

University of Macedonia
Interdepartmental Programme
of Postgraduate Studies in Information Systems

**Supporting learners' Capacity for Autonomous Decisions
using Learning Analytics**

Zacharoula K. Papamitsiou

A Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

Supervising Professor: Anastasios A. Economides
Preliminary Examiners: Maria Satratzemi, Thrasyvoulos Tsiatsos

Thessaloniki, 2018

*In memory of
George E. and George N.*

ABSTRACT

Freedom of choice. As old as humanity. Freedom of choice implies accepting responsibility (either consciously or not). Freedom of choice implies making decisions that facilitate the individual's self-set goals. We are free to choose, but *do we know how to make the right choices?*

Contemporary online learning environments provide more opportunities to the learners to freely choose *what, where, and how* to learn, compared to the traditional classrooms or the blended learning environments. These self-directed learning decisions, which imply that the learners have absolute control and accept responsibility of their own learning, are synopsised under the term "autonomous learning". And when the learner is autonomous, the learning is more possible to be efficient: the learner sets her goal, monitors her progress, critically reflects on her learning, becomes aware of her own learning needs and makes self-enforced choices independently, guided by her own goals, because her learning is important for her. The more the learning turns online, the higher the need for learners to gain and sustain autonomous learning competences.

However, learners do not intuitively know how to develop their capacity to achieve autonomy; they need to be trained in building the competences for efficient autonomous choices. It has been argued that practising self-regulated learning strategies could progressively lead to conquering autonomy. Self-regulated learning is a process whereby learners employ strategies (e.g., set goals for their learning, regulate their efforts, seek-help to achieve goals) to regulate their cognition, motivation, and behavior, in line with their goals and the contextual features in the environment. However, it is a "black box" how the usage of these strategies develops autonomous capacity.

Meanwhile, massive amounts of learning and learner data, capturing the learners' behavior and actions within the online learning environments, have progressively become available. Learning analytics (LA) has risen as a knowledge discovery paradigm that provides valuable insights and facilitates stakeholders to understand the learner, the learning process and its implications. The rapid advancements in learning analytics could shed light to how self-regulated learning strategies are employed by learners whilst building their capacity towards achieving autonomy.

This thesis investigates this perspective, explores its potential through the analysis of the different types of learner and learning data collected within online learning environments, and suggests an analytics-driven model for assessing autonomous learning capacity development.

This research is contextualized in online self-assessment and in online collaborative learning conditions. By emphasizing on the learners' personal choices and control over their learning outcome, self-assessment promotes the development of learners' capacity to self-regulate their behavior and to preserve their autonomy. Moreover, the added value of collaborative learning on individual's development is beyond question.

This thesis contributes to understanding how the individuals develop their capacity for autonomous learning by utilizing three learning analytics-enhanced approaches of guiding and supporting learners' autonomous capacity development: (a) as controlled selection of learning tasks guided by the adaptive online self-assessment environment, allowing the learners to practise self-regulated learning strategies, (b) as on-demand task-related analytics visualizations targeting at enhancing learners' autonomous metacognitive help-seeking and data-driven sense-making, and (c) as game-theoretic group-recommendations to motivate learners' autonomous decision-making in collaborative learning conditions. Taken together, these three approaches provide a comprehensive and holistic view of how the online learning environment can assist learners to develop their capacity for autonomous learning. An analytics-driven model that captures and assesses autonomous learning capacity development in online learning self-assessment conditions is suggested, as well.

However, prior to guiding and supporting autonomous learning capacity, it is a pre-requisite to understand the learner, to justify her achievement behavior in online learning conditions and to shape accurate learner models, accordingly. Thus, this thesis also (a) identifies a set of highly informative factors that sufficiently and accurately justify achievement behavior, in a meaningful manner, and (b) develops coherent learner models using the abovementioned factors, and enhanced with temporal dynamics for granting current-awareness.

Overall, the developed analytics throughout the different stages of this research, extend current theoretical understandings on how the autonomous learner makes decisions, from a data-driven perspective of autonomy and self-regulation. Different quantitative measures are extracted from the learners' interactions during online self-assessments and explored with different analysis techniques. This empirical research has significant theoretical and methodological implications for research and practice in the context of autonomous online self-assessment. The discussion of the impact of the findings, as well as the directions for future research, concludes this thesis.

Acknowledgements

*"There comes a moment where our majesty, ourselves,
needs to pass into the Throne Room with its courtiers, the mind and the heart,
to decide upon the charter of his life."*

Tasos Athanasiadis, "The Throne Room"

Here, I would like to thank several "fellow travellers" for their kind help and continuous contributions, without which this dissertation would not have been possible.

Foremost, I would like to express my earnest appreciation to my research supervisor, Professor Anastasios A. Economides, for broadening my research horizons with valuable inspiration, guidance, and support of my study, and mostly with his patience, admonition, sincerity and criticism. His willingness to give his time so generously helped me in all the – light and dark – time of research and writing of this thesis. Dear Professor Anastasios, you welcomed me to the University of Macedonia five years ago and have never stopped supporting me ever since. Thank you for the opportunity you offered me to make my dream come true.

I would also like to thank my preliminary examiners, Professor Maria Satratzemi and Assistant Professor Thrasyvoulos Tsiatsos, whose insightful feedback, comments, and remarks were essential for the improvement of this thesis.

During this journey called doctoral research, I was fortunate enough to work with exceptional colleagues and researchers who honoured me with their fellowship and friendship. Vasileios Terzis, Eirini Karapistoli, Michalis Giannakos, Ilias Pappas, please accept my most sincere "Thank you" for sharing with me not only your knowledge and expertise, but your kindness and understanding, as well, and looking forward to our future endeavors.

Special gratitude goes to the "unseen" contributors of this research, the people who silently reviewed my work in its early stages and provided their fruitful insights. Christos Moridis, Eleni Dalla, Eirini Giannakidou, Foteini Papamitsiou, thank you for standing by me and motivating me in ways that no-one else could.

I would also like to thank my closest friends for their infinite support and understanding all these five years. Areti Ampatzoglou, Nikos Tsoulakis, Ioanna Balodimou, Panagiotis Floros, you are the best friends one could have. Thank you for being there for me, for sharing your cookies with me, for playing basketball with me, for working out with me, for challenging me.

I owe my deepest gratefulness to my parents, Konstantinos and Euaggelia, my life-heroes, for their unconditional love, for making me seek knowledge and truth, for making me who I am. Dad, Mom, I will always owe you...

Finally I would like to express my unconditional thankfulness to the two people who are the "light" in my "darkness", who incessantly believed in my efforts and kept reminding me to breathe, who shared with me the laughs, and who make me smile even when everything seem dark. My beloved sister, Foteini Papamitsiou, and my significant other, Christos Tserkezis, the "voices" who whispered "you can", when I howled "I can't!". Thank you.

Contents

ABSTRACT	3
Acknowledgements	5
List of Figures	8
List of Tables.....	10
Chapter 1 : Introduction	12
Chapter 2 : The State-of-the-Art in Learning Analytics – A decade of empirical research.....	25
Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence	25
Chapter 3 : Learning Analytics for Explanation of Performance in online self-assessment tests.	42
Explaining learning performance using response-time, effort, self-regulation and satisfaction from content: analytical approaches with machine learning and fsQCA	42
Chapter 4 : The “testing analytics” paradigm.....	65
Explaining learners’ performance in computer-based tests using learning analytics ...	65
Chapter 5 : From prediction of performance to learner models – the role of personality.....	86
Exhibiting achievement behavior during Computer-based testing: what temporal trace data and personality traits tell us?	86
Chapter 6 : “Current-awareness” – Enhancing the learner models with temporal dynamics	111
Towards currently-aware learner models	111
Chapter 7 : From motivational profiles to help-seeking strategies.....	138
Which help-seeking strategies do the learners use according to their motivational profiles? A pattern-based approach using learning analytics	138
Chapter 8 : Please Help! What I need to know is.....	157
Fostering learners’ engagement and performance with on-demand metacognitive help: the case of task-related analytics visualizations	157
Chapter 9 : Practice makes perfect – Building capacity for efficient help-seeking	173
The impact of metacognitive help-seeking on engagement and performance: A longitudinal study using learning analytics	173
Chapter 10 : Taking control of the self-assessment.....	186
Exploring autonomous learning capacity from a self-regulated learning perspective using learning analytics	186
Chapter 11 : Towards autonomous decision making.....	198
A learning analytics approach on the impact of learners’ autonomy on performance, response-time and effort	198
Chapter 12 : Modeling autonomous learning capacity development	217
Towards an analytics-driven model for assessing autonomous learning capacity development in online self-assessment conditions	217
Chapter 13 : Supporting groups with recommendations.....	225

Motivating students in collaborative activities with game-theoretic group recommendations	225
Chapter 14 : Conclusions and future directions.....	247
Overall Discussion of contributions, implications and future research directions	247
Bibliography	261
Appendix A : The Learning Analytics and Educational Recommender System	294
Appendix B : Algorithms and Formulas.....	299
Appendix C : Questionnaires & Instruments	301

List of Figures

Figure 1-1. Developing capacity for autonomous learning with self-regulated learning strategies	13
Figure 1-2. A conceptual framework for autonomy development using learning analytics.....	18
Figure 3-1. Venn diagram of the conceptual model.....	53
Figure 4-1. Overall research model and factor relationships with hypotheses in summative fixed and adaptive tests.....	74
Figure 4-2. Path coefficients of the research model, overall variance explained (R^2) for test score and cross-validated predictive relevance (Q^2).....	80
Figure 5-1. TLA for predicting performance during CBT (Papamitsiou & Economides, 2014b).....	89
Figure 5-2. Overall research model and variables relationships.....	90
Figure 5-3. Research model and hypothesis.	92
Figure 5-4. Path coefficients of the research model and overall variance (R^2).....	100
Figure 5-5. Graphical exploratory analysis on classes' characteristics: (a) the five classes according to their time-spent, (b) the five classes according to goal-expectancy, and (c) the five classes according to their level of certainty	102
Figure 6-1. The overall learner modeling method.....	120
Figure 6-2. Outline of real-time model update (testing and training).	123
Figure 6-3. Assignment of students to clusters according to goal-expectancy and self-efficacy.	127
Figure 6-4. Final cluster centers.	127
Figure 6-5. Accuracy of the HAT algorithm for student models' configuration in real-time, for each skill being assessed.....	129
Figure 7-1. The task-related analytics visualizations - information about an easy task.....	144
Figure 7-2. The task-related analytics visualizations - information about a hard task.....	144
Figure 7-3. Clusters and configurations of goal-expectancy and self-efficacy scores.....	147
Figure 7-4. Clusters and configurations of help-seeking strategy.....	147
Figure 7-5. Final motivation-based cluster centers.....	148
Figure 7-6. Final help-seeking strategies cluster centers	148
Figure 7-7. Average requests for task-related analytics visualizations per task, per motivational profile	149
Figure 7-8. Average time-spent on task-related analytics visualizations per task, per motivational profile.....	149
Figure 7-9. Percentages of learners assigned to help-seeking strategy clusters according to motivational profiles.....	150
Figure 8-1. Overall research model and factor relationships with hypotheses.....	162
Figure 8-2. Path coefficients of the research model, overall variance explained (R^2) for test score and cross-validated predictive relevance (Q^2).....	166
Figure 8-3. Average requests for task-related analytics visualizations per task.....	167
Figure 8-4. Average time-spent on task-related analytics visualizations per task.....	167
Figure 9-1. Overview of the longitudinal experimental study – Phases, duration and participants.....	178
Figure 10-1. Overall research model and factor relationships with hypotheses.....	190
Figure 10-2. Path coefficients of the research model, overall variance explained (R^2) for test score and cross-validated predictive relevance (Q^2).....	194
Figure 11-1. Overall research model and factor relationships with hypotheses.....	203
Figure 11-2. Path coefficients of the research model, overall variance explained (R^2) for test score and cross-validated predictive relevance (Q^2).....	210

Figure 12-1. Developing capacity for autonomous learning with self-regulated learning strategies.....	219
Figure 12-2. Overview of the model of autonomous learning capacity development	220
Figure 12-3. A model for assessing autonomous learning capacity development in online self-assessment conditions	224
Figure 13-1. Architecture of the non-cooperative game-theoretic group recommender system for educational resources	230
Figure 13-2. The experimental activity process.....	235
Figure 13-3. (average) Effectiveness of group recommendations with respect to the aggregation strategy and the group inner similarity.....	239
Figure 13-4. Diversity of group recommendations with respect to the aggregation strategy and the group inner similarity.....	239
Figure 13-5. Groups' actual engagement with the items during the collaborative phases of the activity, with respect to the aggregation strategy and the group similarity.....	240
Figure 13-6. Distance of group motivation with respect to individual motivation.	241
Figure 13-7. Means of learning performance per recommendation strategy per group homogeneity type.....	242
Figure A-1. The abstract architecture of the full LAERS version.....	294
Figure A-2. The LAERS student interface	295
Figure A-3. Data flow in the application logic layer	296

List of Tables

Table 1-1. Overview of the thesis research questions by individual chapters.....	21
Table 2-1. Inclusion/exclusion criteria	27
Table 2-2. Classification of case studies according to the learning settings.....	29
Table 2-3. Classification of case studies according to the analysis method	30
Table 2-4. Classification of case studies according to the research objectives	31
Table 2-5. Classification of the results of LA/EDM case studies (algorithmic)	35
Table 2-6. Classification of the results of LA/EDM case studies (pedagogical)	36
Table 2-7. SWOT of LA/EDM research	41
Table 3-1. A summary of our experiment	49
Table 3-2. Features used for training and testing	49
Table 3-3. Covariance matrix for all predictor variables	49
Table 3-4. Performance metrics for cross-validation 10% with three features	49
Table 3-5. Performance metrics for test set size 10% with 4 and 5 features	50
Table 3-6. List of factors considered in the conceptual model	56
Table 3-7. Configurations for high performance	60
Table 3-8. Configurations for medium/low performance	61
Table 4-1. List of factors considered in testing analytics.....	75
Table 4-2. Results for the Latent Constructs of the Measurement Model	77
Table 4-3. Measurement Model – fixed testing.....	77
Table 4-4. Measurement Model – adaptive testing.....	78
Table 4-5. Hypothesis testing results	78
Table 4-6. R ² , Q ² and Direct, Indirect and Total effects	79
Table 5-1. Description of achievers’ classes and their characteristics	93
Table 5-2. Results for the Latent Constructs of the Measurement Model	98
Table 5-3. Discriminant Validity for the Measurement Model	99
Table 5-4. Hypothesis testing results	99
Table 5-5. R ² , Q ² and Direct, Indirect and Total effects	100
Table 5-6. A summary of the classification approach	101
Table 5-7. Performance metrics for cross-validation 10% with seven features	102
Table 5-8. Achievers’ classes and their characteristics (reconsideration).....	107
Table 6-1. Features in learner models.....	118
Table 6-2. Synopsis of the exploratory study	124
Table 6-3. Synopsis of the collected dataset	126
Table 6-4. Results for Validity of the Latent Constructs	126
Table 6-5. k-means ANOVA	127
Table 6-6. Performance Metrics for the HAT Classifier	128
Table 7-1. Measurements used in the study.	143
Table 7-2. Motivational constructs and items from the questionnaire	143
Table 7-3. Results for the Latent Constructs.....	146
Table 7-4. Discriminant validity	146
Table 7-5. k-means ANOVA	147
Table 7-6. ANOVA results for the learning analytics factors on the different performance-based student clusters	148
Table 8-1. Measurements used in the study.....	163
Table 8-2. Results for the Latent Constructs of the Measurement Model	165
Table 8-3. Measurement Model (Discriminant validity) for the treatment group (n=88).....	165
Table 8-4. Hypothesis testing results	165

Table 8-5. Fit indices, Direct, Indirect and Total effects.....	166
Table 8-6. Descriptive statistics for performance for the treatment and control groups	166
Table 8-7. Independent samples t-test results for learning performance.....	166
Table 8-8. ANOVA results for the LA factors on the performance-based clusters.....	167
Table 9-1. Measurements used in the study.....	179
Table 9-2. The final hierarchical linear mixed model for explaining the change in learning performance	181
Table 9-3. Statistical differences between the phases of self-assessment with respect to fixed effects of the response-times variables on learning performance	182
Table 10-1. Measurements used in the study.	191
Table 10-2. Results for the Latent Constructs of the Measurement Model.....	193
Table 10-3. Measurement Model (Discriminant validity)	193
Table 10-4. Hypothesis testing results.....	193
Table 10-5. R ² , Q ² and Total effects	194
Table 11-1. Synopsis of the experimental study.....	204
Table 11-2. Measurements used in the study.....	205
Table 11-3. Results for the Latent Constructs of the Measurement Model.....	208
Table 11-4. Measurement Model (Discriminant validity) for the full-autonomous group.....	208
Table 11-5. Measurement Model (Discriminant validity) for the semi-autonomous group	208
Table 11-6. Hypothesis testing results.....	208
Table 11-7. R ² , Q ² and Direct, Indirect and Total effects.....	209
Table 11-8. Descriptive statistics for learning performance for the treatment and control groups.....	210
Table 11-9. Independent samples t-test results for learning performance	210
Table 11-10. ANOVA results for the LA factors on the different performance-based clusters	211
Table 12-1. The autonomous learning capacity analytics - ALCA.....	223
Table 13-1. The individual students' predicted motivation from the educational resources ...	230
Table 13-2. The payoff (motivation) for each student from all the possible strategies	230
Table 13-3. Questions for measuring motivation.....	232
Table 13-4. Description of Groups in the Second Phase	234
Table 13-5. Metrics for Homogeneous/Heterogeneous Groups	238
Table 13-6. Effect of Recommendation strategy on Performance	241
Table A-1. Features from the raw log files	296
Table A-2. Variables used in this study and short description	297
Table C-1. Chapter 3: Measuring Goal-Expectancy in Study 1.....	301
Table C-2. Chapter 3: Measuring self-regulation and satisfaction from content in Study 2 – Constructs and items from the questionnaires.....	301
Table C-3. Chapter 4: The testing analytics paradigm: Measuring non-cognitive/motivational factors – Constructs and items from the questionnaires	301
Table C-4. Chapter 5: The Big Five Inventory (BFI)	301
Table C-5. Chapters 6, 7: Measuring non-cognitive/motivational factors – Constructs and items from the questionnaires	302
Table C-6. Chapter 7: Multiple Comparisons (Bonferroni test) for the motivational factors	303
Table C-7. Chapter 7: Multiple Comparisons (Bonferroni test) for the help-seeking factors	303
Table C-8. Chapter 8: Measuring perceived usefulness of the visualizations	303
Table C-9. Chapter 10: Measuring self-regulation in online self-assessment – Constructs and items from the questionnaires	304

Chapter 1 : Introduction

*“The more decisions that you are forced to make alone,
the more you are aware of your freedom to choose”*

Thornton Wilder

1.1. General overview

Adaptivity and adaptive learning environments are in the epicentre of the Technology Enhanced Learning (TEL) research community. In the recent 2018 NMC Horizon Report Preview, the emergence of these systems is highlighted. In the same report, it is acknowledged that the focus of these systems is on modifying the instruction anytime and providing the best possible support to the learners “to accurately and logically move [students] through a learning path, empowering active learning” (New Media Consortium, 2018, p. 9). The core idea is to increase their “awareness” regarding the learners’ cognitive and emotional states, as well as regarding their degree of non-cognitive skills and competences acquisition, and to accurately predict what kind of personalized assistance the learners would need, accordingly (Brusilovsky et al., 2016). The adaptive systems adjust their features to meet the learners’ characteristics and offer them a personalized learning experience. This chain of adaptive interactions results in continuously engaging the learners in controlling their own learning, as they move along the self-regulated learning continuum towards autonomy. In a sense, adaptivity is a means of achieving autonomy.

And, the more the learning turns online, the higher the need for learners to develop and sustain autonomous learning competences. Contemporary online learning environments provide more opportunities for autonomous interactions than traditional classrooms or blended learning environments (Broadbent & Poon, 2015; Xu & Jaggars, 2014). Through the “lenses” of autonomy, the learners understand their needs, are aware of their self-directed learning goals, take control of and become responsible for their learning choices, monitor their progress, and critically reflect on their learning (Benson, 2001; Cotterall, 1995; Dickinson, 1995; Holec, 1981; Little, 1991; Littlewood, 1996; White, 1995).

