Time	16	0.16 *	0.29 *	0.55 * *
Heart rate	21	0.16 *	0.28 *	0.56 * *
Pupil diameter	34	0.17 *	0.29 *	0.56 * *
Fixations	26	0.16 *	0.29 *	0.56 * * #
Fixations duration	14	0.17 *	0.3 *	0.56 * * #
Time between fixations	15	0.16 *	0.28 *	0.55 * *
Distance between fixation points	15	0.16 *	0.29 *	0.55 * *
Fixation points change speed	13	0.17 *	0.29 *	0.56 * *
Clicks	19	0.16 *	0.29 *	0.55 * *
Time between clicks	14	0.16 *	0.29 *	0.55 * *
Keystrokes	20	0.16 *	0.28 *	0.56 * *
Time between keystrokes	14	0.16 *	0.29 *	0.56 * *
Visible keystrokes	11	0.15	0.28 *	0.55 * * #
Time between visible keystrokes	10	0.15	0.28 *	0.55 * * #
Backspaces	25	0.17 *	0.29 *	0.57 * *
Time between backspaces	15	0.16 *	0.29 *	0.56 * *
Actions	25	0.17 *	0.29 *	0.56 * * #
Time between actions	15	0.15	0.28 *	0.54 * * #
Attempts	10	0.15	0.28 *	0.55 * * #
Previous exercises	29	0.16 *	0.29 *	0.56 * *
Borys et al. [21] (types)	67	0.18 *	0.28 *	0.55 * *
Herbig et al. [73] (types)	78	0.17 *	0.27 *	0.54 * * #
Eye activity	43	0.19 *	0.29 *	0.56 * *
Interactions	109	0.24 *	0.32 *	0.59 * *
Performance	22	0.18 *	0.29 *	0.58 * *
Behavioral	97	0.21 *	0.29 *	0.57 * *

Table K.2: Ordinal regression’s accuracy while predicting students’ effort with engagement features (continues in the next page)

(continuation)

Subset	#Features	1-7 Recall	1-7 Accuracy	1-3 Accuracy
Physiological	78	0.17 *	0.27 *	0.54 * #
All	182	0.28 *	0.34 *	0.61 *
All but effort	177	0.26 *	0.32 *	0.58 *

Table K.2: Ordinal regression’s accuracy while predicting students’ effort with engagement features

K.3 Regression results

Subset	#Features	RMSE
Dummy	10	1.68 #
Empty	10	1.65
Effort	15	1.51 *
Score	16	1.55 *
Time	16	1.56 *
Heart rate	21	1.52 *
Pupil diameter	34	1.49 *
Fixations	26	1.47 *
Fixations duration	14	1.55 *
Time between fixations	15	1.55 *
Distance between fixation points	15	1.55 *
Fixation points change speed	13	1.54 *
Clicks	19	1.55 *
Time between clicks	14	1.61 *
Keystrokes	20	1.57 *
Time between keystrokes	14	1.62
Visible keystrokes	11	1.64
Time between visible keystrokes	10	1.65
Backspaces	25	1.53 *
Time between backspaces	15	1.62
Actions	25	1.56 *
Time between actions	15	1.61 *
Attempts	10	1.65
Previous exercises	29	1.64
Borys et al. [21] (types)	67	1.37 *
Herbig et al. [73] (types)	78	1.34 *
Eye activity	43	1.38 *
Interactions	109	1.41 *
Performance	22	1.47 *
Behavioral	97	1.46 *
Physiological	78	1.34 *
All	182	1.33 *
All but effort	177	1.34 *

Table K.3: Regressions’ RMSE while predicting students’ effort with engagement features

Appendix L

Full results of the recommendation models performance

In this Annex we present the full results obtained during the evaluation of our recommendation models. In the following tables, each line represents a recommendation model presented in Chapter 6. In turn, each column represents the respective foot-in-the-door function tested:

- Column N represents the *NoFITD* function. In other words, the recommendation models that do not formalize the foot-in-the-door.
- Column Z represents the *ZpdFITD* function.