According to Holec’s (1981, p. 3) definition, autonomous learning is “the *capacity* to take charge of one’s own learning”. Littlewood (1999, p. 73) defined autonomous learning as “involving students’ *capacity* to use their learning independently of teachers”. A broader conceptualization of autonomy focuses on learners’ *capacity* that allows them to accept responsibility and take control of their own learning processes (Vanijdee, 2003). Autonomy targets at fostering learners’ responsible self-initiative and allows them to determine the *selection* of what shall be learned, as well as the *critical evaluation* (reflection) of the learning tasks that were selected (Candy, 1991). In Benson’s (2001) conceptualization of learner autonomy, the autonomous and self-directed learners take control over the cognitive, emotional, motivational, and behavioral processes of learning, as well as the independent use of learning material and technology. This approach implies that the autonomous learners exhibit self-regulation strategies, and that they make self-enforced decisions independently. Self-regulated learning

refers to learning that is guided by metacognition (thinking about one's thinking), strategic action (planning, monitoring, and evaluating personal progress against a standard), and motivation to learn (Paul R. Pintrich, 2000; Zimmerman, 2001).

However, autonomy is beyond self-regulation only; autonomous learners are also capable of taking responsibility for their learning choices (Oxford, 2015), as well. Furthermore, Little (1995) argues that autonomy is also beyond independence; the learners should not only feel that they are independent, but to be guided to develop their capacity for autonomy, as well. “Capacity development (capacity building) is the process by which individuals, organizations, institutions and societies develop abilities to perform functions, solve problems and set and achieve objectives” (United Nations Economic and Social Council, 2006, p. 7).

However, learners do not intuitively know how to develop their capacity and achieve autonomy; they need to be trained in building the competences for efficient autonomous choices (McDevitt, 1997; White, 1995). Learners exercise autonomy when they make choices and act on them: freedom of choice is central to the idea of autonomy – it is always the learners who choose *what, where, and how* to learn. For Oxford (2008), the use of self-regulated learning strategies can develop learner’s capacity for autonomous learning. Figure 1-1 illustrates this conceptualization.

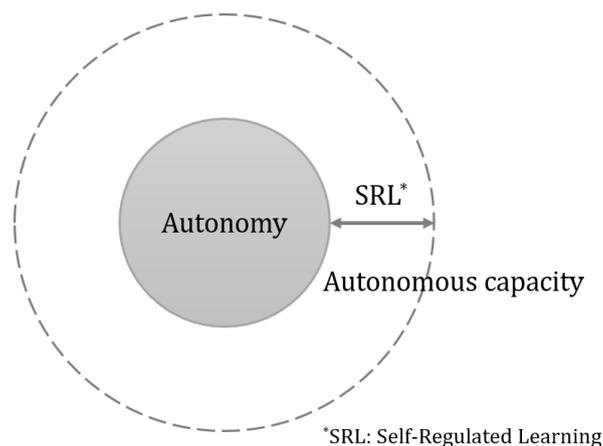


Figure 1-1. Developing capacity for autonomous learning with self-regulated learning strategies

Andrade (2014) prompted that towards achieving autonomy, there is a need to develop Technology Enhanced Learning environments within which the learners shall be given the opportunity for exercising control over their self-regulated learning processes and to be consciously involved in their own learning. However, although such forms of learning may *require* the exercise of autonomy, yet it is a “black box” how their usage *develops* this capacity.

Meanwhile, massive amounts of learning and learner data that reflect the behavior and actions of learners within the digital learning environments have progressively become available. Prior to the availability of massive amounts of educational data, the areas of Intelligent Tutoring Systems (Corbett, Koedinger, & Anderson, 1997), Educational Hypermedia (De Bra, 2002), and Adaptive Hypermedia (Brusilovsky, 2001) used technology mediation to increase the support

learners received while participating in a learning experience. Contemporary technology enhanced learning environments provide increased opportunities for better understanding the learners, their abilities, their needs, their goals, etc.

Learning analytics (LA) has risen as a knowledge discovery paradigm that provides valuable insights and facilitates stakeholders to understand the learner, the learning process and its implications. According to the definition introduced during the 1st International Conference on Learning Analytics and Knowledge, LA is “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and environments in which it occurs” (<https://tekri.athabascau.ca/analytics/>). Like any context-aware system, a learning analytics procedure monitors, tracks and records data related to the context, interprets and maps the real current state of these data, organizes them (e.g., filter, classify, prioritize), uses them (e.g., decide adaptations, recommend, provide feedback, guide the learner) and predicts the future state of these data (Papamitsiou & Economides, 2016).

Since its first mention in the 2012 NMC Horizon Report (Johnson, Adams, & Cummins, 2012), learning analytics has gained an increasing relevance. LA has emerged as a new instance of computers use in education with the purpose of improving teaching and learning, assisting data-driven decision-making, as well as enhancing the delivery of educational experiences (D. Knight, Brozina, & Novoselich, 2016). LA is essentially a specialized paradigm that processes educational data to produce useful information for decision-making. The available data is collected and analysed, and the gained insights are used to understand the behavior of the learners in order to provide additional support to them (Gašević, Dawson, & Siemens, 2015). The LA initiatives can all be connected to improvements that affect learners, yet, their focus is primarily on the steps to collect, analyze, and report data (Papamitsiou & Economides, 2014a). Papamitsiou and Economides (2014a) highlighted the need for the LA community to focus more precisely on the *actions derived from the use of data*.

1.2. Motivation and Purpose of the research

As apparent, it is a challenge to address the issue of how to appropriately support learners in building their capacity towards achieving autonomy. The rapid advancements in recent digital learning environments as well as in learning analytics could assist in facilitating this objective. This thesis investigates this perspective, explores its potential through the analysis of the different types of learner and learning data collected by online learning environments, and suggests an analytics-driven model for assessing autonomous learning capacity development.

Specifically, the large body of this research is contextualized in online self-assessment. Self-assessment leads students to a greater awareness, by training them to self-regulate their motivation and behavior, as well as by fostering reflection on their own progress in knowledge or skills, and finally, to understanding themselves as learners (McMillan & Hearn, 2008; Nicol & Macfarlane-Dick, 2006). It promotes the development of learners' capacity to self-regulate their

own performance and preserves students' autonomy by emphasizing on personal choices and control over their outcome and experience mastery. Moreover, it encompasses learners' evaluation of the quality of their own cognition and their own behavior. The role of feedback in self-assessment procedures is central: feedback will guide and support the learners to make informed decisions. Thus, designing appropriate forms of feedback for achieving this objective is not only necessary, but challenging as well. This thesis attempts to address this issue.

For gaining holistic insights from other online learning contexts – yet, in smaller extent – online collaboration is considered, as well. The added value of collaborative learning on individual's development is beyond question (Kreijns, Kirschner, & Jochems, 2003; Roschelle & Teasley, 1995). In this research, the focus is on exploring the degree of autonomous individuals' choices within the group, on understanding how the other members' autonomous decisions affect the individual's choices, as well as on investigating to what extent the individuals' capacity for autonomous choices can be supported and developed in these conditions.

This thesis explores three learning analytics-enhanced approaches of guiding and supporting learners' autonomous capacity development: (a) as controlled selection of learning tasks guided by the adaptive online learning environment, allowing the learners to practise self-regulated learning strategies in self-assessment conditions, (b) as on-demand task-related analytics visualizations targeting at enhancing learners' autonomous metacognitive help-seeking and data-driven sense-making in online self-assessment conditions, and (c) as game-theoretic group-recommendations to motivate learners' autonomous decision-making in collaborative learning conditions. Taken together, these three approaches provide a comprehensive and holistic view of how the online learning environment can assist learners to develop their capacity for autonomous learning. An analytics-driven model that captures and assesses autonomous learning capacity development in online learning self-assessment conditions is suggested.

However, prior to guiding and supporting autonomous learning capacity, it is a prerequisite to understand the learner, to justify her achievement behavior in online self-assessment conditions and to shape accurate learner models, accordingly. This thesis attempts to address these issues, as well.

1.3. Research objectives and research questions

The work presented in this thesis was conducted with five overarching research objectives in mind.

The *first goal* is in line with the need to shift the discussion about learning performance prediction and interpretation beyond the performance itself, to a deeper understanding of the underlying factors that justify learners' actions in self-assessment. As such, the research objective is to seek and determine the most informative factors that explain learners' behavior in self-assessment activities and affect the learning outcome, using learning analytics, as well as to model causal relationships between them. In doing so, the *first research question* is defined as follows:

Research Question 1:

(a) which learning analytics factors explain sufficiently the multiple aspects of learners' interactions with the assessment tasks? (b) How can we model the cause-effect relationships between these factors towards explaining the variance in the learners' performance?

The *second goal* of the thesis is in line with the demand that contemporary adaptive learning systems have highlighted, i.e., the accurate adaptation to the learners' needs in real-time. Understanding the learner and modeling her, shall next lead to better and timely supporting her. Thus, there is a need to develop coherent learner models, using the previously identified factors, and to enhance these models with a notion of "temporal dynamics", in order to deliver improved and timely personalized learning experience. With this in mind, the *second research question* is:

Research Question 2:

How feasible is it to maintain "currently-aware" learner models in real-time, by refitting their parameters in run-time, and how accurately can these models approximate the learners' next cognitive states from the current ones?

The *third goal* of the thesis is in line with Nelson-Le Gall's (1985) claims that help-seeking can promote autonomy. The focus of the present research is on designing and delivering task-related metacognitive instrumental help (i.e., aided in understanding) to the learners, in order to foster their on-task engagement, data-driven decision making and performance, by considering their motivational profiles. Thus, the *third research question* is twofold and is defined as follows:

Research Question 3.1:

(a) Are there any differences in the usage of metacognitive help with respect to the learners' motivational profiles? If yes, (b) how significant are these differences? (c) Which are the emerging help-seeking strategies? (d) How are these strategies associated with the motivational profiles?

Research Question 3.2:

(a) Can learners make-sense from the task-related analytics visualizations? If yes, how the actual usage of visualizations is related to the learners' perceptions of visualizations' usefulness? (b) Are there any differences in the usage of task-related analytics visualizations with respect to the learners' level of performance? If yes, how significant are these differences? (c) Which is the effect of metacognitive help on learners' performance? (d) Does the exploitation of task-related metacognitive information enhance the learners' performance? If yes, how significant is its effect on learning performance? (e) Do learners' interpretations of the metacognitive help actually help them to deeper engage with the task? How significant is this effect?

Research Question 3.3:

Are there any changes in learners' engagement and performance due to receiving metacognitive help, over time? If yes, how significant are these changes?

The *fourth goal* of the present study is aligned with Andrade's (2014) prompt that practising autonomy shall progressively result in developing capacity for autonomous learning.

For Oxford (2008), the use of self-regulated learning strategies can promote learner autonomy. This thesis targets at taking us beyond current knowledge on *how* this could be achieved, using insights from learning analytics. This study aims at identifying which strategies and which facilitating factors/conditions result in increasing learners' autonomous capacity, and to introduce an analytics-driven model for assessing autonomous learning capacity development in online self-assessment conditions. Therefore, the *fourth research question* is threefold:

Research Question 4.1:

(a) Which is the effect of self-regulated learning strategies on the learners' control of autonomous learning? (b) Which is the impact of autonomous control on learners' performance, response-times and effort? (c) Does the exploitation of autonomous control (measured with utilized analytics) contribute to enhancing the learners' performance? If yes, how significant is its effect on learning performance?

Research Question 4.2:

(a) Are there any differences in the analytics parameters of autonomous control, with respect to the learners' level of performance? If yes, how significant is the effect of each one of these parameters? (b) Are there any differences in the learners' engagement with the self-assessment task, with respect to their level of performance? If yes, how significant is the effect of each one of these parameters?

Research Question 4.3:

To what extent can we exploit learning analytics to assess learners' autonomous capacity development?

The *fifth goal* of this thesis is to explore autonomous capacity development within collaborative, group-learning conditions. Specifically, the objective is to consider the individuals' self-enforced preferences and decisions within the group they belong, and support them with appropriate group-recommendations of educational resources to motivate their participation. Thus, the *final research question* of this PhD is:

Research Question 5:

(a) Can we accurately and efficiently recommend sequences of educational resources to homogeneous and heterogeneous groups of students, with respect to both the individuals' and the group's motivation, and intention to use the resources? (b) What is the impact of a recommendation on individual students' persistence as well as on the groups' learning performance in the collaborative problem-solving activity?

1.4. Theoretical background

This research is contextualized in online self-assessment and in online collaborative learning conditions. The theoretical basis of the research is grounded in three pillars: (1) the learning approach, (2) the learning strategies, and (3) the assessment framework.

The first pillar specifies the **learning approach** used throughout the research to *adaptive learning*, implemented from three different perspectives: (a) as Computerized Adaptive Testing by utilizing a version of the Measurement Decision Theory (Rudner, 2003): the system adapts the self-assessment to the learner' s proficiency, (b) as learner' s self-directed choices of the self-

assessment tasks by integrating the principles of autonomous learning (Benson, 2001): the adaptation is learner-driven, according to the learner’s self-directed decision making, and (c) as group-recommendations of educational resources by exploiting game-theory for resolving “conflict of interest” in collaborative learning conditions (Myerson, 1985; Nash, 1951): the system’s decisions are driven by considering each individual’s preferences within the group.

The second pillar concerns the specification of **learning strategies** in adaptive, online self-assessment contexts, and is particularized in two dimensions: (a) a theoretical model of self-regulated learning (Barnard, Lan, To, Paton, & Lai, 2009), and, (b) an empirical model of self-regulation that exploits learning analytics parameters for modelling the respective strategies.

The third pillar outlines how the LA models and their components fit together to provide a coherent and holistic **framework for the assessment** of students’ autonomous and self-regulated learning in online self-assessment contexts and is based on the evidence-centered design (Mislevy, Almond, & Lukas, 2003), involving the identification of what should be assessed (in terms of knowledge, skills, or other learner attributes), and the configuration of the learning analytics factors.

These theoretical models are being assembled and seen from the learning analytics perspective, share the same contextual underpinning and contribute to a conceptual framework for autonomy development using learning analytics. These three pillars complement each other, as shown in Figure 1-1 (bottom-up).

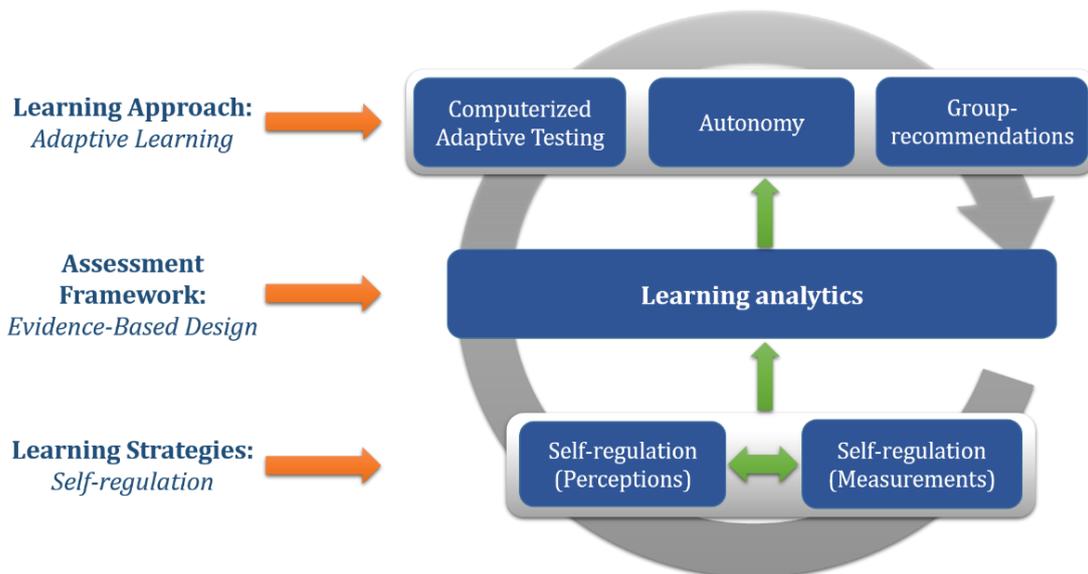


Figure 1-2. A conceptual framework for autonomy development using learning analytics

1.5. Research methodology

1.5.1. Studies design and research participants

A series of experiments – exploratory, experimental and longitudinal – were designed and conducted both in Higher and in Secondary education. In all experiments, evidence-based design guided the design, development and implementation of the research (Mislevy et al., 2003).

Overall, seven (7) exploratory studies (Setia, 2016), four (4) experimental studies (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003), and one (1) longitudinal study (Ployhart & Vandenberg, 2010) provided the data for further analyses. Three studies were conducted at a Greek High School and the rest of them took place at the University of Macedonia. In total, 2888 participants (aged 16-28 years old) attended the experiments. All adult participants signed an informed consent form prior to their participation. The informed consent explained to them the procedure and was giving the right to researchers to use the data collected for research purposes. Students were aware that their interactions had been anonymized prior to being tracked and analysed. For the under-aged High School students, the consent form was signed by their guardians.

1.5.2. Data collection and Measurements

Data were collected with an online self-assessment environment, developed in stages during this research, and configured according to the needs of the studies. In all experiments, measures commonly used in the field of learning analytics and indicative of measuring learners' behavior during the interactions explored (e.g., response-times, frequencies) (Joksimović, Gašević, Loughin, Kovanović, & Hatala, 2015; Kizilcec, Pérez-Sanagustín, & Maldonado, 2017; Kovanović, Gašević, Joksimović, Hatala, & Adesope, 2015) were computed from the logged interactions trace data. In addition, a set of instruments (questionnaires) extensively used in literature, were adopted and particularized according to the needs of the studies. Specifically, measures were adopted from the Technology Acceptance Model (TAM) (Davis, 1989; Venkatesh, Morris, Davis, & Davis, 2003), Computer Based Assessment Acceptance Model (CBAAM) (Terzis & Economides, 2011), Big Five Inventory (John & Srivastava, 1999), Online Self-Regulated Learning Questionnaire (OSLQ; Barnard et al., 2009), and Motivated Strategies for Learning Questionnaire (MSLQ; Paul R Pintrich, Smith, Garcia, & Mckeachie, 1993).

1.5.3. Data analysis methods

The primary means for conducting the research in this thesis are quantitative learning analytics methods and the investigation of empirical data from real-world online self-assessment. For addressing **Research Question 1**, three different data analysis techniques were employed for exploring relationships between the considered factors: (a) supervised machine learning for classification purposes (Alpaydin, 2010), (b) fuzzy-set qualitative comparative analysis (fsQCA) (Ragin, 2008), and (c) Partial Least Squares – Structural Equation Modeling (PLS-SEM) (Chin, 1998). These methods were applied on data extracted from the logged learners' interactions with the self-assessment environment, as well as from the learners' responses to questionnaires. Next, these quantitative metrics were used to shape the learner models in real-time. For addressing **Research Question 2**, initially, supervised classification was applied. Next, for enhancing the models with a notion of temporal dynamics, an adaptive data stream classification approach was employed. Data were adaptively mined as a data-stream, using the HoeffdingAdaptiveTree (HAT) classifier (Bifet & Gavaldà, 2009). For addressing **Research Question 3.1**, unsupervised learning (Tan, Steinbach, & Kumar, 2005) was used to determine the different help-seeking strategies and

the motivational profiles of the learners, variance-based analysis was conducted to investigate differences in the actual usage of help-seeking, and pattern-based analysis was employed to explore the distinctiveness of help-seeking strategies using the motivational profiles. For addressing **Research Question 3.2**, PLS-SEM was used for the construction of a path diagram that contains the structural and measurement model for explaining the effect of metacognitive help on learners' performance, and variance-based analyses were employed for investigating individual differences. The same methods were applied for addressing **Research Question 4**, as well. Finally, for addressing **Research Question 5**, heuristics commonly used in the field of recommender systems (e.g., the Normalized Discounted Cumulative Gain (Järvelin & Kekäläinen, 2002), the Hamming Distance metric (Zhou, Jiang, Su, & Zhang, 2008)) were employed.

1.6. Overview of Contributions

Building upon the inherent potential of self-assessment to support self-improvement, this dissertation demonstrates how training students to choose the learning tasks/recommended materials according to their goal-orientation, and to extract actionable insights from data-analytics visualizations can (a) lead them to efficient, informed, self-enforced decision-making, (b) promote autonomous learning capacity, and (c) guide them to become responsible learners.

The overall contribution of this PhD study involves: a) the identification of the most informative factors that sufficiently and accurately explain performance and achievement behavior in self-assessment contexts, in a meaningful manner (e.g., response-times, effort, satisfaction from content, goal-expectations, time-awareness, autonomous capacity, help-seeking capacity, visualizations translation capacity, personality traits), and the construction of a conceptual model, further evaluated with multiple measurements, using different data-analysis techniques (e.g., PLS, fsQCA, machine learning); b) the development of coherent learner models using the abovementioned factors, and enhanced with temporal dynamics for granting current-awareness (i.e., allowing for accurate temporal predictions in real-time – predicting future states from the current ones – and continuously keeping up-to-date the progress of the learners' multiple skill/knowledge acquisition in a single profile); and c) the support of learners' capacity for autonomous decision-making in three ways: i) with on-demand metacognitive feedback as visualizations of learning task-related analytics, ii) with semi-autonomous control as adaptive self-assessment, and iii) with game-theoretic group recommendations of educational resources, in collaborative learning contexts. Another contribution of this research is the introduction of an analytics-driven model for assessing autonomous learning capacity development in online self-assessment conditions.

Furthermore, during this doctoral research a self-assessment system (LAERS, Learning Analytics and Educational Recommender System) has been developed and evaluated, dedicated to facilitating continuous, autonomous capacity building activities. The LAERS started as a small-scale initiative in 2013 that initially sought to facilitate reflection from self-assessment

procedures targeting at aligning learning goals with learning outputs. Progressively, the system integrated the abovementioned functionalities and was used as the basic tool for data collection.

1.7. Thesis in brief

The dissertation consists of six parts and 14 chapters in total, and is structured across four learning analytics development stages, with different chapters corresponding to one or more stages. Each chapter focuses on one or more research questions– except for the first two – (Table 1-1), and includes at least one peer-reviewed publication, as the core of the chapter.

Table 1-1. Overview of the thesis research questions by individual chapters.

Chapter	Title	Research Question				
		RQ1	RQ2	RQ3	RQ4	RQ5
Chapter 2	The State-of-the-Art in Learning Analytics					
Chapter 3	Learning Analytics for Explanation of Performance in online self-assessment tests	✓				
Chapter 4	The “testing analytics” paradigm	✓				
Chapter 5	From prediction of performance to learner models – the role of personality		✓			
Chapter 6	Enhancing the learner models with temporal dynamics		✓			
Chapter 7	From motivational profiles to help-seeking strategies			✓		
Chapter 8	Please Help! What I need to know is....			✓		
Chapter 9	Practice makes perfect – Building capacity for efficient help-seeking			✓		
Chapter 10	Taking control of the self-assessment					✓
Chapter 11	Towards autonomous decision making					✓
Chapter 12	Towards assessing autonomous learning capacity development					✓
Chapter 13	Supporting groups with recommendations					✓

Specifically, the first part includes **Chapter 2**. This chapter provides the review of the state-of-the-art in the field of learning analytics and educational data mining from 2008 to 2014, and highlights the current trends, opportunities and critical topics in the area.

List of relevant publications

1. Papamitsiou, Z. & Economides, A. A. (2014). Learning Analytics and Educational Data Mining in practice: a systematic literature review of empirical evidence, *Educational Technology and Society*, 17(4), 49-64.
2. Papamitsiou, Z. & Economides, A. A. (2016). An assessment analytics framework for enhancing students’ progress, In S. Caballé Editor and R. Clarisó Editor (Eds.), "ICT-FLAG" *Enhancing ICT education through Formative assessment, Learning Analytics and Gamification*, (117-133), Elsevier. [DOI:10.1016/B978-0-12-803637-2.00007-5](https://doi.org/10.1016/B978-0-12-803637-2.00007-5)
3. Papamitsiou, Z. & Economides, A.A. (2016). Learning Analytics for Smart Learning Environments: A Meta-analysis of Empirical Research Results from 2011 to 2015, In J.M. Spector et al. (eds.), *Learning, Design, & Technology*, (1-23), Springer International Publishing Switzerland, [DOI: 10.1007/978-3-319-17727-4_15-1](https://doi.org/10.1007/978-3-319-17727-4_15-1).

The second part (corresponding to the first stage of learning analytics development) includes **Chapters 3** and **4**, focusing on addressing **Research Question 1**. In particular, **Chapter 3** explores a set of variables regarding their appropriateness to explain learning performance, using supervised classification and fuzzy set qualitative comparative analysis. **Chapter 4** builds on the previous results and considers and explores additional factors, and suggests and evaluates a structural and measurement model for holistically explaining learning performance both in fixed and adaptive assessment tests.

List of relevant publications

1. Papamitsiou, Z., & Economides, A.A. Explaining learners' performance in computer-based tests using learning analytics – The “testing analytics” paradigm (under review)
2. Papamitsiou, Z., & Economides, A.A. (2014). Temporal Learning Analytics for adaptive assessment, *Journal of Learning Analytics*, 1(3), 165-168. [DOI:10.18608/jla.2014.13.13](https://doi.org/10.18608/jla.2014.13.13)
3. Papamitsiou, Z., Economides, A. A., Pappas, I. O., & Giannakos, M. N. (2018). Explaining learning performance using response-time, self-regulation and satisfaction from content: An fsQCA approach. *8th Int. Conf. on Learning Analytics and Knowledge*, (181–190) NY, USA: ACM. [DOI:10.1145/3170358.3170397](https://doi.org/10.1145/3170358.3170397)
4. Papamitsiou, Z., Karapistoli, E. & Economides, A.A. (2016). Applying classification techniques on temporal trace data for shaping student behavior models, *6th Int. Conference on Learning Analytics & Knowledge*, (299-303) NY, USA: ACM. [DOI:10.1145/2883851.2883926](https://doi.org/10.1145/2883851.2883926)
5. Papamitsiou, Z. & Economides, A.A. (2015). A temporal estimation of students' on-task mental effort and its effect on students' performance during computer based testing, *IEEE 18th International Conference on Interactive Collaborative Learning*, (1136 – 1144). [DOI:10.1109/ICL.2015.7318194](https://doi.org/10.1109/ICL.2015.7318194)
6. Papamitsiou, Z., Terzis, V. & Economides, A. A. (2014). Temporal learning analytics for Computer-based testing, *4th Int. Conference on Learning Analytics and Knowledge*, (31-35). [DOI:10.1145/2567574.2567609](https://doi.org/10.1145/2567574.2567609)
7. Papamitsiou, Z. & Economides, A. A. (2014). Students' perception of performance vs. actual performance during computer based testing: a temporal approach, *8th Int. Technology, Education and Development Conference (INTED14)*, (401-411).

The third part (corresponding to the second stage of learning analytics development) includes **Chapters 5** and **6**, aiming at addressing **Research Question 2**. In brief, **Chapter 5** considers the previously identified analytics parameters, explores the role of learners' personality in achievement behavior, and suggests five learner models using machine learning techniques. **Chapter 6** goes a step further by enhancing the previous models with a notion of temporal dynamics, and opens the discussion towards “currently-aware” learner models. These models are generated in real-time and their accuracy is validated in this chapter.

List of relevant publications

1. Papamitsiou, Z., & Economides, A.A. Towards currently-aware learner models, *User Modeling and User-Adapted Interaction (UMUAI)* (under revisions)

2. Papamitsiou, Z., & Economides, A.A. (2017). Exhibiting achievement behavior during Computer-based testing: what temporal trace data and personality traits tell us? *Computers in Human Behavior*, 75, 423–438. [DOI:10.1016/j.chb.2017.05.036](https://doi.org/10.1016/j.chb.2017.05.036)
3. Papamitsiou, Z., & Economides, A.A. (2017). Student modeling in real-time during self-assessment using stream mining techniques, *17th IEEE Int. Conf. on Advanced Learning Technologies (ICALT2017)*, (286-290). [DOI:10.1109/ICALT.2017.90](https://doi.org/10.1109/ICALT.2017.90)
4. Papamitsiou, Z. & Economides, A. A. 2014. The effect of personality traits on students' performance during Computer-Based Testing: a study of the Big Five Inventory with temporal learning analytics, *14th IEEE Int. Conf. on Advanced Learning Technologies (ICALT2014)*, (378-382). [DOI:10.1109/ICALT.2014.113](https://doi.org/10.1109/ICALT.2014.113)

The fourth part (corresponding to the third stage of learning analytics development) includes **Chapters 7, 8, and 9** targeting at answering **Research Question 3**. More precisely, **Chapter 7** extracts the learners' motivational profiles and associates these profiles with strategies of seeking task-related metacognitive help. **Chapter 8** extracts analytics related to the actual usage of task-related metacognitive help-seeking and explores the impact of requesting for this type of assistance on learners' on-task engagement and learning performance. **Chapter 9** demonstrates a longitudinal study, and investigates the changes in learners' on-task engagement and responsible learning before providing them the treatment (i.e., the task-related metacognitive help) and after their exposure to the treatment.

List of relevant publications

1. Papamitsiou, Z. & Economides, A.A. Which help-seeking strategies do the learners use according to their motivational profiles? A pattern-based approach using learning analytics (under review).
2. Papamitsiou, Z. & Economides, A.A. Fostering learners' engagement and performance with on-demand metacognitive help: the case of task-related analytics visualizations, *Journal of Learning and Instruction (JLI)* (under revisions).
3. Papamitsiou, Z. & Economides, A.A. The impact of metacognitive help-seeking on engagement and performance: A longitudinal study using learning analytics (under review).

The fifth part (corresponding to the fourth stage of learning analytics development) is dedicated to the empirical evaluation of learners' autonomous control and includes **Chapters 10, 11 and 12**, addressing **Research Question 4**. In **Chapter 10**, analytics about learners' autonomous choices are extracted and the role of self-regulated learning strategies is explored with respect to the learners' autonomous control. **Chapter 11** uses these analytics and associates them with the learners' actual on-task engagement and performance, by considering the factors identified in Chapter 4. Differences in performance due to autonomy are also explored. Next, **Chapter 12** integrates the previously identified factors that affect autonomy, and suggests an analytics-driven model for assessing autonomous capacity development.

List of relevant publications

1. Papamitsiou, Z. & Economides, A.A. Exploring autonomous learning capacity from a self-regulated learning perspective using learning analytics, *British Journal of Educational Technology* (under revisions).
2. Papamitsiou, Z. & Economides, A.A. Taking control of the self-assessment: a learning analytics approach on the impact of learners' autonomy on performance, response-time and effort, *Computers in Human Behavior* (under revisions).
3. Papamitsiou, Z. & Economides, A. A. An analytics-driven model for assessing autonomous learning capacity development in online self-assessment conditions (under review).

The sixth part includes **Chapter 13**, addressing **Research Question 5**. This Chapter provides empirical evidence on the effectiveness of employing Game-Theory in order to decide upon the recommendation of learning resources to homogeneous, mildly heterogeneous and heterogeneous groups of learners. In this Chapter, the individuals' self-enforced preferences are considered for guiding the recommendation, and the appropriateness of the recommended resources is explored both from the aspect of the group and the individuals.

List of relevant publications

1. Papamitsiou, Z., & Economides, A.A. Motivating students in collaborative activities with game-theoretic group recommendations, *IEEE Trans. on Learning Technologies* (under revision)
2. Papamitsiou, Z., & Economides, A.A. (2018). Group-recommendation of educational resources: a game-theoretic approach, *9th IEEE Global Engineering Education Conference (EDUCON2018)* (760-767) [DOI:10.1109/EDUCON.2018.8363307](https://doi.org/10.1109/EDUCON.2018.8363307)
3. Papamitsiou, Z., & Economides, A.A. 2018. Can't get more satisfaction? Game-theoretic group-recommendation of educational resources, *8th Int. Conference on Learning Analytics and Knowledge* (409-416) NY, USA: ACM. [DOI:10.1145/3170358.3170371](https://doi.org/10.1145/3170358.3170371)

Finally **Chapter 14** presents the research conclusions, study contribution, implications for practice and recommendations for future research.

The description of the LAERS system that was developed during this research, and was used for data collection, analysis and reporting, is provided in Appendix A. A brief description of the Computerized Adaptive Testing algorithms and other formulas used in this study, as well as all factors employed in this study (tracked or measured with other instruments), and the definitions of all core terms are available in Appendices B and C, respectively.

List of relevant publications

1. Papamitsiou, Z. & Economides, A. A. (2013). Towards the alignment of computer-based assessment outcome with learning goals: the LAERS architecture, In *Proceedings of the IEEE e-Learning, e-Management and e-Services* (13-17). [DOI: 10.1109/IC3e.2013.6735958](https://doi.org/10.1109/IC3e.2013.6735958)
2. Papamitsiou, Z. & Economides, A. A. Towards exploiting learning analytics for supporting learners' autonomous decisions in online environments – The case of LAERS (To submit for review).

Chapter 2 : The State-of-the-Art in Learning Analytics – A decade of empirical research

"What is to give light must endure burning"
Viktor Frankl

Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence

2.1. Introduction

The information overload, originating from the growing quantity of “Big Data” during the past decade, requires the introduction and integration of new processing approaches into everyday objects and activities (“ubiquitous and pervasive computing”) (Cook & Das, 2012; Kwon & Sim, 2013). Handling large amounts of data manually is prohibitive. Several computational methods have been proposed in the literature to do this analysis.

In commercial fields, business and organizations are deploying sophisticated analytic techniques to evaluate rich data sources, identify patterns within the data and exploit these patterns in decision making (Chaudhuri, Dayal, & Narasayya, 2011). These techniques combine strategic planning procedures with informational technology instruments, summarized under the term “Business Intelligence” (Eckerson, 2006; Jourdan, Rainer, & Marshall, 2008). They constitute a well-established process that allows for synthesizing “vast amount of data into powerful decision making capabilities” (Baker, 2007, p. 2).

Recently researchers and developers from the educational community started exploring the potential adoption of analogous techniques for gaining insight into online learners’ activities. Two areas under development oriented towards the inclusion and exploration of big data capabilities in education are Educational Data Mining (EDM) and Learning Analytics (LA) and their respective communities.

EDM is concerned with “developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist” (Romero & Ventura, 2013, p. 12). Respectively, LA is an area of research related to business intelligence, web analytics, academic analytics, action analytics and predictive analytics. According to the definitions introduced during the 1st International Conference on Learning Analytics and Knowledge (LAK), LA is “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and environments in which it occurs” (<https://tekri.athabascau.ca/analytics/>).

Explaining the previous definitions, LA and EDM constitute an ecosystem of methods and techniques (in general procedures) that successively gather, process, report and act on machine-

readable data on an ongoing basis in order to advance the educational environment and reflect on learning processes. In general, these procedures initially emphasize on measurement and data collection and preparation for processing during the learning activities. Next, they focus on further analysis, reporting of data and interpretation of results, targeting to inform and empower learners, instructors and organization about performance and goal achievement, and facilitate decision making accordingly.

Both communities share similar goals and focus where learning science and data-driven analytics intersect. However, they differ in their origins, techniques, fields of emphasis and types of discovery (Chatti, Dyckhoff, Schroeder, & Thüs, 2012; Romero & Ventura, 2013; Siemens & Baker, 2012). Romero and Ventura (2013) presented an up-to-date comprehensive overview of the current state in data mining in education. In their review, the authors do not present research results as empirical evidence, but focus on the objectives, methods, knowledge discovery processes and tools adopted in EDM research. Analogous attempts were presented by Ferguson (2012) and Bienkowski et al. (2012) in the state of LA in 2012 and in an issue brief respectively.

All of these previous studies claim that as far as it concerns the approach to gaining insights into learning processes, LA adopts a holistic perspective, seeking to understand systems in their full complexity. On the other hand, EDM adopts a reductionistic viewpoint by analyzing individual components, seeking for new patterns in data and modifying respective algorithms. In other words, these two research areas are complementary and in order to capture the whole picture, someone should follow their traces alongside each other.

2.2. Motivation and Rationale of the study

The motivation for this review derived from the fact that empirical evidence is required for theoretical frameworks to gain acceptance in the scientific community. A search in relevant literature did not reveal any review of empirical evidence of the added value of research in both domains. Consequently, there was a need to supply the audience with an accredited overview. This survey aims to fill that gap.

The value of any single study is derived from how it fits with and expands previous work, as well as from the study's intrinsic properties. Thus, putting together all the unbiased, credible results from previous research would be a step towards understanding the whole picture and construct a map of our knowledge on the domain. In a sense, the rationale of this study was to manage the overwhelming amount of publications through a critical exploration, evaluation and synthesis of the previous empirical results that worth reflection.

This chapter's goal is to carry out a systematic review of empirical evidence in order to contribute towards: a) a complete documentation of the applied research approaches so far, b) a feasibility study that captures the strengths and weaknesses of research in the domain, and c) the identification of possible threats, and thus motivate the research community redefine or refine related questions or hypotheses for further research (opportunities).

2.3. The research questions

The following research questions need to be addressed, and are distinguished into primary (set to fulfill the goals of the review) and secondary (refine the primary, i.e., explanatory):

RQ1 (Primary): *Which are the basic research objectives of LA/EDM so far (in terms of measurable metrics), and which methods do researchers follow to achieve these goals?* **RQ1.1**

(Secondary): *What are the significant results from previous research that constitute empirical evidence regarding the impact of LA/EDM implementation (Efficacy of implementation)?*

RQ1.2 (Secondary): *What do these results indicate regarding the added value of this technology (Interpretation of the results)?*

RQ2 (Primary): *Which other emerging research technologies should be explored through the LA/EDM viewpoint?*

2.4. Research methodology

The followed methodology qualifies this survey as a systematic qualitative review of empirical research results concerning LA/EDM (Okoli & Schabram, 2010). In order to conduct the literature review we defined a review protocol, consisting of four discrete stages: a) searching the literature – data collection, b) reviewing and assessing the search results – selection of primary studies, c) analyzing, coding and synthesizing the results, and d) reporting the review.

During the first stage, our goal was to collect the appropriate studies. For that reason, we determined and accessed the article pool and declared the key search terminology. We extensively and iteratively searched international databases of authoritative academic resources and publishers, including Scopus, ERIC, Google Scholar, Science Direct, DBLP and ACM Digital Library. We also scanned International Journals and selected Conference Proceedings. The search terms included *learning analytics, learning analytics tools, learning analytics case studies, educational data mining, knowledge discovery in education*. The search process spanned from March 2013 to August 2013. The time frame of the search was bound within the last six years (2008-2013), in which emergence and adoption of LA/EDM has grown.

Due to the orientation of our work towards the practical implementation and exploitation of LA/EDM, at the end of the data collection stage, we explicitly determined the article inclusion/exclusion criteria (Table 2-1).

Table 2-1.Inclusion/exclusion criteria

Include	Exclude
<ul style="list-style-type: none"> • Articles published in Journals with Impact Factor • Full-length articles published in International Conference/Workshop Proceedings • Present quantitative results • Date from 2008 to 2013 	<ul style="list-style-type: none"> • Articles that do not present empirical data (e.g. theoretical and conceptual articles, essays, tool demonstration, etc.) • Short papers from conferences/workshops • Book chapters

The search procedure, and after deleting the duplicate records, yielded 209 results. 40 of them are published in International Journals and 169 articles were presented at International Conferences. Then we assessed the quality of the collected literature according to the following rigorous quantitative/qualitative rules:

- Number of citations
- Degree of underlying innovation (e.g. significant changes in techniques/equipment or software, as proposed by UNESCO¹)
- In depth illustration of the followed methodology (e.g. clear settings, fully-explained experimental procedure, etc.)
- Sufficient presentation of the findings (e.g. analytical discussion of findings and interpretation of results, use of figures and tables when needed, etc.)

Throughout the assessment procedure we identified that among the 209 retrieved articles, 40 of them were considered more central to our review (*key studies*), based on the combination of the above rules. Next, we proceeded on with an article classification according to the adopted research strategy (*category*), research discipline (*topic*), learning settings, research objectives (*goals*), data gathering (*sources and data-types*) and analysis technique (*method*) and results. Finally, we used non-statistical methods to evaluate and interpret findings of the collected studies, and conduct the synthesis of this review.

2.5. Limitations

Although there are papers concerning EDM that are published before 2008, this year was a landmark for the independent growth of this research domain (1st Conference in EDM - <http://www.educationaldatamining.org/EDM2008/>). For that reason we decided to examine the literature on experimental case studies conducted in the domain from 2008 to 2013. Furthermore, indicative reviews on former work can be found in Romero and Ventura (2007) and Romero and Ventura (2010).

We should note that, despite the fact that appreciated research works have been published in respectable conferences like Intelligent Tutoring Systems (ITS), Artificial Intelligence in Education (AIED) and User Modeling, Adaptation and Personalization (UMAP), in this review we included papers only from the EDM and LAK conferences. We acknowledge the leadeness of the previously mentioned conferences, but, in this work, we wanted to isolate the LA/EDM research and focus on its strengths, weaknesses, opportunities and threats.

We should also mention that the papers presented at the 6th International Conference on Educational Data Mining were excluded from the review process, since at the time of the search process these articles had had no citations. However, we decided to include recently published Journal articles (published in 2013) with lower number of citations, as indicative of current trends in the domain.

¹ <http://www.uis.unesco.org/ScienceTechnology/Pages/st-data-collection-innovation.aspx>

2.6. Results

In this section, we present our findings based on the analysis of the published case studies. We used non-statistical methods to evaluate and interpret findings of the collected studies.

According to the followed *research strategy*, most of the published case studies are exploratory or experimental studies. Some of them are evaluation studies, while others are empirical studies or surveys. Furthermore, the *research topics* differ from study to study, but most of them focus on science, technology, engineering, and mathematics (STEM).