In the following tables we can also see five different symbols, each one of them marking the presence of a statistically significant difference when carrying out the following comparisons:

1. Asterisk (*): It marks the comparison between the model that formalizes the foot-in-the-door with its respective model that does not (i.e., original model).
2. Diamond suit (\diamond): It marks the comparison between the recommendation model and the original sequence (column N , row *Original sequence*),
3. Club suit (\clubsuit): It marks the comparison between the models that use the *ZpdFITD* function (column Z) and the same function applied over the original sequence (column Z , row *Original sequence*).
4. Heart suit (\heartsuit): It marks the comparison between the recommendation model and random recommendations (column N , row *Random*),

The color of all of the aforementioned symbols is **blue** if the baseline (the model represented by the column and row) has a larger value, and in **red** if it is smaller.

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	0.5	◇		0.79	*	◇		♡
– Difficulty	0.53	◇		0.76	*	◇		♡
– Grade	0.5	◇		0.74	*	◇		♡
– Actions	0.53	◇		0.71	*	◇		♡
– Similarity	0.5	◇		0.79	*	◇		♡
Top N								
– Compliance	0.86	◇	♡	0.88		◇		♡
– Difficulty	0.93	◇	♡	0.96		◇		♡
– Grade	0.58			0.47				
– Actions	0.73	◇	♡	0.79		◇		♡
– Similarity	0.67	◇	♡	0.77		◇		♡
Original sequence	0.45		♡	1				
Random	0.52							

Table L.1: Proportion of accepted exercises on which students presented an effort increase

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	0.83			0.92	*			♡
– Difficulty	0.8			0.93	*			♡
– Grade	0.83			0.92	*			♡
– Actions	0.79		♡	0.89	*			♡
– Similarity	0.83			0.92	*			♡
Top N								
– Compliance	0.58		♡	0.85	*			
– Difficulty	0.52		♡	0.61				♡
– Grade	0.57		♡	0.6				
– Actions	0.56		♡	0.6				♡
– Similarity	0.69		♡	0.79	*			
Original sequence	1			1				
Random	0.83							

Table L.2: Proportion of recommendations accepted

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	122	◇		37	*	◇	♣	♡
– Difficulty	104.2	◇		41.6	*	◇		♡
– Grade	119.6	◇		41.6	*	◇		♡
– Actions	97	◇	♡	49.8	*	◇		♡
– Similarity	119	◇		37	*	◇	♣	♡
Top N								
– Compliance	49.6	◇	♡	31.8	*	◇	♣	♡
– Difficulty	44.6	◇	♡	35.2		◇	♣	♡
– Grade	53	◇	♡	10.8	*	◇	♣	♡
– Actions	46.8	◇	♡	36.6	*	◇	♣	♡
– Similarity	66.6	◇	♡	45.6	*	◇		♡
Original sequence	165.2		♡	51.6	*	◇		♡
Random	118.6							

Table L.3: Total number of recommendations

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	5.98	◇		1.82	*	◇	♣	♡
– Difficulty	5.11	◇		2.04	*	◇		♡
– Grade	5.86	◇		2.04	*	◇		♡
– Actions	4.75	◇	♡	2.45	*	◇		♡
– Similarity	5.83	◇		1.82	*	◇	♣	♡
Top N								
– Compliance	2.43	◇	♡	1.56	*	◇	♣	♡
– Difficulty	2.18	◇	♡	1.72		◇	♣	♡
– Grade	2.59	◇	♡	1.05	*	◇	♣	♡
– Actions	2.3	◇	♡	1.79	*	◇	♣	♡
– Similarity	3.26	◇	♡	2.23	*	◇		♡
Original sequence	8.09		♡	2.52	*	◇		♡
Random	5.82							

Table L.4: Number of recommendations per student

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	0.45			0.66	*	◇	♣	♡
– Difficulty	0.48	◇		0.67	*	◇	♣	♡
– Grade	0.46			0.66	*	◇	♣	♡
– Actions	0.5	◇		0.65		◇	♣	♡
– Similarity	0.46			0.67	*	◇	♣	♡
Top N								
– Compliance	0.7	◇	♡	0.74		◇	♣	♡
– Difficulty	0.93	◇	♡	0.95		◇	♣	♡
– Grade	0.29	◇	♡	0.14		◇	♣	♡
– Actions	0.55	◇		0.58		◇	♣	
– Similarity	0.61	◇	♡	0.72		◇	♣	♡
Original sequence	0.41		♡	0.45	*			
Random	0.47							