Based on the *learning settings* of the studies (illustrated in Table 2-2), most studies are conducted within Virtual Learning Environments (VLEs) and/or Learning Management Systems (LMSs). Other popular learning settings are Cognitive Tutors (CTs), computer-based and web-based environments, mobile settings, and more recently, Massive Open Online Courses (MOOCs) and social learning platforms.

Table 2-2. Classification of case studies according to the learning settings

Learning setting*	Authors & Year (Paper Ref.)
<i>VLEs / LMSs</i>	Lin, Hsieh, & Chuang, 2009; Lykourantzou, Giannoukos, Mpardis, Nikolopoulos, & Loumos, 2009; Lykourantzou, Giannoukos, Nikolopoulos, Mpardis, & Loumos, 2009; Macfadyen & Dawson, 2010; Merceron & Yacef, 2008; Romero-Zaldivar, Pardo, Burgos, & Delgado Kloos, 2012; Romero, Ventura, Espejo, & Hervás, 2008; Tanes, Arnold, King, & Remnet, 2011
<i>MOOC/social learning</i>	Clow & Makriyannis, 2011; Fournier, Kop, & Sitlia, 2011; Kizilcec, Piech, & Schneider, 2013
<i>Web-based education</i>	Abdous, He, & Yen, 2012; Giesbers, Rienties, Tempelaar, & Gijsselaers, 2013; He, 2013; Khribi, Jemni, & Nasraoui, 2009; Li, Cohen, Koedinger, & Matsuda, 2011; Romero, Ventura, Zafra, & Bra, 2009
<i>Cognitive tutors</i>	Baker, Corbett, & Aleven, 2008; Moridis & Economides, 2009b; Pardos, Baker, San Pedro, Gowda, & Gowda, 2013; Shih, Koedinger, & Scheines, 2008
<i>Computer-based education</i>	Ali, Hatala, Gašević, & Jovanović, 2012; Barla et al., 2010; Blikstein, 2011; Jeong & Biswas, 2008; Levy & Wilensky, 2011; Santos, Govaerts, Verbert, & Duval, 2012; Thai-Nghe, Horváth, & Schmidt-Thieme, 2011
<i>Multimodality</i>	Worsley & Blikstein, 2013
<i>Mobility</i>	Chen & Chen, 2009; Leong, Lee, & Mak, 2012

* **VLEs/LMSs:** controlled environment (VLE/LMS), used for gathering learner and activity data, **MOOC/social learning:** informal, social learning setting, **Web-based education:** web-based e-learning environments except from VLEs, LMSs and MOOCs, **Cognitive tutors:** special software, utilized for the needs of the study, **Computer-based education:** other environments that include some type of computer technology (e.g. desktop applications, etc.) except from those belonging to one of the other categories, **Multimodality:** learner data in different modalities, **Mobility:** mobile devices used as the primary learning mediator.

The authors *gathered data* from different *data sources*, including log files from the goal-oriented implemented systems, questionnaires, interviews, Google analytics, open datasets from the dataTEL Challenge (<http://www.teleurope.eu/pg/groups/9405/datatel/>), virtual machines, and many more. In particular, researchers tracked different *types of data* in order to measure students' participation and login frequency, number of chat messages between participants and questions submitted to the instructors, response times on answering questions and solving tasks, resources accessed, previous grades, final grades in courses, detailed profiles, preferences from LMSs, forum and discussion posts, affect observations (e.g. bored, frustrated, confused, happy, etc.) and many more.

Another important parameter is the *data mining method* adopted by authors *to analyze* the gathered data. In the field of LA/EDM, the most popular method is classification, followed by clustering, regression (logistic/multiple) and more recently, discovery with models. In addition, *algorithmic criteria* computed for comparison of methods include precision, accuracy, sensitivity, coherence, fitness measures (e.g. cosine, confidence, lift, etc.), similarity weights, etc. Table 2-3 displays the classification of the key studies according to the data mining method they adopt.

Table 2-3. Classification of case studies according to the analysis method

Data analysis method	Authors & Year (Paper Ref.)
<i>Classification</i>	Baker et al., 2008; Barla et al., 2010; Chen & Chen, 2009; Dejaeger, Goethals, Giangreco, Mola, & Baesens, 2012; Dekker & Vleeshouwers, 2009; Guo, 2010; Guruler, Istanbulu, & Karahasan, 2010; Huang & Fang, 2013; Jeong & Biswas, 2008; Khribi et al., 2009; Kizilcec et al., 2013; Klašnja-Milićević, Vesin, Ivanović, & Budimac, 2011; Li et al., 2011; Lin, Yeh, Hung, & Chang, 2013; Lykourantzou, Giannoukos, Mpardis, et al., 2009; Lykourantzou, Giannoukos, Nikolopoulos, et al., 2009; Moridis & Economides, 2009b; Pardos et al., 2013; Romero et al., 2008; Thai-Nghe et al., 2011
<i>Clustering</i>	Abdous et al., 2012; Chen & Chen, 2009; Khribi et al., 2009; Kizilcec et al., 2013; Klašnja-Milićević et al., 2011; Lykourantzou, Giannoukos, Mpardis, et al., 2009; Romero et al., 2009
<i>Regression</i>	Abdous et al., 2012; Macfadyen & Dawson, 2010; Romero-Zaldivar et al., 2012
<i>Text mining</i>	He, 2013; Leong et al., 2012; Lin et al., 2009
<i>Association rule mining</i>	Merceron & Yacef, 2008; Romero et al., 2009
<i>Social Network Analysis</i>	Fournier et al., 2011; Macfadyen & Dawson, 2010
<i>Discovery with models</i>	Ali et al., 2012; Pardos et al., 2013; Shih et al., 2008
<i>Visualization</i>	Clow & Makriyannis, 2011; Fournier et al., 2011; Santos et al., 2012
<i>Statistics</i>	Giesbers et al., 2013; Guo, 2010

The article classification according to *the research objectives (goals)* is illustrated in Table 2-4. As seen in this table, the majority of studies investigate issues related to student/student behavior modeling and prediction of performance, followed by increase of students' and teachers' reflection and awareness and improvement of provided feedback and assessment services.

Table 2-4. Classification of case studies according to the research objectives

Research objectives (goals)	Authors & Year (Paper Ref.)
<i>Student/Student behavior modeling</i>	Abdous et al., 2012; Baker et al., 2008; Blikstein, 2011; Fournier et al., 2011; He, 2013; Jeong & Biswas, 2008; Kizilcec et al., 2013; Levy & Wilensky, 2011; Li et al., 2011; Pardos et al., 2013; Romero et al., 2008; Shih et al., 2008
<i>Prediction of performance</i>	Abdous et al., 2012; Huang & Fang, 2013; Lykourantzou, Giannoukos, Mpardis, et al., 2009; Macfadyen & Dawson, 2010; Moridis & Economides, 2009b; Pardos et al., 2013; Romero-Zaldivar et al., 2012; Romero et al., 2008; Shih et al., 2008; Thai-Nghe et al., 2011
<i>Increase (self-) reflection & (self-) awareness</i>	Ali et al., 2012; Clow & Makriyannis, 2011; Fournier et al., 2011; Macfadyen & Dawson, 2010; Santos et al., 2012
<i>Prediction of dropout & retention</i>	Dejaeger et al., 2012; Dekker & Vleeshouwers, 2009; Giesbers et al., 2013; Guo, 2010; Guruler et al., 2010; Kizilcec et al., 2013; Lykourantzou, Giannoukos, Nikolopoulos, et al., 2009
<i>Improve assessment & feedback services</i>	Ali et al., 2012; Barla et al., 2010; Chen & Chen, 2009; Leong et al., 2012; Tanes et al., 2011; Wilson, Boyd, Chen, & Jamal, 2011; Worsley & Blikstein, 2013
<i>Recommendation of resources</i>	Khribi et al., 2009; Klašnja-Milićević et al., 2011; Romero et al., 2009; Thai-Nghe et al., 2011; Verbert et al., 2011

2.7. Key studies analysis

In this section we present the findings of the review process and answer on the initially set research questions RQ1 and RQ1.1. The rest of the research questions (mostly the results of the case studies and their comparative evaluation, as well as current and future trends, possible gaps and new research directions) are discussed in next section.

RQ1: *Which are the basic research objectives of LA/EDM so far (in terms of measurable metrics), and which methods do researchers follow to achieve these goals?*

Student/student behavior modeling

As seen from table 2-4, detection, identification and modeling of students' learning behavior is a primary research objective. More specifically, the authors seek to identify learning strategies and when they occur, and model affective and metacognitive states (Abdous et al., 2012; Baker et al., 2008; Blikstein, 2011; Jeong & Biswas, 2008; Levy & Wilensky, 2011; Shih et al., 2008). For example, Abdous, He and Yen (2012) and He (2013) tried to correlate interactions within a Live Video Streaming (LVS) environment to students' final grades in order to predict their performance, discover behavior patterns in LVSS that lead to increased performance, and understand the ways students are engaged into online activities. In another case study, Blikstein (2011) logged automatically-generated data during programming activity in order to understand students' trajectories and detect programming strategies within Open-Ended Learning Environments (OELEs). Furthermore, Shih, Koedinger and Scheines, (2008) used worked

examples and logged response times to model the students' time-spent in terms of "thinking about a hint" and "reflecting on a hint" for capturing behaviors that are related to reasoning and self-explanation during requesting hints within a CT environment. In another self-reasoning example, Jeong and Biswas (2008) tried to analyze students' behavior based on the sequence of actions, and to infer learning strategies within a teachable agent environment.

Another orientation is the discovery and modeling of the respective behaviors within MOOCs (Fournier et al., 2011; Kizilcec et al., 2013). The authors tried to identify meaningful, high-level patterns of participation, engagement and disengagement in learning activities in this recently introduced learning setting.

Updated and extended review on student modeling approaches can be found in Chrysafiadi and Virvou (2013b) and in Pena-Ayala (2014).

Prediction of performance

Authors also explore, identify and evaluate various factors as indicators of performance for prediction purposes. Among these factors, demographic characteristics, grades (in prerequisite courses, during assessment quizzes and their final scores), students' portfolios, multimodal skills, students' participation, enrollment and engagement in activity and students' mood and affective states are acknowledged as the most common ones (Abdous et al., 2012; Huang & Fang, 2013; Lykourantzou, Giannoukos, Mpardis, et al., 2009; Macfadyen & Dawson, 2010; Moridis & Economides, 2009b; Pardos et al., 2013; Romero-Zaldivar et al., 2012). For example, Macfadyen and Dawson (2010) examined the effect of variables tracked within an LMS-supported course (e.g. total number of discussion messages posted, total time online, number of web links visited, etc.) on students' final grade. In another example, Lykourantzou, Giannoukos, Mpardis, et al., (2009) used neural networks to accurately cluster students at early stages of a multiple choice quiz activity.

Moreover, researchers investigated affective factors that influence learning outcomes (within an ITS) (Pardos et al., 2013) and used simulation environments (virtual appliances that appear to learners as regular desktop applications) to monitor students and predict their performance (Romero-Zaldivar et al., 2012). Pardos et al. (2013) employed discovery with models, post-hoc analysis of tutor logged data and sensor-free detectors of affect (based on classification algorithms), while Romero-Zaldivar et al. (2012) tracked events (such as work-time, commands, compile, etc.) and analyzed the gathered data with multiple regression for the estimation of the variance of performance.

Increase (self-)reflection and (self-)awareness

Another crucial issue in EDM/LA research that authors attempt to address is how to increase the instructors' awareness, identify "disconnected" students and evaluate visualizations regarding their capabilities on informing students about their progress and compared to peers. In order to provide instructors with pedagogically meaningful information and to help them

extract such information on their own, the researchers embedded multiple representations of feedback types (Ali et al., 2012) and multiple widget technology for personalization of learning environments (Santos et al., 2012). Alternatively, content analysis for threaded discussion forums was explored regarding its monitoring capabilities (Lin et al., 2013). In particular, the authors aimed to facilitate the automated coding process within a repository of postings in an online course, in order less monitoring of the discussion to be needed by the instructor. Furthermore, Merceron and Yacef (2008) employed association rule mining to extract meaningful association rules to inform teachers about usage of extra learning material. The authors investigated students' usage of learning resources and self-evaluation exercises and its impact on final grades.

In the context of social/open learning, the researchers explored the usefulness and motivation capabilities of dashboard-like applications regarding their self-reflection and self-awareness opportunities (Clow & Makriyannis, 2011). In particular, the request was related to the effect of "expert users" presence on participants' awareness of their own contribution and participation in an online reputation system with positive feedback only. Furthermore, Fournier et al. (2011) searched for crucial moments of learning based on interactions in MOOCs. In this case, the authors examined the impact of visualized provision of useful information to social learners regarding their participation and social interactions.

Prediction of dropout and retention

Prediction of dropout and retention are also key issues for LA/EDM research. In order to predict students' dropout at early stages, Lykourantzou, Giannoukos, Nikolopoulos, et al., (2009) applied a combination of three machine learning techniques on detailed students' profiles from an LMS environment. The authors compared the accuracy, sensitivity and precision measures of the proposed method to others in literature. From a similar point of view, Dekker et al. (2009) tried to predict students' dropout and identify factors of success based on the use of different classification algorithms. The authors compared the accuracy and performance of these algorithms to make a selection between classifiers. In particular, they used classifiers for prediction of dropout based on simple "early" data (from first year enrollment) and boosted accuracy with cost-sensitive learning.

More recently, Kizilcec et al. (2013) classified learners according to their interactions (video lectures and assessment) with course content in learning activities in MOOCs. Next, they clustered engagement patterns, and finally, they compared clusters based on learners' characteristics and behavior.

The issue of motivating engagement in learning activities and consequently increasing students' satisfaction and retention was also explored (Dejaeger et al., 2012; Giesbers et al., 2013; Guo, 2010; Guruler et al., 2010). Demographics and factors like achievement rates and final performance were associated to students' motivation to remain engaged and actively enrolled in courses. Identification of success factors urged Giesbers et al. (2013) to investigate the

relationship between observed student behavior (i.e. actual usage of synchronous tools), motivation, and performance on a final exam. The researchers explored whether actual usage of synchronous tools increases the motivation to participate in online courses that support these tools. Similarly, Guo (2010) used statistical measures and neural network techniques for prediction of students' retention. The researcher examined the number of students enrolled in each course and the distinction rate in final grades. Furthermore, Dejaeger et al. (2012) explored measures of students' satisfaction for retaining student population. The authors investigated a number of constructs of satisfaction (e.g. perceived usefulness of training, perceived training efficiency, etc.) along with class related variables.

Improve feedback and assessment services

Many researchers have explored the use of LA/EDM in producing meaningful feedback. Feedback is strongly related to reflection and awareness and could be informative regarding students' dropout intentions. For that reason, provision of appropriate forms/types of feedback was a major issue for Ali et al. (2012), Clow and Makriyiannis (2011) and Macfadyen and Dawson (2010), formerly presented. Visualization of feedback was also crucial for Tanes et al. (2011). The authors explored instructors' perceptions of feedback types in relation to students' success. Complementary to that, in mobile learning contexts, Leong et al. (2012) explored the impact and usefulness of SMS free-text feedback to teacher regarding the feelings of students, after a lecture. The goal was to visualize positive and negative aspects of the lecture by taking advantage of the limited SMS length and the use of emoticons in order to provide free-text feedback to teacher.

In addition to these studies, an extensive area of LA/EDM research deals with issues related to using LA/EDM for adaptive assessment of goal achievement during activities. The landscape in this domain is quite distributed and diverse. Selection of the most appropriate next task during adaptive testing, students' satisfaction level during mobile formative assessment, as well as construction of sophisticated measures of assessment (Barla et al., 2010; Chen & Chen, 2009; Wilson et al., 2011; Worsley & Blikstein, 2013) have emerged. Barla et al. (2010) focused on assessment capabilities of EDM methods and combined three different classification methods for selection of the most appropriate next task during adaptive testing. In a different context, Chen and Chen (2009) developed a tool that uses six computational intelligence theories according to the web-based learning portfolios of an individual learner, in order to measure students' satisfaction during mobile formative assessment. Furthermore, Worsley and Blikstein (2013) aimed to detect metrics that could be used primarily as formative assessment tools of sophisticated learning skills acquisition in process-oriented assessment. A combination of speech recognition with knowledge tracing was proposed as method for multimodal assessment.

Recommendation of resources

Another major issue in dataset-driven research concerns data resources and their management. Research in this domain focuses on a technical aspect. The approaches include

similarity calculation mechanisms deployment, comparison of the performance of different mining algorithms, aggregation of different datasets in the context of dataset-driven research, suggestion of infrastructures for storing and forwarding learning-resources metadata (Romero et al., 2009; Thai-Nghe et al., 2011; Verbert et al., 2011) for resource recommendation in larger scale and across different contexts.

Examples of algorithmic approaches also include recommendations according to the affective state of the learner (Santos & Boticario, 2012), implementation of collaborative filtering to sequence learning activities, hybrid recommendations based on learner and content modeling (Khribi et al., 2009; Klačnja-Milićević et al., 2011) and more.

Verbert et al. (2011) presented an analysis of publicly available datasets for TEL that can be used for LA in order to support recommendations (of resources or activities) for learning. In addition, the authors evaluated the performance of user-based and item-based collaborative filtering algorithms and measured their accuracy and coverage through metrics implementation. Moreover, Romero et al. (2009) explored user profile information and web-usage mining for recommendation of resources (here, hyperlinks). The authors compared the performance of three different mining algorithms. A comprehensive review on recommender systems in the TEL context can be found in Manouselis et al. (2013).

RQ1.1: *What are the significant results from previous research that constitute empirical evidence regarding the impact of LA/EDM implementation?*

According to the research objectives explored by the authors, Table 2-5 displays a categorization of the algorithmic-oriented findings from the collected studies.

Table 2-5. Classification of the results of LA/EDM case studies (algorithmic)

Objective	Results
<i>Student/ student behavior modeling</i>	<ul style="list-style-type: none"> Quantitative analysis could be applied for reporting on participants' activity, while qualitative analysis could be more effective on revealing deeper concepts related to learning (Fournier et al., 2011). Comprehensibility of the results strongly depends on human judgment - produced models are not equally interpretable by the teachers (Fournier et al., 2011; Romero et al., 2008).
<i>Prediction of performance</i>	<ul style="list-style-type: none"> Adding more predictor variables does not help improve the average prediction accuracy of the mathematical models explored by Huang and Fang (2013) for prediction of performance. However, neural networks method leads to better prediction results compared to those of the regression analysis method (Lykourantzou, Giannoukos, Mpardis, et al., 2009). Reducing the size of the training set by removing very high and very low probabilities of obtaining a correct answer without knowing the skill and obtaining an incorrect answer even though the student knows the skill, and forecasting techniques that embed sequential information (temporality) into the factorization process may improve the predictive model of students' performance (Baker et al., 2008; Thai-Nghe et al., 2011).

<i>Increase (self-) reflection & (self-) awareness</i>	<ul style="list-style-type: none"> • Genre classification methods can automate the coding process in a forum and handle issues like imbalanced distribution of discussion postings (Lin et al., 2009) • LMS log data are not data mining “friendly” (i.e. not stored the same way, data consolidation requires complex manipulations, etc.) (Merceron & Yacef, 2008). • Comparison of measures of interestingness of association rules did not significantly improved decision making for discarding a rule (Merceron & Yacef, 2008) .
<i>Prediction of dropout & retention</i>	<ul style="list-style-type: none"> • Combination of machine learning techniques afforded more reliable results, which however, depend on the level of detail of available students’ data (Lykourantzou, Giannoukos, Nikolopoulos, et al., 2009). • Simple classifiers had higher accuracy than sophisticated ones and cost-sensitive learning helps to bias classification errors (Dekker & Vleeshouwers, 2009). • While investigating disengagement in MOOCs, the cross-cluster comparison can help understanding the reasons learners remain to a cluster (Kizilcec et al., 2013).
<i>Recommendation of resources</i>	<ul style="list-style-type: none"> • A combination of students’ clustering and sequential pattern mining is suitable for the discovery of personalized recommendations (Romero et al., 2009), while content based filtering and collaborative filtering approaches are valid recommendation strategies (Khribi et al., 2009), but further research should be conducted (Verbert et al., 2011).

Table 2-6 displays a categorization of the pedagogy-oriented findings. The learning context of the studies has been taken under consideration, as well. That is because we wanted to maintain the targeted applicability of the results.

Table 2-6. Classification of the results of LA/EDM case studies (pedagogical)

Objective	Results	
	Formal Learning	Non-Formal Learning
<i>Student/ student behavior modeling</i>	<ul style="list-style-type: none"> • Students’ detected critical moments during programming reflect students’ behavior and their perceived learning benefits, both in Secondary and Higher Education (Blikstein, 2011; Levy & Wilensky, 2011). • In secondary education, learning by teaching provides more opportunities for retaining metacognitive learning strategies (Jeong & Biswas, 2008), while worked examples are effective indicators of self-explanation and learning gain (Shih et al., 2008). • Specifying the moments that teacher should intervene requires to better distinguish between students who use worked examples, how they use them and their response times (Shih et al., 2008). 	<ul style="list-style-type: none"> • The presence of “experts” (that is users or organizations with advanced expertise or reputation on the field of study) has a significant impact on the highly unequal distribution of activities within a functioning social network (Clow & Makriyannis, 2011). • At a level of interactivity among learners or between learners and teachers, questions posed to instructors and chat messages posted among students (both in number and their content) are correlated (Abdous et al., 2012). • The discovery of four trajectories (auditing, completing, disengaging, sampling learners) (Kizilcec et al., 2013) roughly describes engagement that makes sense in MOOCs.
<i>Prediction of performance</i>	<ul style="list-style-type: none"> • Both in Secondary and Higher Education, the number of quizzes 	<ul style="list-style-type: none"> • Abdous, He and Yen (2012) couldn’t predict performance based on

passed is the main determinant of performance (i.e. the final grade), while others, such as number of posts, frequencies of the events and time-spent could identify activities that are related to higher or lower marks (Romero-Zaldivar et al., 2012; Romero et al., 2008; Shih et al., 2008).

- In Secondary Education, affective states like engaged concentration and frustration are correlated with positive learning outcomes, while boredom and confusion are negatively correlated with performance (Pardos et al., 2013)
- In Secondary Education, and from the instructors' perspective, coding discussion posts in a forum can assist the teacher to automatically monitor the forum and maintain its quality (Lin et al., 2009).
- In Higher Education, meaningful rules increases teachers' awareness regarding the students' usage of additional material within LMSs (Merceron & Yacef, 2008).
- SMS text increases instructor's awareness on students' affective states in order to modify the lecture (Leong et al., 2012).

Increase (self-) reflection and (self-) awareness

- Dashboard-like applications and multiple feedback representations could increase (self-)awareness and perceived value of provided feedback (Ali et al., 2012; Macfadyen & Dawson, 2010; Santos et al., 2012).
- From the learners' point of view, students want to be aware of what their peers are doing, but they don't like to be tracked outside course environment due to privacy concerns (Santos et al., 2012).
- Identification of disconnected students based on their networking activity ended up to clusters of students with similar participatory behavior (Macfadyen & Dawson, 2010).

Prediction of dropout and retention

- Monitoring students' activity with virtual machines and applying data-driven machine learning methods on students' profiles and log files (mostly grades and assessment quiz scores) from LMS allow detecting students at-risk at an early stage (Lykourantzou, Giannoukos, Nikolopoulos, et al., 2009; Romero-Zaldivar et al., 2012).
- Students want to feel that they belong to the course in order to engage and enroll (Guo, 2010). Improving students' course satisfaction can be used to reduce the dropout (Dejaeger et al., 2012; Guo, 2010).

Improve feedback and

- In Higher Education, adaptive selection of the most appropriate next task improved testing outcomes mostly for below-average students (Barla et al., 2010).

students' participation and online interactions.

- Giesbers et al (2013) and Macfadyen and Dawson (2010) found a significant positive relationship between participation and grades.

- In MOOCs, the most common detected disengagement reasons were personal commitments, work conflict and course overload (Kizilcec et al., 2013).
- In web-videoconference settings there was not found a relation between motivation to participate and dropout (Giesbers et al., 2013).
- Types of registration to the university as well as the family income seem to affect more the students' retention (Guruler et al., 2010).

<i>assessment services</i>	<ul style="list-style-type: none"> • In Elementary Education, web-based learning portfolios of an individual learner during mobile formative assessment granted similar results to those of summative assessment (Chen & Chen, 2009).
<i>Recommendation of resources</i>	<ul style="list-style-type: none"> • Additional learner attributes (e.g. experience level indicators, learning interests, learning styles, learning goals and competences and background information), student's expected performance on tasks, his recent navigation history (within a number of resources) or learner's affective traits should be taken under consideration in recommendation processes (Khribi et al., 2009; Klašnja-Milićević et al., 2011; Santos & Boticario, 2012; Thai-Nghe et al., 2011; Verbert et al., 2011).

2.8. Discussion and future research

From the former analysis it becomes apparent that recently, the educational research community has started applying sophisticated algorithmic methods on gathered (mostly raw) data for understanding learning mechanisms through an in-depth exploration of their relations and meaning. As seen in Tables 2-5 and 2-6, the landscape of the LA/EDM research combines diverse and often conflicting aspects and results related to gaining insight into learning processes. However, the above results have highlighted four distinct major axis of the LA/EDM empirical research including:

a) *Pedagogy-oriented issues* (e.g. student modeling, prediction of performance, assessment and feedback, reflection and awareness): several studies focus on pedagogically meaningful analysis on collected students' data in order to shed light to the whole picture from students/students' behavior modeling to self-regulated learning.

b) *Contextualization of learning* (e.g. multimodality, mobility, etc.): a number of studies gathered data from the learning context itself and focus on positioning learning within specific conditions and attributes.

c) *Networked learning* (e.g.: MOOCs, social learning platforms, etc.): some case studies try to identify patterns within the social aspect of learning and the MOOCs, where the number of participants rapidly increases and the interactions between learners and the learners and the content are text/video-based.

d) *Educational resources handling*: fewer, but not neglected studies raise the issue of organizing and recommending educational resources from data pools, and selecting the most appropriate algorithmic method for making suggestions.

However, these four axis are not completely autonomous, since significant overlaps may occur. For example, student modeling (i.e. a pedagogy-oriented issue) can still be explored in MOOCs (i.e. a form of Networked learning). However, this statement could only constitute a limitation which does not deduce the added value of the findings.

RQ1.2: *What these results indicate regarding the added value of this technology?*

One of the most important goals of the systematic review was to reveal the added value of the field explored. From the above analysis of findings derives that analysis of user interactions in order to “control” the information generated through technology has always been a request. LA/EDM research results indicate that data integration from multiple sources can improve the accuracy of a learner profile and subsequent adaptation and personalization of content. Exploration of students’ behavior within educational contexts that support multimodality and mobility could lead to shaping a holistic picture of how, when and where learning occurs.

Researchers set the educational context within limits in which previously it was almost impossible to infer behavior patterns, due to their high levels of granularity. In such advanced learning contexts, LA/EDM research community determines simple and/or sophisticated factors as predictors of performance and explores their predictive value and capabilities by tracking actual data and changes on behavioral data. The goal is to identify the most significant factors in order to develop better systems. These systems will allow students to monitor their own progress and will help them evaluate and adjust their learning strategies to improve their performance in terms of learning outcomes.

Moreover, the social dimension of learning and the opportunity of selectively participating in MOOCs are also explored with encouraging results. Consequently, the research community could gain insight into the learning mechanisms that previously were a “black box”.

RQ2: *Which other emerging research technologies should be explored through the LA/EDM viewpoint?*

Complementary, the literature overview has revealed a number of unexplored issues in this rapidly grown domain, including (but not limited to) the following:

Suggested incorporation of other emerging research technologies with LA/EDM

1. Game-based learning (GBL) has been acknowledged for its positive impact on learners. According to Collony et al. (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012, p. 1), “playing computer games is linked to a variety of perceptual, cognitive, behavioral, affective and motivational impacts and outcomes”. One interesting research question is if and how LA/EDM methods could be applied to report and visualize learning processes during GBL. In other words, how can LA/EDM be applied on GBL to detect patterns and construct measures that are transferable to other OELEs, in order to assess advanced skills development.
2. Another field evolving in a rapid pace is mobile and ubiquitous learning. Mobile learning has been acknowledged for the unique opportunity of offering authentic learning experiences anytime and anywhere (Tatar, Roschelle, Vahey, & Penuel, 2003). Although two of the selected studies were conducted in a mobile context (Chen & Chen, 2009; Leong et al., 2012), none of them associated or explored the effect of the context on the attained results. LA/EDM

research could investigate the appropriateness of the popular methods in the above context in order to provide sophisticated, personalized learning services in mobile applications.

3. Furthermore, according to Piaget's theory of cognitive development, sensorimotor learning is the first stage of human learning (Piaget, 1952). Sensorimotor learning refers to improvement, through practice, in the performance of sensory-guided motor behavior (Krakauer & Mazzoni, 2011). Due to its high relevance to the brain anatomy and functionality, sensorimotor learning has recently been under the lenses of neuroscience research (e.g., Catmur, 2013). LA/EDM has not been previously examined for sensorimotor learning or combined to neuroscience research. It would be interesting to study transformation of learning experience into strategy development (knowledge transfer) by exploring big neuroscience data.
4. Technology acceptance is also a well addressed issue in educational research. Regarding learning analytics acceptance, Ali et al. (2012) proposed a model that considers only two parameters – ease of use and perceived usefulness. However, more parameters should be explored in order to create a reliable learning analytics acceptance model. An appreciated model for computer based assessment acceptance was proposed by Terzis and Economides (2011). Researchers from the LA/EDM domain could also examine respective models that are suitable for the purposes of LA tools.
5. Finally, the review process didn't yield any article related to learning "meta"-analytics (i.e. feeding machine readable results from the LA/EDM procedures to another data-driven system for diving decision making without the mediation of the human judgment parameter). It would be interesting to take advantage of the plethora of results from LA/EDM research towards introducing innovative intelligent tutoring systems or fully automated educational recommender systems.

2.9. Conclusions

Previous reviews on LA/EDM research provided significant insight into the conceptual basis of this rapidly growing domain. However, these studies were either focused solely on LA or EDM, or they did not conduct an analysis of empirical research results. The current study presents a systematic review of empirical evidence of LA/EDM research. We searched the literature and gathered representative, mature and highly-cited articles of real case studies with actual data, both from LA and EDM domains. The analysis of selected case studies and their results shed light on the approaches followed by the respective research communities and revealed the potential of this emerging field of educational research. Along with the arising opportunities, we discovered a number of gaps that require the researchers' attention. Table 2-7 illustrates our findings regarding the strengths, weaknesses, opportunities and threats (SWOT) of LA/EDM research.

Table 2-7. SWOT of LA/EDM research

<p>Strengths</p> <ul style="list-style-type: none"> • <i>Large volumes of available educational data → increased accuracy of experimental results.</i> • <i>Use of pre-existing powerful and valid algorithmic methods.</i> • <i>Interpretable multiple visualizations to support learners/teachers.</i> • <i>More precise user models for guiding adaptation and personalization of systems.</i> • <i>Reveal critical moments and patterns of learning.</i> • <i>Gain insight to learning strategies and behaviors.</i> 	<p>Weaknesses</p> <ul style="list-style-type: none"> • <i>Misinterpretation of results due to human judgment factors - focus on reporting, not decision.</i> • <i>Heterogeneous data sources: not yet a unified data descriptive vocabulary – data representation issues.</i> • <i>Mostly quantitative research results. Qualitative methods have not yet provided significant results.</i> • <i>Information overload – complex systems.</i> • <i>Uncertainty: “are we ready yet?” So far, only skilled teachers/instructors could interpret the results correctly.</i>
<p>Opportunities</p> <ul style="list-style-type: none"> • <i>Use of Open Linked Data for data standardization and compatibility among different tools and applications → generalized platform development.</i> • <i>Multimodal and affective learning opportunities based on sophisticated metrics.</i> • <i>Self-reflection/ self-awareness/ self-learning in intelligent, autonomous and massive systems.</i> • <i>Feed machine readable results from the LA/EDM procedures to other data-driven systems for diving decision making.</i> • <i>Acceptance Model: e.g. perceived usefulness, perceived ease of use, perceived playfulness, trust, goal expectancy, social influence.</i> 	<p>Threats</p> <ul style="list-style-type: none"> • <i>Ethical issues – data privacy.</i> • <i>Over-analysis: the depth of analysis becomes profound and the results lack generality. The “over-granularity” approaches so far might threaten the holistic picture being explored; look at the tree and miss the forest.</i> • <i>Possibility of pattern misclassification.</i> • <i>Trust: contradictory findings during implementations.</i>

Beyond learning perceptions and attitudes collected through questionnaires, every “click” within an electronic learning environment may be valuable actual information that can be tracked and analyzed. Every simple or more complex action within such environments can be isolated, identified and classified through computational methods into meaningful patterns. Every type of interaction can be coded into behavioral schemes and decoded into interpretable guidance for decision making. This is the point where learning science, psychology, pedagogy and computer science intersect. The issue of understanding the deeper learning processes by deconstructing them into more simple, distinct mechanisms remains in the middle of this cross-path.

We believe that this active research area will continue contributing with valuable pieces of work towards the development of powerful and mostly accurate learning services both to learners and teachers.

Chapter 3 : Learning Analytics for Explanation of Performance in online self-assessment tests

“It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”

Arthur Conan Doyle

Explaining learning performance using response-time, effort, self-regulation and satisfaction from content: analytical approaches with machine learning and fsQCA

3.1. Introduction

Computer Based Assessment (CBA or e-assessment) is the use of information technologies to mechanize and facilitate assessment and feedback processes. CBA has been introduced to alleviate the practical problems emerged by large classes and to help both teachers and students evaluate the students' strengths and weaknesses (Gvozdenko & Chambers, 2007; Joosten-ten Brinke et al., 2007; Triantafyllou, Georgiadou, & Economides, 2008). The use of e- assessment purposes allows for remote progress tracking and evaluation and is strongly correlated to increased student and teacher (self-) awareness regarding student's learning achievements (Leony, Muñoz-Merino, Pardo, & Kloos, 2013). This is one of the reasons why improvement of CBA services has been under the lenses of learning analytics research and constitutes one of its main objectives (Chatti et al., 2012; Papamitsiou & Economides, 2014a); Assessment analytics concerns applying sophisticated, fine-grained analytic methods on multiple types of data for revealing the intelligence held in CBA systems.

Researchers attempt to detect and understand the most significant factors (i.e. with the higher impact or predictive capabilities) that affect students' performance during CBA. A body of literature explored, identified and evaluated factors as indicators of performance that are more significant for justification purposes. For example, research shown a significant positive relationship between participation and grades (Triantafyllou et al., 2008), whereas performance is also related to the type, content and nature of feedback (formative/summative) students receive during assessment (Gvozdenko & Chambers, 2007; Leony et al., 2013).

The interest in explaining performance in online assessment tests derive from the fact that these tests are a typical and popular format for the evaluation of knowledge acquisition (Arnold, 2016; Kim, Smith, & Maeng, 2008). In general, testing procedures are treated by the teachers worldwide as “diagnostic tools” to gradually mark their students' progress on the course and measure the learning gain, i.e., the learning performance (Challis, 2005). It is common practice to use tests to measure academic performance, since they setup a preamble of students' overall achievements on a specific course; grades are required at the end of the course and they are critical to the students' academic success. Thus, tests could be regarded as a “mean” to early distinguish students who are likely to achieve high or dropout.

The existing methods (i.e., Classical Test Theory and Computerized Adaptive Testing) have

provided well-established testing formats. However, assessment tests have received comprehensive criticism; chasing grades may distract students from deeper learning (Wolsey, 2008), yet good grades do not necessarily reflect mastery (Davis, 1999) and put academic honesty in question, since they are conducive to cheating (Arnold, 2016). Gaining in-depth insight of students' interactions and seeking for explanation of their actions in testing contexts is a demand to further interpret the test result, and the overall learning gain and performance.

Towards understanding students' behavior during assessment tests, prior studies have contributed by holistically exploring students' response-time, i.e., by analyzing the amounts of time the students allocate on test items (W. J. van der Linden, 2009). It was claimed that response-time should be treated as a fixed predictor (Wang & Hanson, 2005). It was also suggested that considering additional students' attributes – beyond response-time – might provide more concise prediction of their score (Xiong, Pardos, & Heffernan, 2011). The investigation of symmetric dependencies between goal expectations, correctness of answers, response-time and performance, provided encouraging findings (Papamitsiou, Terzis, & Economides, 2014). Previous research also revealed that the students' performance is strongly associated with their perception of task difficulty and on-task mental effort (Capa, Audiffren, & Ragot, 2008; Papamitsiou & Economides, 2015). Self-regulatory strategies and motivation (Beck, 2005; Broadbent & Poon, 2015; Hodges & Kim, 2010; Kitsantas, 2002), content validity and satisfaction from assessment items (Doll & Torkzadeh, 1988; Fitzpatrick, 1983; Lo, 2010; Puzziferro, 2008), demographic backgrounds (Lee & Haberman, 2016) and personality traits (Papamitsiou & Economides, 2017) have been identified as factors that affect learning performance as well.

However, it is quite unclear which combinations of the previously identified factors better explain the obtained assessment outcome. Towards seeking for specific patterns and configurations that foretell and explain students' performance and achievement level on assessment tests, two different studies were conducted, employing different analytical methods, i.e., a series of supervised learning techniques (e.g., Artificial Neural Networks – ANNs, Support Vector Machines – SVM, Naïve Bayes – NB, k-Nearest Neighbors – kNN and the treeBagger) (Alpaydin, 2010), and fuzzy set qualitative comparative analysis (fsQCA) (Ragin, 2008).

3.2. Related work

Explaining the students' learning performance and achievements is a timeless research topic. Over the past decades, several studies have reported results for addressing this objective, with respect to students' response-time (Hornke, 2000; Papamitsiou et al., 2014), self-regulatory factors (Hodges & Kim, 2010; Puzziferro, 2008), as well as non-cognitive students' perceptions (Chua, 2012; Shee & Wang, 2008). This section briefly reviews relevant literature for identifying the core factors to be further explored regarding their capacity to reason students' performance; i.e., all measures to be used in this study are carefully extracted from prior related work.

3.2.1. Response-time and behavioral factors for explaining students' performance

Scholars from the fields of Psychometrics and Intelligent Tutoring Systems have extensively explored time-related factors and investigated their appropriateness for explaining students' behavioral aspects during assessment (Lee, 2011). For example, response-time were associated to lack of test-taking motivation and guessing behavior, coded in time-driven students' test-taking effort (Chang, Plake, Kramer, & Lien, 2011; Setzer, Wise, van den Heuvel, & Ling, 2013; Wise & Kong, 2005). Results have shown that students tend to be more engaged in the beginning of a session, while guessing is more likely to occur when students are less engaged (Beck, 2005). No direct relation of response-time to test score was found, though (Chang et al., 2011; Hornke, 2000). The result from exploring the efficiency of using students' previous response-time for directly predicting the correctness of next actions and test score was statistically insignificant (Xiong et al., 2011). Exploring indirect dependencies between correctness of answers, goal expectancy, response-time and scores provided promising results (Papamitsiou et al., 2014).

The studies that associate efficient use of available time with performance are limited; time-management seems to reduce the need to game test completion strategies (Burrus, Jackson, Holtzman, Roberts, & Mandigo, 2013). Usually, lack of time-management can create a bad cycle for performance in the test, whereas, good time-management implies gaining control over the items and eventually, less stress. Thus, more items will probably be answered correctly, which is likely to be reflected on the test score. Indeed, high achieving students often exhibit strong time-management skills (Macan, 1990).

Time-management has been regarded as a self-regulation strategy. In general, self-regulatory strategies are acknowledged as significant determinants of students' overall performance and academic success (Broadbent & Poon, 2015). In fact, high achieving students tend to demonstrate more self-regulatory skills (Hodges & Kim, 2010; Kitsantas, 2002). Results from exploring the adoption of self-regulation for prediction of performance (Kitsantas, 2002; Sundre & Kitsantas, 2004) indicate that reviewing responses and efficiently using the available time are the strategies that (during testing) affect test score more, whereas goal expectations are more predictive of performance prior to taking the test (Kitsantas, 2002).

Moreover, students' perceived comprehensibility of the test items was explored regarding its effect on test result: since the test items are designed to measure knowledge acquisition, their validity and clarity are critical for students' response strategies (Doll & Torkzadeh, 1988; Fitzpatrick, 1983; Sun, Tsai, Finger, Chen, & Yeh, 2008). If students understand the items, they are more likely to be successful (Puzziferro, 2008). Clarity of content was proposed as a determinant of student satisfaction (Kurucay & Inan, 2017; Wang, 2003). The subjective perception of how well learning items meet the student's expectations for learning and supports success (Lo, 2010) may help refine our insight on the test result.

Furthermore, the students' performance during testing is strongly associated with their perception of task difficulty and on-task mental effort (Capa et al., 2008; Papamitsiou &

Economides, 2015). Wise & Kong (2005) introduced a method, the Response Time Effort (RTE), for measuring examinee test-taking effort based on item response time. It was found that RTE is a determinant of learners' performance in low-stakes test (Lee & Jia, 2014).