Table L.5: Proportion of accepted exercises on which students presented a grade increase

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	49.43	◇		62.22	*	◇	♣	♡
– Difficulty	50.49	◇		64.26	*	◇	♣	♡
– Grade	49.38	◇		64.81	*	◇	♣	♡
– Actions	50.22	◇		58.93	*	◇	♣	♡
– Similarity	49.5	◇		62.2	*	◇	♣	♡
Top N								
– Compliance	49.98			50.46				
– Difficulty	72.31	◇	♡	72.47		◇	♣	♡
– Grade	42.03		♡	46.51				
– Actions	30.83	◇	♡	30.09		◇	♣	♡
– Similarity	63.65	◇	♡	65.56		◇	♣	♡
Original sequence	41.51		♡	45.8	*			
Random	50.4							

Table L.6: Average grade on accepted exercises

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	0.47	◇		0.64		◇	♣	
– Difficulty	0.53	◇		0.73	*	◇	♣	♡
– Grade	0.48	◇		0.69	*	◇	♣	♡
– Actions	0.53	◇		0.66		◇	♣	♡
– Similarity	0.47	◇		0.64		◇	♣	
Top N								
– Compliance	0.46			0.46				
– Difficulty	0.83	◇	♡	0.83		◇	♣	♡
– Grade	0.32		♡	0.29				
– Actions	0.3		♡	0.3				♡
– Similarity	0.7	◇	♡	0.73		◇	♣	♡
Original sequence	0.38		♡	0.42	*			
Random	0.47							

Table L.7: Proportion of accepted exercises on which the students' had good grades

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	0.25	◇		0.62	*	◇		♡
– Difficulty	0.28	◇		0.56	*	◇		♡
– Grade	0.26	◇		0.57	*	◇		♡
– Actions	0.3	◇		0.52	*	◇		♡
– Similarity	0.26	◇		0.62	*	◇		♡
Top N								
– Compliance	0.67	◇	♡	0.72		◇	♣	♡
– Difficulty	0.89	◇	♡	0.93		◇	♣	♡
– Grade	0.12	◇	♡	0.04		◇	♣	♡
– Actions	0.4	◇	♡	0.46		◇		♡
– Similarity	0.46	◇	♡	0.61		◇		♡
Original sequence	0.18		♡	0.45		◇		♡
Random	0.25							

Table L.8: Proportion of accepted exercises on which students presented an increase on effort and grade

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	0.31	◇		0.42		◇	♣	
– Difficulty	0.34	◇		0.47	*	◇	♣	♡
– Grade	0.3	◇		0.43	*	◇	♣	♡
– Actions	0.34	◇		0.43		◇	♣	♡
– Similarity	0.31	◇		0.41		◇	♣	
Top N								
– Compliance	0.32			0.33				
– Difficulty	0.56	◇	♡	0.55		◇	♣	♡
– Grade	0.25		♡	0.29				
– Actions	0.18		♡	0.19				♡
– Similarity	0.45	◇	♡	0.46		◇	♣	♡
Original sequence	0.25		♡	0.22	*			♡
Random	0.32							

Table L.9: Proportion of accepted exercises that lead to learning

Model	N	◇	♡	Z	*	◇	♣	♡
Q-learning								
– Compliance	0.93			0.96				
– Difficulty	0.93			0.96				
– Grade	0.93			0.97				
– Actions	0.93			0.95				
– Similarity	0.93			0.96				
Top N								
– Compliance	0.92			0.92				
– Difficulty	0.97			0.97				
– Grade	0.89			0.85				
– Actions	0.91			0.91				
– Similarity	0.97	◇		0.98		◇	♣	
Original sequence	0.9			0.87	*			
Random	0.92							

Table L.10: Proportion of accepted exercises that lead to engagement

Appendix M

Full comparison results of the recommendation models that formalize the foot-in-the-door technique

In this Annex we present the full results obtained during the comparison between all of the proposed models that use the *ZpdFITD* function. In the following tables, each line represents a recommendation model presented in Chapter 6. In turn, each column represents one of the models that best apply the foot-in-the-door technique, according to the results presented in Chapter 6 and Annex L.