However, considering response-time, effort, self-regulation, or satisfaction exclusively can be unreliable to explain performance. More research is necessary on the interpretation of learning outcome, along with new methods that will offer fresh insight into the existing literature. The present study takes a different methodological approach by implementing supervised learning analyses, as well as configurational analysis.

3.3. First Study: Applying Supervised Learning algorithms for explaining performance

Recent technologies for data collection and machine learning could offer new insights into human learning (Beck & Woolf, 2000; Blikstein & Worsley, 2016). The purpose of machine learning (either in its supervised or unsupervised form) in learning analytics research is: 1) to learn statistical models (functions) out of observed learning events and learning outcomes, 2) generalise of future, similarly structured unobserved data. Suppose there is a data set containing learning events (observations) with measurements on different variables (called predictors) and their known learning output (class) labels. *If predictor values are obtained for new observations, could the classes those observations belong to be determined?* This is the problem of supervised classification: the task of assigning objects to one of several predefined classes. Each classification technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and the class label of the input data, and operates in two phases: the training phase and the testing phase (Alpaydin, 2010; Tan et al., 2005). In this study we examine how accurately students can be classified according to their assessment test scores using as predictors a set of logged interactions data.

3.3.1. Methodology

3.3.1.1. Research participants and data collection

In this study, data were collected from a total of 259 undergraduate students (108 males [41.7%] and 151 females [58.3%], aged 20-27 years old [M=22.6, SD=1.933, N=259]) at a European University. Twelve randomly generated groups of 20 to 25 students attended an assessment procedure for the Computers II course (related to databases, information system, introduction to e-commerce) at the University computer lab. For the assessment, we used 34 tasks (multiple-choice questions). Each task had two to four possible answers, but only one was the correct. During the design of the tasks, two instructors agreed on their difficulty (easy, medium, hard).

All students had to answer 10 tasks within 30 mins. The participants could skip or review the tasks and/or alter the submitted answer. For the test score computation, only the correct answers were considered, without penalizing the incorrect answers (i.e., without negative scores). Further, each task was weighted based on its difficulty level, and contributed differently to the overall

test score, ranging from 0.8 points (easy) to 1.2 points (medium) to 2 points (hard). In the case that students chose not to submit an answer to a task, they received a score of zero for this one.

The participation to the assessment procedure was optional. As an external motivation to increase the students' effort, we set that their score would participate up to 30% to their final grade. All participants signed an informed consent form prior to their participation, explaining to them the procedure and giving the right to researchers to use the data collected for research purposes. Students were aware that their answers were being tracked, but not their time-spent, because we wanted them to act spontaneously.

3.3.1.2. Data collection and measurements

Data were collected with the LAERS environment (Appendix A). According to the previous findings from literature, measures that either are commonly used in the field of learning analytics and acknowledged to satisfactorily explain students' performance (e.g., response-times) or have been found to predict achievement (Capa et al., 2008; Papamitsiou et al., 2014; Xiong et al., 2011) were computed from the logged interactions trace data. Specifically, these measures included: (a) Time to answer correctly (TTAC), (b) Time to answer wrongly (TTAW), (c) Idle Time (TIT), and (d) Effort. TTAC and TTAW are defined as the total time that students spend on viewing the assessment items and submitting the correct and wrong answers respectively. By definition, they indicate the respective response-time the students constantly aggregate on answering the assessment questions (Papamitsiou et al., 2014). TIT is the idle time the students aggregate on viewing each item, without submitting an answer (Appendix A). For the effort calculation, the Response Time Effort (RTE) measures the proportion of items which the students try to solve (solution behavior) instead of guessing the answers (Wise & Kong, 2005) (Appendix B). Moreover, before taking the assessment test, the students had to answer to a pre-test questionnaire that measured their goal-expectancy (GE), a measure of goal-orientation, introduced in Computer Based Assessment Acceptance Model (Terzis & Economides, 2011), and particularized to measure this self-regulation strategy in assessment procedures. We adopted three items that measure this construct. These items are measured in a 5 point Likert-like scale (1 = strongly disagree to 5 = strongly agree, Appendix C)

3.3.1.3. Feature Subset Selection

The initial raw log file contained a sample of the five attributes (features) to be used in this study. Our pre-experimental thoughts were that some of the attributes were "noisy"; i.e. contain signals not related to the target of classification. Therefore, we first attempted to remove spurious attributes using feature subset selection. Feature selection reduces the dimensionality of data by selecting only a subset of features (i.e., predictor variables) to create a model. Selection criteria usually involve the minimization of a specific measure of predictive error for models fit to different subsets. Algorithms search for a subset of predictors that optimally model measured responses, subject to constraints such as required or excluded features and the size of the subset. Note that the number of attributes to select is crucial in the analysis of the data. In this experiment,

we ranked the attributes from most to least informative. The attributes were ranked using the sequential feature selection method of MATLAB. This method has two components: a) an objective function, called the criterion, which the method seeks to minimize over all feasible feature subsets, and b) a sequential search algorithm, which adds or removes features from a candidate subset while evaluating the criterion.

3.3.1.4. Analysis Methods

In the present study, we explored 5 different advanced supervised learning techniques for classifying students based on their time-based characteristics (predictors) and according to their actual learning performance (class label). In particular, we tried Artificial Neural Networks (ANNs), Support Vector Machines (SVM), Naïve Bayes (NB), k-Nearest Neighbors (kNN) and the treeBagger method. These are some of the well-known classifiers used in the machine learning field, and the most common approaches explored with a plurality of different attributes in the learning analytics and educational data mining research domain.

Artificial Neural Networks (ANNs) are computational systems based on the structure, processing method and learning ability of the brain (Haykin, 1998). When performing classification analysis with an existing dataset, a commonly adapted approach, named holdout validation, is used to split the data into a larger set for training the ANN and a smaller set for testing the model. In this work, a Feed Forward neural network has been created and trained.

Support Vector Machines (SVM) is a supervised learning method for linear modeling. For classification purposes, nonlinear kernel functions are often used to transform the data into a feature space of a higher dimension than that of the input before attempting to separate them using a linear discriminator (Cortes & Vapnik, 1995). In this work, a third degree polynomial kernel function was employed.

Naïve Bayes (NB) are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the predictors within each class. During the training step, the method estimates the parameters of a probability distribution. Next, during the prediction step, the method computes the posterior probability of that sample belonging to each class, and classifies the test data accordingly (Tan et al., 2005).

k-Nearest Neighbors (kNN) is a non-parametric method used for classification. Given an unknown sample, a kNN classifier searches the pattern space for the k training samples that are closest to the unknown sample. kNN is based on the principle that the samples within a dataset will generally exist in close proximity to others that have similar properties (Altman, 1992).

Bagging Bagging (Bootstrap Aggregating) is an ensemble method that creates separate samples of the training dataset and creates a classifier for each sample. In fact, bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation does a plurality vote when predicting a class. The multiple

versions are formed by making bootstrap replicates of the learning set and using these replicates as new learning sets (Breiman, 1996).

3.3.1.5. Measures and Performance Criteria

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model, tabulated in a confusion matrix. Generally speaking, the (i,j) element in the confusion matrix is the number of samples whose known class label is class i and whose predicted class is j . The diagonal elements represent correctly classified observations. However, the confusion matrix is not convenient to compare the performance of different models. Accuracy is a single-value performance metric defined as the proportion of correct predictions to the total predictions. Further, the performance of a model can be expressed in terms of its error rate, which is given as the proportion of wrong prediction to the total predictions (Alpaydin, 2010; Tan et al., 2005). The errors committed by a classification model are divided into resubstitution errors (training errors) and generalization errors (test errors). The resubstitution error is the proportion of misclassified observations on the training set, whereas the test error is the expected prediction error on an independent set. A good model must have low resubstitution error as well as low test error (Mitchell, 1997; Tan et al., 2005).

A method commonly used to evaluate the performance of a classifier is cross validation. The k-fold cross validation method segments the data into k equal-sized partitions. This procedure is repeated n times so that each partition is used the same number of times for training and exactly once for testing. We used a stratified k=10-fold cross validation with n=100 iterations for estimating the misclassification (test) error (Alpaydin, 2010; Tan et al., 2005). Moreover, sensitivity analysis is a method for identifying the “cause-and-effect” relationship between the inputs and outputs of a prediction model. This method is often followed in machine learning techniques to rank the variables in terms of their importance according to the sensitivity measure (Mitchell, 1997). Finally, F-score (or F-measure) is a measure of a test's accuracy. It considers the precision and the recall of the test to compute the score. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant, while high recall means that an algorithm returned most of the relevant results (Alpaydin, 2010; Mitchell, 1997). The F-score can be interpreted as a weighted average of the precision and recall, where an F-score reaches its best value at 1 and worst score at 0 (Tan et al., 2005).

3.3.2. Results

Table 3-1 outlines the supervised learning algorithms (SLA) we applied on the input data, the number of classes being predicted (i.e., the different categories of students' performance results), the overall accuracy of the prediction (for training and testing respectively) together with the respective sample sizes (90% for training and 10% for testing for all SLA methods), and the tool used during the analysis.

Table 3-1. A summary of our experiment

SLA used	# of classes predicted	Sample size	Accuracy of prediction	Simulation tool used
ANNs, SVMs, NB, kNN, treeBagger	7-class	259 samples in total 233 for training 26 for testing	100% for training 76% for testing	MATLAB

3.3.2.1. Exploratory data analysis

Table 3-2 illustrates the variables (features) used to train and test the machine learning networks, as well as the range of their values. Table 3-3 illustrate the covariance matrix for all five input variables. As it can be seen, there are no strong correlations between the variables.

Table 3-2. Features used for training and testing

	Variable	Description	Type	Value Range
Temporal	TTAC	Time to answer correctly	Simple	≥0 (msec)
	TTAW	Time to answer wrongly	Simple	≥0 (msec)
	TIT	Idle time	Simple	≥0 (msec)
	RTE	Effort	Computed	0-1
Time-varying	LP*	Learning Performance	Simple	0-1.5
Self-reported	GE	Goal expectancy	Latent	0-5

*LP: target (output)-dependent variable

Table 3-3. Covariance matrix for all predictor variables

	1	2	3	4	5
1. TIT	1.000				
2. TTAC	-0.082	1.000			
3. TTAW	-0.313	0.357	1.000		
4. RTE	-0.353	0.056	0.259	1.000	
5. GE	0.128	0.098	0.055	0.564	1.000

TIT: Idle Time, TTAC: Time to answer correctly, TTAW: Time to answer wrongly, RTE: Effort, GE: Goal-expectancy

3.3.2.2. Classification results

In this study, we initially explored the previously described methods with an input dataset consisting of three predictors: TTAC, TTAW and GE. We chose to examine these variables based on the formerly reported results (Papamitsiou et al., 2014). Table 3-4 presents the performance results (resubstitution error, true test error, sensitivity, and F-score) for all the methods used to develop a classification model with 10% of the initial dataset as testing sample size.

Table 3-4. Performance metrics for cross-validation 10% with three features

Test Set Size	cvpartition = 10% (k-fold=10)				
Classifier	ANN	SVM	kNN	NB	ENS**
Resub Error	0.34	NaN	0.30	0.38	0.00
True Test Error*	0.24	0.27	0.28	0.24	0.24
Sensitivity	0.96	0.95	0.95	0.96	0.96
F-score	0.87	0.85	0.86	0.85	0.88

*True test error=cross-validation error, **ENS:ensembles of decision trees

These results demonstrate that all methods achieve high prediction performance, since the true test error varies from 0.24 (ENS method) to 0.28 (kNN method). Further to that, the sensitivity measure is close to 1 in most cases (0.95-0.96) and the F-score is also high (0.85-0.88). Moreover, from this table it becomes apparent that the ENS method provides better results compared to the other methods, while the kNN and NB methods also achieve satisfactory results.

Based on this finding, we examined how the highly performing methods (ENS, kNN and NB) change their output when applied to more input variables (predictors). We explored this question with two additional features: TIT and RTE. Table 3-5 illustrates the performance metrics for the ENS, kNN and NB methods with 4 (initially we added TIT) and 5 (finally we added RTE) features and testing sample set to 10% of the initial dataset.

Table 3-5. Performance metrics for test set size 10% with 4 and 5 features

Forward Feature Selection	'TTAW', 'TTAC', 'GE', 'RTE'			'TIT', 'TTAW', 'TTAC', 'RTE', 'GE'		
Classifier	kNN	NB	ENS	kNN	NB	ENS
Resub Error	0.30	0.37	0.00	0.30	0.36	0.00
True Test Error	0.28	0.28	0.28	0.24	0.32	0.24
Sensitivity	0.95	0.95	0.95	0.96	0.94	0.96
F-score	0.85	0.86	0.85	0.88	0.82	0.84

TIT: Idle Time, TTAC: Time to answer correctly, TTAW: Time to answer wrongly, RTE: Effort, GE: Goal-expectancy

ENS does not seem to be affected by the additional features, providing results similar to the previous ones (conducted using with three features only). On the contrary, the performance of the other two methods is slightly reduced when the number of predictors increases.

3.3.3. Discussion and Conclusions from the first study

In this study, we explored student-generated temporal trace data for classifying students' according to their learning performance score in an assessment testing procedure. Our goal was to investigate whether the temporal data (response-time, effort, idle-time) and the self-reported perceptions of goal-orientation could predict and explain the students' achievement level. The motivation for our experimentation was based on previous research studies that analyzed the same temporal parameters for explaining performance, using other statistical approaches (e.g., PLS-SEM) and reported significant results (Papamitsiou et al., 2014). During our experimentation, we applied 5 advanced SLA techniques.

Our findings verify formerly reported results (Belk, Germanakos, Fidas, & Samaras, 2014; Shih et al., 2008) regarding the capability of temporal data to represent, describe and model the students' behavior. In particular, our findings indicate that the time to answer correctly and the time to answer wrongly in combination with the goal expectancy could satisfactorily be used for classification of students in computer-based testing procedures. The low misclassification rates are indicative of the accuracy of the proposed method. Further to that, from tables 3-4 and 3-5 it becomes apparent that the ensemble learning (treeBagger) method provided the most accurate classification results compared to the other methods. However, an interesting finding that

requires more investigation is that most algorithms perform worse when two additional features are included in the analysis. We still have to explore why this is happening and whether these additional features are appropriate for classification purposes.

Based on the findings, we suggest that one can identify a set of functional temporal (or behavioral) factors/parameters that could constitute the core components of an assessment system's architecture. For example, TTAC, TTAW, TIT, RTE and GE are only indicative variables that could be embedded into a testing system in order to model test-takers' answering behavior and to guide adaptation and personalization of the assessment services. For example, such a service could be the recommendation of the next most appropriate task according to the student's behavior and the detected level of expertise (based on the corresponding timely predicted performance). In this case, the system should be "trained" in order to "recognize" and model its current users based on their temporal and behavioral data. Then, it should "choose" the appropriate task (among the collection of tasks from an item bank) that best corresponds to the needs and meets the abilities of the user, in order to improve the expected outcome. Finally, the system should inform the users about their progress and either suggest the selected task (as a CAT system) or allow the users to make their own choice of the next task (as a CBT system).

The approach suggested in this study was applied on a dataset collected during an assessment procedure in the context of mid-term exams. However, the nature of the data collected (time-based parameters) and the general-purpose methodology followed for the analysis of these data, render this approach replicable and/or transferable to other contexts, and eliminate the restriction of using it only during testing. The temporal factors are not contextualized to the LAERS assessment environment, but a similar tracker could be embedded in any adaptive learning system. For example, time-related parameters could be tracked to measure the duration of solving sub-activities or sub-tasks in the context of project-based learning, or to measure the duration of studying and exercising with learning modules during inquiry-based learning, along with the number of repeating the intermediate, facilitating steps (e.g. watch educational videos, view/use educational resources, participate in discussions, etc.).

As a next step, we are planning to deeper explore the patterns of these classes in terms of time-spent, i.e., which are the specific characteristics of the time-spent behavior of the examinee that belong to each one of the classes, and using alternative analytical methods.

3.4. Second Study: Explaining learning performance with fsQCA

In this study, we employed fuzzy set qualitative comparative analysis (fsQCA) (Ragin, 2008), a striking alternative to traditional variance-based approaches (Woodside, 2013). When fsQCA is applied together with complexity theory, researchers have the opportunity to gain deeper and richer perspectives on their data (Fiss, 2011; Pappas, Mikalef & Pavlou, 2017; Pappas, Giannakos, & Sampson, 2017; Woodside, 2013). In the technology enhanced learning context, fsQCA suits for explaining complex combinations and interdependencies between various forms of learning analytics and performance data, and can lead to interpreting the observed learning

outcome (Pappas, Giannakos, & Sampson, 2016, 2017; Sergis, Sampson, & Giannakos, 2017). Here, the detected asymmetries between the identified factors expand the results from previous studies (Papamitsiou et al., 2014) that elaborated on symmetric relationships and contribute to better explaining performance.

3.4.1. fsQCA in technology enhanced learning and learning analytics

In Technology Enhanced Learning (TEL), a typical approach to knowledge construction is to evaluate innovative technologies empirically with user studies. In most of the cases, such evaluation is driven by front-loaded research questions/hypotheses and concludes into the acceptance or rejection of the front-loaded assumptions. To do so, many studies in TEL analyse data using variance-based approaches such as analysis of variance (ANOVA) or multiple regression analysis (MRA). Using these methods, it is possible to examine net effects between variables, and conceptually each statistical test offers a single solution/model to explain the observed outcome. The key advantage of such methods is their ability to test hypotheses, whereby researchers use the method to test an assumption and obtain a single concrete assessment (e.g. mean, p-value, etc.).

Although learners with different characteristics may form different sub-groups within a sample, these sub-groups have the potential to be analysed in-depth and receive targeted and learner-centric design recommendations. For example, if in a sample we have performance data for participants, one possible configuration would be "learners with high response time, who spent a lot of time each day with system A and had taken similar courses in the past". Intuitively, when we consider the different configurations in our sample independently, it is likely that different configurations may lead to the same or to different outcomes. The inability of quantitative methods to account for such "relativist flexibility" in analysis has been a weakness that we often account for through qualitative data such as interviews.

Conversely, fsQCA has been designed to embrace the notion of configurations, and the fact that different ways of slicing the data may tell different stories. The main benefit of the technique is that it can identify multiple unique configurations that explain a large part of the sample. While variance-based approaches also explain parts of a sample, it is often the case that their models have a relatively low R^2 value, meaning that the model explains or predicts only a portion of the sample (Woodside, 2013). In such cases, it can be beneficial to use fsQCA to identify multiple configurations that jointly explain a much larger portion of the sample. As such, fsQCA explains certain parts of the sample that otherwise would had been considered as outliers. This is an important methodological difference: fsQCA can help us identify how to design learning technologies for all (Pappas et al., 2016; Pappas, Giannakos, & Sampson, 2017), unlike variance-based methods that test competing models to identify the fittest.

3.4.2. Conceptual Model and Research Propositions

Apparently, much of the research into response-time has endeavored to identify item factors and human factors that determine and affect the students' response-time strategies in

testing conditions (Hornke, 2000; Kahraman, Cuddy, & Clauser, 2013; Papamitsiou et al., 2014; Xiong et al., 2011). It was claimed that response-time should be treated as fixed predictor, because time-limit may affect students' performance (Wang & Hanson, 2005). Nonetheless, response-time only reflect the time-spent on individual items and do not tell the full story about how students complete a test.

To this end, self-regulation is expected to provide additional evidence regarding students' behavioral strategies. However, although reports of self-regulation in prior research have been found to be predictive of learning performance (Zimmerman & Pons, 1986), very little is known about the role of self-regulation in test performance, and thus, estimating whether students' test outcomes are influenced by the use of a self-regulatory strategy is still an open issue (Kitsantas, 2002). Among the key self-regulatory processes expected to affect test performance are goal expectations, self-monitoring (reviewing responses), and time-management (Macan, 1990; Sundre & Kitsantas, 2004).

Moreover, students' satisfaction from content has been acknowledged for reflecting the consistency between expected gain and the actual experience (Lo, 2010); more research is required in how perceived clarity of content influences performance.

This study posits that there is a synergy among self-regulation (goal expectations, time-management), response-time and satisfaction from content (perceived clarity of content) in predicting students' learning performance. Indeed, there is not one unique, optimal, configuration of such values. Instead, multiple and equally effective configurations of causal conditions exist, which may include different combinations of self-regulation, response-time and clarity of content. Depending on how they combine they may or may not explain students' high or medium/low performance. High performance refers to the presence of a condition, and medium/low to the absence of the condition. The absence is examined as the negation of a condition (i.e., not present), thus we examine the non-high performance, that is medium/low performance. This approach allows the identification of asymmetrical relations among the examined factors and the outcome.

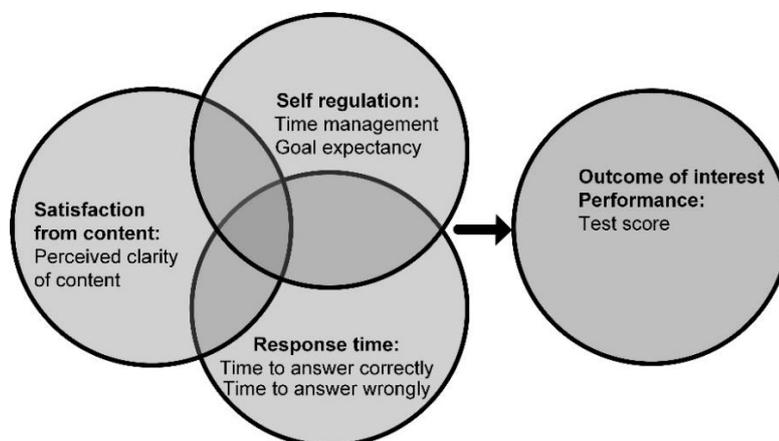


Figure 3-1. Venn diagram of the conceptual model.

To conceptualize these relationships, we propose a theoretical model (Figure 3-1) illustrating three constructs, their intersections, and the outcome of interest. On the left we present self-regulation (i.e., time management, goal expectancy), response-time (i.e., time to answer correctly and time to answer wrongly), and satisfaction from content (i.e., students' perception about the clarity of content). On the right, the outcome of interest is presented, that is students' performance on assessment tests. The overlapped areas represent possible combinations among factors, that is, areas that one factor may exist together with the other factors. Furthermore, to identify such patterns of factors in complex systems dedicated to educational measurement and assessment of learning (e.g., Computerized Adaptive Testing systems implementing complex Item Response Theory - IRT - solutions for large-scale high-stakes tests (Wainer, 2000)), formulating hypotheses, common in variance based methods that are framed as correlational expressions, does not allow for a holistic approach that will lead to the identification of multiple solutions.

Indeed, in configuration theory approaches, research propositions are formulated as causal recipes to capture the different combinations among factors, and theoretically specify which should be present or absent from the causal recipe (Doty, Glick, & Huber, 1993; Ordanini, Parasuraman, & Rubera, 2014).

The principle of *equifinality* is inherent in both complexity theory and configuration theory, based on which a result may be equally explained by alternative sets of causal conditions (Doty et al., 1993; Fiss, 2011). In a complex system, relations among factors (i.e., causes) are also complex and depending on how they combine, both high and low conditions of a certain factor may explain high scores of an outcome. These conditions may be combined in sufficient configurations to explain the outcome (Fiss, 2011; Woodside, 2013). For example, students may respond quickly and correctly because of a lucky guess, or because of high self-confidence due to high self-preparation to take the assessment test (Papamitsiou et al., 2014; Setzer et al., 2013). However, self-confident students who believe that they can perform well, tend to be more careful when they answer the test items: these students are more likely to persist longer in their efforts to accomplish the tasks successfully (higher response-time) than less self-reliant students (Tabak, Nguyen, Basuray, & Darrow, 2009). Similarly, students may respond slowly and correctly because of lower perception regarding the clarity of items' content or because of carefulness in time-management and self-regulation. Studies have shown that questions' content and students' test performance are indirectly associated with each other, mediated by response-time (Chang et al., 2011). Moreover, students who lack self-regulatory skills, tend to exhibit guessing behavior at the end of a session, not-trying to understand the test items and exhibiting low time-management (Beck, 2005; Eilam & Aharon, 2003). As configurations can include different combinations of the examined constructs, they lead to the following proposition:

Proposition 1: *No single configuration of self-regulation, response-time, and satisfaction from content is sufficient for explaining high performance; instead, multiple, equally effective configurations exist.*

Further, configuration theory proposes the principle of *causal asymmetry*, which means that, for an outcome to occur, the presence and absence of a causal condition depends on how this condition combines with the other conditions (Fiss, 2011; Pappas, Giannakos, & Sampson, 2017). A predictor variable may have an asymmetric relation with the outcome, which means that even if one variable is insufficient for the outcome to occur, it is still able to serve as a necessary condition for the outcome variable (Fiss, 2011; Woodside, 2013). For example, using students' previous response-time for prediction of correctness of their next actions and, consequently, their test performance provided only statistically insignificant results (Xiong et al., 2011). However, when the response-time were associated to self-regulatory strategies, e.g., goal expectations, the result regarding the prediction of test score was significantly improved (Papamitsiou et al., 2014). In addition, high goal-expectations exclusively do not imply high score; unless the students use the available time efficiently, they might achieve a low test score although they have been well-prepared. On the contrary, high time-management could lead to answering correctly those items that seem more clear and understandable, beyond the students' prior self-preparation to take the test. Hence, we form the following propositions:

Proposition 2: *Single conditions of self-regulation, response-time, and satisfaction from content can have opposite effects on performance, depending on how they combine with other conditions to form a solution.*

Proposition 3: *Configurations of self-regulation, response-time, and satisfaction from content for high performance are no mirror opposites of configurations for its negation (i.e., medium/low performance).*

3.4.3. Research methodology

3.4.3.1. Research participants and data collection

Data were collected with LAERS at a European University during a progress assessment test with 452 undergraduate students (211 males [46.7%] and 241 females [53.3%], aged 20-28 years old (M=21.18, SD=1.47, N=452)). The students attended the testing procedure for the Microeconomics II course (related to monopolistic competition, oligopoly, competitive strategy, and general equilibrium theory) at the University computer lab, for 75 min. each group.

For the assessment needs, 60 multiple choice tasks were used in total, distributed in 5 equivalent tests of 15 tasks each (some of the items were shared in more than two assessment tests). Each task had four possible answers, but only one was the correct, and corresponded to one of the first five levels of the factual, conceptual and procedural domains of the knowledge dimension according to the revised Bloom's taxonomy. We chose to evaluate students' knowledge by using tasks of these five levels only, due to the available time of the assessment test. Moreover, each task participated on the score computation, contributing in accordance to its level of difficulty. The difficulty of the tasks had been previously determined by the course instructor.

Before taking the assessment and right after the completion of the procedure, each participant had to answer to a pre-test and a post-test questionnaire that measure each student's

goal expectancy and time-management, and their perceived clarity of the items' content respectively. The participation to the procedure was mandatory. Similar to study 1, all participants signed an informed consent form prior to their participation, explaining to them the procedure and giving the right to researchers to use the data collected for research purposes. Students were aware that their answers were being tracked, but not their time-spent, because we wanted them to act spontaneously.

3.4.3.2. Measures

As stated in section 3.4.2, the identified set of factors to be included in the conceptual model consisted of actual variables (i.e., response-time) and latent behavioral factors and perceptions (i.e., self-regulation and satisfaction from content). The targeted outcome of interest was the students' learning performance, calculated for each student as: $LP = \sum_{i=1}^N d_i z_i$, where $z_i \in (0,1)$ is the correctness of the student's answer on item i , and d_i is the difficulty of the item.

More precisely, in this study, the response-time variables and the goal-expectancy factor were the same with study 1. Furthermore, grading self-regulation, another latent factor, i.e., time-management, was measured via pre-test questionnaire. Time-management (TM) (Macan, 1990) reflects students' perception of their own planning abilities and it is associated with students' exercising conscious control over the amount of time spent on items during assessment. For measuring students' satisfaction from content, the latent factor for perceived clarity of items (CONT) (Terzis & Economides, 2011) was measured via post-test questionnaire. CONT stores information related to whether the students considered the items to be clear, understandable and relative to the course's content, allowing students to evaluate the quality of the items and to self-reflect on their understandings of this content. All items from the questionnaires are measured in a 7 point Likert-like scale (1 = not at all to 7 = very much, Appendix C). Table 3-6 summarizes all factors included in the conceptual model with a short description, their type and value range.

Table 3-6. List of factors considered in the conceptual model

Factor	Description	Type	Value
Pre-test (Self-regulation)			
Time management (TM)	perception of exercising conscious control over the amount of time spent	Latent – measured via questionnaire	1-7
Goal expectancy (GE)	perception of preparation and motivation to succeed	Latent – measured via questionnaire	1-7
During test (response-time)			
Time to answer correctly (TTAC)	response-time aggregated on submitting correct answers	Simple – computed from actual data	≥ 0 (msec)
Time to answer wrongly (TTAW)	response-time aggregated on submitting wrong answers	Simple – computed from actual data	≥ 0 (msec)
Post-test (satisfaction from content)			
Perceived clarity of the test tasks (CONT)	perception about the clearness of test tasks	Latent – measured via questionnaire	1-7
Learning Performance (LP)	The test result	Computed	1-10

3.4.3.3. *FsQCA*

3.4.1 Data calibration. FsQCA analysis was performed based on Pappas et al. (2017). When performing fsQCA, the researcher starts by defining the outcome of interest and the independent measures. Next all measures must be recoded into fuzzy sets, that is receiving values from 0 to 1. This process is called data calibration, and defines the extent to which cases are members of a certain group (or set) (Ragin, 2008). Every case of a dataset has a distinct place as determined by its fuzzy-set membership. A value of 1 means that a case is a full member of a set, and a value of 0 means that a case is fully non-member of the set. A value of 0.5 is exactly in the middle, thus a case is both a member and a non-member of the set, creating the intermediate set membership.

Data calibration can be done either directly or indirectly. Direct calibration means that three qualitative thresholds need to be chosen, which define the level of membership in the fuzzy set for every case. On the other hand, indirect calibration means that the measurements need to be rescaled following qualitative assessments. Either method can be followed, as it depends on one's substantive knowledge of the data and the underlying theory (Ragin, 2008; Rihoux & Ragin, 2009). Data calibration is critical, because disparities in the calibration may lead to disparities in the outcome, thus cases in the dataset should be transformed into membership scores following a well-documented and qualitatively justified manner. The direct method of setting three values, corresponding to full-set membership, full-set non-membership, and intermediate-set membership is recommended (Ragin, 2008).

Next, the question is how to choose the three thresholds. The simplest way is to choose the values of 1, 0.5, and 0. For instance, in a 7-point Likert scale, the values 7, 4, and 1 would be calibrated into 1, 0.5, and 0 respectively, with the rest (6, 5, 3, 2) following accordingly. For 7-point Likert scales multiple studies suggest that the values of 6, 4, and 2 should be used as thresholds (Pappas, Mikalef & Pavlou, 2017; Ordanini et al., 2014; Pappas, Giannakos, & Sampson, 2017). Also, measures can be calibrated by using percentiles. In this case, the following percentiles can be set as the full-set membership, intermediate-set membership, and full-set non-membership, 80%, 50%, and 20%, respectively (Pappas, Giannakos, & Sampson, 2017). However, as it is up to the researcher to choose the three thresholds, these values can be changed accordingly. In this study, since data are skewed to the right, data calibration is done based on percentiles. Calibration based on the survey scale might lead to less meaningful results, producing a single solution with all the conditions identified as necessary (Pappas, Papavlasopoulou, Giannakos, & Sampson, 2017; Pappas, Giannakos, & Sampson, 2017).

Once running the analysis, fsQCA creates a truth table of 2^k rows, where k represents the number of outcome predictors (i.e., independent variables) and each row represents every possible combination. For instance, a truth table between five variables (i.e., conditions) would provide thirty-two possible logical combinations. For each combination, fsQCA computes the minimum membership value (i.e., the degree to which a case supports the specific combination). The threshold of 0.5 is used to identify the combinations that are acceptably supported by the

cases. Thus, all combinations that are not supported by at least one case with membership larger than the threshold of 0.5 are automatically removed from further analysis.

Next, the truth table must be sorted based on frequency and consistency (Ragin, 2008). Frequency refers to the number of observations for each possible combination, and consistency refers to “the degree to which cases correspond to the set-theoretic relationships expressed in a solution” (Fiss, 2011). Since fsQCA computes all logical combinations, many combinations will have a frequency of zero. It is important to set a frequency cut-off point which will ensure that a minimum number of empirical observations is obtained for the assessment of subset relationships. A higher frequency threshold means that every combination will refer to more cases in the sample, but it will reduce the percentage (i.e., coverage) of the sample explained by the solutions. On the other hand, a small frequency threshold will increase the coverage of the sample, although each combination will refer to fewer cases in the sample. For small and medium-sized samples, a cut-off point of 1 is appropriate, but for larger samples (e.g., 150 or more cases), the cut-off point should be set higher (Ragin, 2008). The researcher can decide if a larger cut-off point should be set for very large datasets. Low frequency combinations are removed from further analysis and the truth table must be sorted based on “raw consistency.”

A consistency threshold needs to be set, with the minimum recommended value being 0.75 (Rihoux & Ragin, 2009). A good indication for choosing this threshold is to identify big changes in the consistency of each combination. For instance, a combination may have a consistency of 0.82 and the next may have 0.79. Although both values are above the recommended threshold of 0.75, this is an indication of what the consistency threshold should be. In any case, it is up to the researcher to choose the exact threshold. A low consistency threshold may produce more necessary conditions, reducing type II errors (i.e., false negatives), but increasing type I errors (i.e., false positives), and vice versa (Dul, 2016). The last step is to insert the value of 1 or 0 in the column with the outcome variable, depending on the consistency threshold that has been chosen. Combinations with consistency higher than the threshold will get the value of 1, otherwise, 0.

3.4.2 Obtain the solution sets. Following the sorting of the truth table, fsQCA computes the following three sets of solutions: complex, parsimonious, and intermediate; “solution” is a combination of conditions that is supported by a high number of cases, and the rule “the combination leads to the outcome” is consistent. The complex solution presents all possible combinations of conditions when traditional logical operations are applied. The number of complex solutions can be large, including configurations with several terms, their interpretation is difficult and often impractical (Mendel & Korjani, 2012). Thus, they are simplified automatically into parsimonious and intermediate solutions.

The parsimonious solution is a simplified version of the complex solution and presents the most important conditions that cannot be left out from any solution. These are called “core conditions” (Fiss, 2011) and are identified automatically by fsQCA. Finally, the intermediate solution is computed when performing counterfactual analysis on the complex and parsimonious

solutions (Pappas, Giannakos, & Sampson, 2017; Ragin, 2008). FsQCA uses simplifying assumptions to compute the parsimonious and intermediate solutions, and if needed the researcher may employ more assumptions, regarding the connection between each causal condition and the outcome, based on theoretical or substantive knowledge (Fiss, 2011; Ragin, Drass, & Davey, 2006). The intermediate solution is a part of the complex solution and includes the parsimonious solution. Conditions that are part of the intermediate solution but of the parsimonious solution are called “peripheral conditions” (Fiss, 2011). A more detailed description of the steps in counterfactual analysis is provided by (Mendel & Korjani, 2012).

3.4.3 Interpretation of the solutions. FsQCA computes the complex and parsimonious solutions regardless of any simplifying assumptions employed by the researcher, while the intermediate solution depends directly on these assumptions. For better interpreting the results, combining the parsimonious and intermediate solutions is recommended. A table that will include both core and peripheral conditions should be created (Fiss, 2011; Pappas, Giannakos, & Sampson, 2017). To do this, the researcher should identify the conditions of the parsimonious solution in the intermediate solution. This leads to a combined solution, which will include all core and peripheral conditions, thus helping in the interpretation of the findings. Further, to improve the visualization of the results, the presence of a condition is presented with a black circle (●), the absence with a crossed-out circle (⊗), and the “do not care” condition with a blank space. The distinction between core and peripheral is done by using large and small circles, respectively. The overall solution consistency and the overall solution coverage are presented. Consistency measures the degree to which a subset relationship has been approximated, and overall coverage describes the extent to which the outcome is explained by the different configurations, and is comparable with the R-square reported on regression-based methods (Rihoux & Ragin, 2009; Woodside, 2013).

3.4.4. Results

The results for high performance and medium/low performance are shown in Tables 3-7 and 3-8 respectively. Both tables also present consistency values for the overall solution and for each solution separately. All values are higher than the recommended threshold (> 0.75) (Ragin, 2008). An overall solution coverage of .77 and .79 suggests that the five solutions account for a substantial proportion of high performance and medium/low performance respectively. FsQCA also estimates the empirical relevance of every solution, by calculating raw and unique coverage. The raw coverage describes the amount of the outcome explained by a specific alternative solution, while the unique coverage describes the amount of the outcome that is exclusively explained by a specific alternative solution. Solutions for high performance explain a large amount of users’ performance, ranging from 22% to 40% of cases associated with the outcome (Table 3-3). Similarly, solutions for medium/low performance explain a vast amount of the absence of performance: 22% - 58% of cases associated with the outcome (Table 3-4).

For high performance (Table 3-7), solutions A1-A5 present combinations for which the different factors may be present or absent depending on how they combine with each other.

- *Solution A1*: Students achieved high performance when they did not spend a lot of time to answer the questions, either the answers were correct or wrong, but they did good time management. This solution explains the behaviour of 22% of the high performing students.
- *Solutions A2, A3 and A4*: These solutions show that spending a lot of time to find the correct answers is important for a high performance, but not enough. This is an intuitive finding as it shows that students who give all their focus only in finding the correct answer will achieve high performance. However, it is interesting to note that this happens when students were neither sufficiently prepared for the progress assessment nor they believe that they can manage their time properly (solution A2). Also, this happens when students have good time management skills, but they had not understood the questions very well (solution A3). In fact, according to the results, solution A2 explains 28% of the high performing students, whereas solution A3 explains a bigger sub-population of the high performing students, since row coverage is 36%. Nonetheless, spending enough time, to find the correct answer will lead to high performance even if the questions are unclear or not so relative to the syllabus. On the other hand, if the students understand well the questions, then they can achieve high test score, even when spending a lot of time both on the questions answered wrongly and correctly (Solution A4). This solution explains 26% of the high performing cases.
- *Solution A5*: Finally, the students can achieve high performance regardless of how much time they spend to answer the questions: they have set high goal expectations, they have high time management skills, and they have a good understanding of the questions. This solution explains the larger part of the cases of high performing students (40%).

Table 3-7. Configurations for high performance

Solutions for high performance		A1	A2	A3	A4	A5
Configuration						
Response-time	Time to answer correctly	⊗	●	●	●	
	Time to answer wrongly	⊗			•	
Self-regulation	Goal Expectancy		⊗			●
	Time Management	●	⊗	•		•
Satisfaction	Clarity of Content			⊗	•	●
Consistency		.82	.86	.90	.88	.88
Raw Coverage		.22	.28	.36	.26	.40
Unique Coverage		.04	.08	.01	.01	.01
Overall Solution Consistency		.81				
Overall Solution Coverage		.77				

Note: Black circles (●) indicate the presence of a condition, and circles with “x” (⊗) indicate its absence. All circles indicate core conditions. Blank spaces indicate don’t care conditions.

Next, Table 3-8 presents the solutions for not achieving a high performance, that is achieving medium/low performance. The findings show that the solutions that explain medium/low performance are not perfect opposites of the solutions that explain high performance. Specifically:

- *Solutions B1 and B2:* Students that do not have high goal expectancy will have a low or medium performance, when they do not spend a lot of time to answer the questions (either correctly or wrongly) (Solution B1), or when they have low time management skills, leading them to not using efficiently the available time (Solution B2). These solutions explain 34% and 58% of the cases of medium/low performing students.
- *Solution B3:* Students that have spent a lot of time to answer the questions wrongly, and they are not well prepared, they will have a medium or low performance, even if they have understood the questions. This behaviour is observed in 33% of the cases of medium/low performing students.
- *Solution B4:* Students that give correct answers fast, but spend a lot of time to questions that they do not know the answer, will have a medium or low performance even if they have high goal expectancy and believe that they can use their time properly. This solution explains 22% of the cases for medium/low performing students.
- *Solution B5:* Students who perceive the questions as clear and relative to the syllabus, and spent a lot of time in answering them (both correctly and wrongly), will not achieve a high performance unless they have a good time management. This finding highlights the importance of time management in high performances. However, this behaviour for medium/low performing students is not very common (17% of the cases).

Table 3-8. Configurations for medium/low performance

Solutions for medium/low performance						
	Configuration	B1	B2	B3	B4	B5
Response-time	Time to answer correctly	⊗	⊗		⊗	●
	Time to answer wrongly	⊗		●	●	●
Self-regulation	Goal Expectancy	⊗	⊗	⊗	●	
	Time Management		⊗		●	⊗
Satisfaction	Clarity of Content			●		●
Consistency		.82	.85	.85	.89	.85
Raw Coverage		.34	.58	.33	.22	.17
Unique Coverage		.03	.16	.02	.07	.01
Overall Solution Consistency						.81
Overall Solution Coverage						.79

Note: Black circles (●) indicate the presence of a condition, and circles with “x” (⊗) indicate its absence. All circles indicate core conditions. Blank spaces indicate don’t care conditions.