In the following tables, a statistically significant difference is marked by an asterisk in **blue** if the model represented by the given row and columns has a larger value, and in **red** if it is smaller.

	Q-learning		Top N	
	Compliance	Similarity	Compliance	Difficulty
Q-learning				
– Compliance				*
– Difficulty				*
– Grade				*
– Actions				*
– Similarity				*
Top N				
– Compliance				*
– Difficulty	*	*	*	
– Grade	*	*		
– Actions	*	*	*	
– Similarity	*	*		*

Table M.1: Comparison between the recommendation models that formalize the foot-in-the-door regarding the metric proportion of recommendations accepted

	Q-learning		Top N	
	Compliance	Similarity	Compliance	Difficulty
Q-learning				
– Compliance				
– Difficulty				
– Grade				
– Actions				
– Similarity				
Top N				
– Compliance				
– Difficulty				
– Grade	*	*	*	*
– Actions				
– Similarity				

Table M.2: Comparison between the recommendation models that formalize the foot-in-the-door regarding the metric total number of recommendations

	Q-learning		Top N	
	<i>Compliance</i>	<i>Similarity</i>	<i>Compliance</i>	<i>Difficulty</i>
Q-learning				
– Compliance				
– Difficulty				
– Grade				
– Actions				
– Similarity				
Top N				
– Compliance				
– Difficulty				
– Grade	*	*		*
– Actions				
– Similarity				

Table M.3: Comparison between the recommendation models that formalize the foot-in-the-door regarding the metric number of recommendations per student

	Q-learning		Top N	
	<i>Compliance</i>	<i>Similarity</i>	<i>Compliance</i>	<i>Difficulty</i>
Q-learning				
– Compliance				*
– Difficulty				*
– Grade				*
– Actions				*
– Similarity				*
Top N				
– Compliance				*
– Difficulty	*	*	*	
– Grade	*	*	*	*
– Actions				*
– Similarity				*

Table M.4: Comparison between the recommendation models that formalize the foot-in-the-door regarding the metric proportion of accepted exercises on which students presented a grade increase

	Q-learning		Top N	
	Compliance	Similarity	Compliance	Difficulty
Q-learning				
– Compliance				*
– Difficulty				*
– Grade			*	*
– Actions				*
– Similarity				*
Top N				
– Compliance				*
– Difficulty	*	*	*	
– Grade				*
– Actions	*	*	*	*
– Similarity				*

Table M.5: Comparison between the recommendation models that formalize the foot-in-the-door regarding the metric average grade on accepted exercises

	Q-learning		Top N	
	Compliance	Similarity	Compliance	Difficulty
Q-learning				
– Compliance				
– Difficulty				
– Grade				
– Actions				*
– Similarity				
Top N				
– Compliance				
– Difficulty				
– Grade				*
– Actions	*	*		*
– Similarity				

Table M.6: Comparison between the recommendation models that formalize the foot-in-the-door regarding the metric proportion of accepted exercises on which the students' had good grades

	Q-learning		Top N	
	<i>Compliance</i>	<i>Similarity</i>	<i>Compliance</i>	<i>Difficulty</i>
Q-learning				
– Compliance				*
– Difficulty				*
– Grade				*
– Actions				*
– Similarity				*
Top N				
– Compliance				*
– Difficulty	*	*	*	
– Grade	*	*	*	*
– Actions			*	*
– Similarity				*

Table M.7: Comparison between the recommendation models that formalize the foot-in-the-door regarding the metric proportion of accepted exercises on which students presented an increase on effort and grade

	Q-learning		Top N	
	<i>Compliance</i>	<i>Similarity</i>	<i>Compliance</i>	<i>Difficulty</i>
Q-learning				
– Compliance				
– Difficulty				
– Grade				
– Actions				
– Similarity				
Top N				
– Compliance				
– Difficulty				
– Grade				
– Actions	*	*		*
– Similarity				

Table M.8: Comparison between the recommendation models that formalize the foot-in-the-door regarding the metric proportion of accepted exercises that lead to learning

	Q-learning		Top N	
	<i>Compliance</i>	<i>Similarity</i>	<i>Compliance</i>	<i>Difficulty</i>
Q-learning				
– Compliance				
– Difficulty				
– Grade				
– Actions				
– Similarity				
Top N				
– Compliance				
– Difficulty				
– Grade				
– Actions				
– Similarity				

Table M.9: Comparison between the recommendation models that formalize the foot-in-the-door regarding the metric proportion of accepted exercises that lead to engagement

Abstract

Data exploitation is a growing phenomenon that is present in different scenarios, including the educational scenario, where it holds the promise of advancing our understanding and improving the learning process. From this promise emerged the learning analysis research field that, ideally, takes advantage of technology and educational theories to explore the educational data. On the technological side, we are interested in recommendation systems because they can help students, teachers and other stakeholders to find the best learning resources and thus achieve their learning goals and develop competencies in less time. On the theoretical side, we are interested in the social influence technique foot-in-the-door, which consists in making consecutive requests with an increasing cost. This technique seems particularly relevant to the educational context because it can not only be formalized into a recommendation system, but it is also compatible with the zone of proximal development that states that the challenge presented by the learning resources need to increase gradually in order to keep students motivated. However, we do not know to what extent explicitly applying this technique via recommendations can influence students. Therefore, in this thesis, we investigate such influences assuming that students' effort is a good indicator of the cost of the requests, since not only every learning activity requires a certain level of effort and, but it is often cited as a key factor for students' success. For this, we modeled the measurement and prediction of the students' effort through machine learning models using data that can be used in real life and exploited it in order to explicitly apply the foot-in-the-door technique in a recommendation system. Our results show that, compared to recommendation models that do not formalize this technique, the proposed recommendation models have a positive influence on the students' effort, compliance, performance and engagement. This suggests that this approach has the potential to improve the learning process as students will present the aforementioned behaviors.

Keywords: Learning analytics, recommendation systems, cognitive load, students' effort, students' engagement, multimodal data

Résumé

L'exploitation des données est un phénomène croissant qui est présent dans différents scénarios, y compris le scénario éducatif, où il tient la promesse de faire progresser notre compréhension et d'améliorer le processus d'apprentissage. De cette promesse a émergé le domaine de recherche sur l'analyse de l'apprentissage qui tire idéalement parti de la technologie et des théories éducatives pour exploiter les données éducatives. Sur le plan technologique, nous nous intéressons aux systèmes de recommandation car ils peuvent aider les étudiants, les enseignants et les autres parties prenantes à trouver les meilleures ressources d'apprentissage, et ainsi atteindre leurs objectifs d'apprentissage et développer des compétences en moins de temps. Sur le plan théorique, nous nous intéressons à la technique d'influence sociale pied-dans-la-porte, qui consiste en faire des demandes consécutives avec un coût croissant. Cette technique semble particulièrement pertinente dans le contexte éducatif car elle peut non seulement être formalisée dans un système de recommandation, mais elle est également compatible avec la zone de développement proximal

qui stipule que le défi présenté par les ressources d'apprentissage doivent augmenter progressivement afin de garder les étudiants motivés. Cependant, nous ne savons pas comment l'application de cette technique via des recommandations peuvent influencer les étudiants. Par conséquent, dans cette thèse, nous étudions ces influences en supposant que l'effort des étudiants est un bon indicateur du coût des demandes, car non seulement chaque activité d'apprentissage nécessite un certain niveau d'effort, mais il est souvent cité comme un facteur clé pour la réussite des étudiants. Pour cela, nous avons modélisé la mesure et la prédiction de l'effort des étudiants grâce à des modèles d'apprentissage automatique utilisant des données pouvant être utilisées dans la vie réelle, et les avons exploitées afin d'appliquer explicitement la technique du pied-dans-la-porte dans un système de recommandation. Nos résultats montrent que, par rapport aux modèles de recommandation qui ne formalisent pas cette technique, les modèles de recommandation proposés ont une influence positive sur l'effort, la conformité, la performance et l'engagement des étudiants. Cela suggère que cette approche a le potentiel d'améliorer le processus d'apprentissage, car les élèves présenteront les comportements susmentionnés.

Mots-clés: Analyse de l'apprentissage, systèmes de recommandation, charge cognitive, effort des étudiants, engagement des étudiants, données multimodales