The results provide support for all three propositions. In detail, multiple configurations lead to high performance, verifying equifinality (Proposition 1). Also, the results provide configurations that explain performance in which conditions may be either present or absent, depending on how they combine with each other, verifying the existence of causal asymmetry (Proposition 2). Finally, the findings support proposition 3, i.e., configurations that explain high performance are not the exact opposites of those explaining medium/low performance.

3.4.5. Discussion and Conclusions from the second study

Understanding the factors that affect assessment outcome, as well as their interrelationships, could contribute to the sufficient explanation and interpretation of the obtained high or medium/low performance. Previous research identified critical factors that affect the performance (e.g., response-time, self-regulatory strategies, non-cognitive perceptions of satisfaction from the content), but failed to reveal asymmetric relationships between these factors, mostly due to the variance-based analysis methods they employed for exploring the data (Kitsantas, 2002; Papamitsiou et al., 2014; Puzziferro, 2008). This study focuses on compiling students' response-time allocated to answer correctly or wrongly, their self-regulation, as well as their satisfaction from the content, targeting at explaining high or medium/low performance achieved on the assessment test. For this purpose, fuzzy set qualitative comparative analysis (fsQCA) was applied for exploring multiple configurations of causal conditions which may include different combinations of goal-expectations, time-management, response-time and clarity of content. Data were collected during a progress assessment test with 452 undergraduate students from a European University. The results provided several interesting findings.

Firstly, as seen from Table 3-7, highly performing students who believe that they have good time-management skills, ended up spending little time in giving answers, and not spending their time in a meaningless way (solution A1). This finding is in agreement with (Macan, 1990), who supported that high achieving students often exhibit strong time-management skills. Moreover, according to solutions A2, A3, and A4, students who aggregated non-neglectable response-time for correct answers, although they were neither sufficiently prepared for the progress assessment nor they managed their time efficiently (solution A2), however, they finally achieved a high score in the test. This finding contradicts with (Papamitsiou et al., 2014) which identified that poorly-prepared students (i.e., scoring low in goal-expectations) achieve low scores, and indicates that regardless of preparation and time-management, the students may still get high grades if they engage more on answering the questions. This contradiction, however, might be due to the fact that in (Papamitsiou et al., 2014) the authors investigated only symmetric solutions. Thus, the current approach sheds more light into the interrelationships between preparation, response-time and test-score. This also happens when students did a good time management but they had not understood all questions very well (solution A3). This means that probably, these students managed their time efficiently in order to answer correctly on those items that were clear to them, and although they struggled to understand the rest of the items, they

finally delivered more correct answers. Nonetheless, spending enough time, in an appropriate manner, to find the correct solution will lead to high performance even if the questions are unclear or not so relative to the syllabus.

On the other hand, if the students understand well the questions they can achieve high performance even when spending a lot of time for all questions (both the ones answered wrongly and correctly) (Solution A4). This is an expected finding, since perceiving the test content as comprehensible does not necessarily mean that it is trivial or easy to answer, and as such, wrong answers are likely to occur, but they are not the dominant ones, leading to an overall high performance.

These two findings are interesting and innovative in terms of the rather limited literature on the issue of the effects of time-management along with content comprehensibility on the assessment test outcome. To the best of our knowledge, this is the first study to explore this combinational/conditional interrelationship, and extends previous work (Lo, 2010), which focuses solely on the effect of content on performance. It should be noted that spending a lot of time to answer correctly is a present as a core factor in these solutions, highlighting its importance in achieving high performance. This is in full agreement with (Papamitsiou et al., 2014) and provides additional evidence regarding the role of aggregated response-time to interpreting performance.

Finally, solution A5 verifies that well-prepared students are quite likely to perform well and get high scores, regardless of how much time they need to answer any of questions (Hodges & Kim, 2010).

Regarding the results for medium/low performance, solution B1 and B2 claim that poorly prepared students answer the questions relatively quick are not expected to have a high performance, which seems intuitive and further complies with the literature (Eilam & Aharon, 2003). However, comparing these findings with solution A1 from Table 3-7, highlights how important is time management in achieving high performance.

Moreover, solution B3 describes students that had a good understanding of the questions, but because they were not well prepared they did not know the answers, thus making them to use a lot of time on answering wrongly, leading to medium or low performance. There are two interesting clues about this finding: (a) it is in contrast to previous results that comprehensibility of content is directly reflected on performance (Wang, 2003), and (b) it is surprising from a slightly different point of view: as seen from table 3-8, this solution explains 58% of the cases of medium/low performing students. The surprising thing is that, these students admitted that the test questions were clear, comprehensible and related to the course's content, yet they were neither prepared to answer them, nor they tried to guess the answers. It would be really valuable to explore a measure of guessing regarding this sub-group of medium/low performing students in order to identify/evaluate their guessing intentions.

Finally, one of the most important implications of this chapter is related to how learning analytics researchers and practitioners can utilize the fsQCA method to make sense of diverse analytics and take design decisions for various user groups (Pappas, Giannakos, Jaccheri, & Sampson, 2017; Sergis et al., 2017). Future studies should combine fsQCA with variance-based techniques to gain a deeper insight into the learning analytics, and combine both methods towards extending current theories and practices as well as developing new ones. As this study is among the first to employ fsQCA in learning analytics context (Pappas et al., 2016; Pappas, Giannakos, & Sampson, 2017), further innovative research is needed to identify complex and important configurations that reveal the full potential of this analysis. Future studies should incorporate data from various learning activities and modalities, making-sense of complex learning interactions and offering a holistic understanding of the potential of this data analysis technique in technology enhanced learning and analytics.

Chapter 4 : The “testing analytics” paradigm

*“Not everything that counts can be counted
And not everything that can be counted counts”*

Albert Einstein

Explaining learners’ performance in computer-based tests using learning analytics

4.1. Introduction

In every educational setting, assessment is used to measure learners’ achievements and progress in learning processes (Farrell & Rushby, 2016; Shute & Rahimi, 2017). e-Assessment is the use of information technologies to automate and facilitate assessment and feedback processes. The benefits from e-assessment for learning are not questionable and their exploitation is in the epicentre of the technology enhanced learning research interest (Gikandi, Morrow, & Davis, 2011; Shute & Rahimi, 2017). Nonetheless, testing procedures should not be neglected; especially in higher education, computer-based testing is a typical and popular format for the evaluation of knowledge acquisition and for the measurement of learners’ progress (Adesope, Trevisan, & Sundararajan, 2017; Arnold, 2016). Besides, grades are still required at the end of the course and are critical to the overall learners’ academic success. In a recent meta-analysis of the gains in learning and retention from practising tests, Adesope et al. (2017) concluded that tests work way better than *any other* strategy they compared it with, and increase learners’ achievement.

A testing procedure can be regarded as formative or summative depending on whether the objective is to assist learning or to assess achievement, respectively (Harlen & James, 1997). Formative assessment in the form of frequent testing is usually treated by the teachers worldwide as a diagnostic tool to regularly mark the learners’ progress during the course (Challis, 2005; Gikandi et al., 2011). It is used to monitor and assess learning by providing quick feedback, and to guide modifications in instruction, until a desired level of the course objectives has been achieved. Moreover, summative tests use grades to describe what the learners have achieved (learning validation) and to certify whether academic objectives have been reached (learning accreditation) after a period of time (Gikandi et al., 2011; Shute & Rahimi, 2017).

Standardized tests on the one hand and Computerized Adaptive Testing (CAT) on the other, have both been extensively applied in higher education to successfully resolve measurement-related issues (Adesope et al., 2017). And, although in standardized tests a fixed set of items is administered to all examinees and graded in the same manner for everyone, in CAT, different examinees receive quite different tests. In this case, the computer-based test is tailored to the examinee’s ability level: the choice of the next item to be administered depends on the correctness of the examinee’s response to the latest item administered. CAT is usually used in large-scale high-stakes/low-stakes exams (high/low impact on learners’ course grade respectively), identification of learners who need

specialized support, formative and summative assessment, etc. (Weiss, 2004; Weiss & Kingsbury, 1984). The main approaches in CAT are Classical Test Theory (CTT) and Item Response Theory (IRT), with the first one being recently superseded by the second one. Unlike CTT analyses that employ item difficulty and item discrimination estimators for the adaptive administration of the items, IRT models probabilistically select optimal items on the basis of information rather than difficulty (Birnbaum, 1968; Wainer, 2000).

Issues related to tests such as scoring, bias, reliability, validity, etc. have attracted increased attention and scholarly debates (Adesope et al., 2017; Duckworth & Yeager, 2015). However, current testing procedures are restricted to only providing a score, failing to look deeper in which learner/learning factors cause this score. There is a need to shift the justification of test scores beyond the observed performance, and to use them to support claims about learners' achievements that generally are not self-evident. This study focuses on determining a set of factors that adequately and evidently reason learners' responses and performance in computer-based tests, as well as on modeling causal relationships between them, using a learning analytics approach.

4.1.1. Motivation of the research

The existing testing methods have provided well-established measurement and assessment formats. However, assessment tests have received comprehensive criticism; (a) they fail to represent the types of authentic situations that learners are likely to deal with in their everyday life (Nortvedt, 2014), (b) they use items that predominantly test lower levels of cognitive skills and do not assess higher-level thinking skills (Zlatović, Balaban, & Kermek, 2015), (c) chasing grades may distract students from deeper learning (Wolsey, 2008), yet good grades do not necessarily reflect mastery (Davis, 1999), are conducive to cheating and put academic honesty in question (Arnold, 2016), and (d) IRT models are not easy to build due to highly demanding pre-requisites (e.g., calibration of the item pool and exposure control algorithms to prevent overuse of a few items, extensive psychometric expertise, and large pilot samples) (Hambleton & Jones, 1993).

Apparently, these testing formats lead to a superficial assessment of learners' skills and knowledge, failing to look deeper in the causation of learners' responses, and encounter drawbacks and limitations. The continuous formative tests could provide sufficient information about the attained learners' progress, knowledge development and possible misconceptions. Yet, when it comes to summative tests, it becomes harder for both teachers and learners to clearly interpret the grades the learners score on a single test. There is a need for enhancing the existing testing formats with information-rich and concise indices that advance our understanding on the interpretation of the test outcome. Gaining in-depth insight of the learners' interactions and seeking for reasoning of their responses in summative computer-based testing contexts (fixed or adaptive), is a demand. This means that it is necessary to dive into the raw learner data, extract the significant information from it, select the most descriptive and informative factors that affect performance, convert them to comprehensible indices, and model causal relationships between these factors.

It has been argued that computational techniques from the emerging learning analytics research have the potential to efficiently capture and analyse data from e-assessment environments (Knight, Shum, & Littleton, 2014). The present study has its origins in the learning analytics research and is inspired from solid Psychometrics approaches. It introduces “*testing analytics*” and investigates their capacity to elucidate and model learners’ responses and performance in formal summative, fixed or adaptive, computer-based testing, in terms of learning analytics.

In accordance with learning analytics, testing analytics is a repetitive, context-aware process that monitors and records detailed data related to the testing context, interprets and maps the real current state of the data, organizes them, uses them, and continues by predicting their future state. The motivation behind our research is to contribute towards exploring the premise of testing analytics and elaborate on the benefits from their adoption.

4.1.2. Objectives and research questions

The main goal of this study is to shift the discussion about testing outcome interpretation beyond the performance itself, to a deeper understanding of the underlying factors that justify learners’ responses. As such, the research objective is to seek for and determine a set of factors that sufficiently explain learners’ responses and test score, as well as to model causal relationships between them in summative fixed and adaptive computer-based testing conditions. Towards that objective, an analytics-driven method is proposed. The research question that guided the study is:

RQ: (a) *which learning analytics factors explain sufficiently the multiple aspects of learners’ interactions with the assessment tasks? (b) How can we model the cause-effect relationships between these factors towards explaining the variance in the learners’ performance?*

4.2. Related work and the “testing analytics” paradigm

Most of the recent research on educational testing aspects is devoted to refining technical facets of CAT, such as comparison of different item-selection methods (e.g., He, Diao, & Hauser, 2014; Yao, 2012), item pool construction (Lee & Dodd, 2012), and test stopping rules (Wang, Chang, & Boughton, 2013). When it comes to modeling and explaining learners’ responses during testing most of the relevant research adopts a learners’ response-time perspective, i.e., exploring the amounts of time the learners allocate on test items (W. J. van der Linden, 2009). Overall, four dimensions are explored: *response-time, motivational factors, response strategies, and emotions*.

(a) Response-time: Scholars from the fields of Psychometrics and Intelligent Tutoring Systems have extensively examined time-related factors, and investigated their appropriateness for explaining learners’ behavioral aspects and refining the testing procedures accordingly (e.g., guiding item selection) (for more evidence, see Schnipke & Scrams (2002), and Lee & Chen (2011)). It was claimed that response-time should be treated as a fixed predictor, because time-limit may affect learners’ score (Wang & Hanson, 2005). Issues related to pacing behavior were also considered, and it was found that test speededness (i.e., the distribution of the learners’ total time on the test (W. J. van der Linden, 2011)) introduces a severe threat to the validity of interpretations based on test scores

(Kahraman et al., 2013). Although response-time reflect the impact of the test items on learners' behavior, however, no direct relation of response-time to test score was found (Chang et al., 2011). The efficiency of using learners' previous response-time for prediction of correctness of their next actions and their test score was examined without statistically significant results. It was suggested that considering other learners' attributes might provide more concise predictions (Xiong et al., 2011). The investigation of causal dependencies between motivational factors (e.g., goal-expectancy), response-times, correctness of answers and test score, provided encouraging findings (Papamitsiou et al., 2014).

(b) Motivational factors: Response-time was associated to learners' lack of motivation and to guessing behavior to represent learners' test-taking effort and explore its impact on test scores (Cao & Stokes, 2007; Setzer et al., 2013; Silm, Must, & Täht, 2013; Wise & Kong, 2005). It was claimed that modeling effort in terms of response-time could contribute to understanding the dynamics of learners' test-taking motivation (Wise & Kong, 2005). Using a response-time formulation for effort, learners' guessing behavior was found to be strongly related to *how seriously* they take the assessment test: they had higher achievement expectation in tests that were in line with their own learning goals, resulting in reduced rapid-guessing behavior (Setzer et al., 2013). Moreover, performance in low-stakes tests was influenced by two test-taking effort parameters: the number of items the learners attempted to solve, and the time they allocated to solve each item (Silm et al., 2013). Other motivational human factors have been examined as well, regarding their effect on test score. Computer self-efficacy and training satisfaction (H. Lu, Hu, Gao, & Kinshuk, 2016) were both found to positively affect test score. The effect of incentives and motivational instruction on engagement and test score was examined, as well (Jalava, Joensen, & Pellas, 2015; Liu, Rios, & Borden, 2015; Nikou & Economides, 2017b). It was found that instruction and extrinsic incentives improved both learners' self-reported test-taking motivation and test scores (Jalava et al., 2015; O. L. Liu et al., 2015).

(c) Response-strategies: Other studies investigated the effect of response strategies and learning strategies on test scores (Schnipke & Scrams, 1997; Zlatović et al., 2015), and how they are associated with personality traits (Papamitsiou & Economides, 2017) and demographics (Lee & Haberman, 2016). The stimulated learning strategies (deep or surface, measured with appropriate questionnaire instruments) influenced the achievement level the learners exhibited in online testing (Zlatović et al., 2015). It was also shown that learners' personality traits were strongly associated with the achievement results in computer-based tests, and specifically, extraversion, agreeableness and conscientiousness had statistically significant indirect impact on learners' response-times and achievement level (Papamitsiou & Economides, 2017). Moreover, considerable research focused on detecting and preventing "*gaming the system*" learner strategies, i.e., when learners exploit properties of the system in their attempt to succeed, instead of using their knowledge to answer correctly (Baker, Corbett, Koedinger, & Wagner, 2004). It was proposed that a better distributional schema of learner response-times could be employed for distinguishing learners who engage in problem solving from learners who engage in gaming the system (Shih et al., 2008). Gaming behavior has been associated

with boredom and confusion (Rodrigo et al., 2007). No correlation to gaming has been found regarding motivation and anxiety (R. Baker, Corbett, Roll, & Koedinger, 2008).

(d) *Emotions*: Test anxiety (H. Lu et al., 2016; Ortner & Caspers, 2011; Putwain, Daly, Chamberlain, & Sadreddini, 2016), fatigue, attention, boredom and achievement emotions (Daniels & Gierl, 2017; Moridis & Economides, 2009b; Wainer, 2000) were also examined with respect to the learners' scores in assessment tests. Findings confirmed the expected negative effect of test anxiety on test scores; higher worry, as well as emotional and physiological arousal predicted lower examination scores (Ortner & Caspers, 2011; Putwain et al., 2016). The negative effect of anxiety was present even after the end of the exam (Daniels & Gierl, 2017).

Apparently, many studies have attempted to shed light to how item factors and human factors affect the testing outcome. However, a model that holistically explains the cause-effect relationships of these factors and the importance and contribution of each one of them on the interpretation of the test score is missing. Nevertheless, none of these studies has directly infused the testing procedures with critical information to foster the instructors' awareness beyond the test score, to the interpretation of this result.

4.2.1. The “testing analytics” paradigm and expected contributions

The primary purpose of this study is to dive into the learner-generated raw data and extract a set of variables that can be directly computed/measured and transformed into informative factors that reason learners' actions and diagnose test score in formal summative computer-based testing conditions. This context-aware process that monitors and records detailed data related to the testing context, interprets and maps the real state of the data, organizes them, uses them, and continues by predicting their future state, is defined as *testing analytics*. The candidate factors to be considered in the testing analytics paradigm emerged from the analysis of previous empirical research findings, are grounded in the learning analytics domain, and are selected according to the following criteria:

1. *The factor has its origin in the learning analytics research*: it is commonly used to model and assess learning in terms of analytics. For example, response-times, frequencies of interactions, aggregated counters, etc. have been previously explored and provided encouraging results (for additional evidence, see Kovanović, Gašević, Joksimović, Hatala, & Adesope, 2015; Papamitsiou, Economides, Pappas, & Giannakos, 2018).
2. *The factor provides fine-grained information about learners' cognitive state*: it has been previously found to statistically significantly explain the variance in learners' achievement level (i.e., knowledge) in learning or assessment activities. As such, it can be inspired or originate from the well-established IRT models that are commonly agreed to predict satisfactorily the learners' progress, needs and achievements in the cognitive domain. The parameters in IRT's 3PL model include the items' difficulty, their discrimination ability and a guessing parameter (Birnbaum, 1968); we consider learners' time-spent according to the item's difficulty, their effort that incorporates guessing (Wise & Kong, 2005), and their

accumulated response-times to answer correctly/wrongly for discriminating their ability (for additional evidence, see Lin, Shen, & Chi, 2016; Papamitsiou et al., 2014; Shih et al., 2008; van der Linden, Entink, & Fox, 2010; T. Wang & Hanson, 2005).

3. *The factor reflects the impact of motivational factors*: it contains learners' evaluation of their own motivation, expectations, and satisfaction. According to the review of related work, the motivational factors employed in this study are self-efficacy and goal-expectancy. In addition, perceived clarity of items is considered because it reflects the learners' evaluation on the test items. Other factors could be explored as well. For example, all self-regulated learning constructs (e.g., help-seeking, effort-regulation) could potentially be incorporated in the model (for additional evidence, see Daniels & Gierl, 2017; Gutman & Schoon, 2013; Lu et al., 2016; McMillan & Hearn, 2008).
4. *The factor reflects the learner's affective state*: it has been established that mood and emotions play a crucial role concerning human reasoning and learning, and as such, factors that reflect the learner's affective state should also be considered. As seen from the related work, test anxiety (Lu et al., 2016), fatigue, attention, boredom (Moridis & Economides, 2009b, 2009a) are all determinants of test scores.
5. *The factor is enhanced with a notion of "dynamics"*: it might be a time-varying factor (e.g., effort) or a temporal measure (e.g., response-times), signaling changes in learners' behavior and allowing for temporal predictions during testing (Kleinberg, 2016; Whiting, 2015). For example, Wise and Kong (2005) modelled effort in terms of response-time aiming at understanding the dynamics of learners' test-taking motivation.

The present study focuses on extracting analytics from testing procedures, as a complementary technology to explain and reason learners' score. In essence, the contribution of this work stems from the holistic consideration of a set of informative factors that correspond to learners' actions during summative computer-based testing into a single cause-effect model that will next simplify the interpretation of the attained score. The findings from this study are expected to advance our knowledge on why learners act the way they do in computer-based testing contexts and how their scores reflect both their learning and actions. Analytics enhanced testing systems, illustrating learners' manipulations of the testing items (as quick progress check or detailed diagnostics), can lead instructors to a greater awareness of the attained outcome.

It should be noted that the affective factors have not been included in the current version of the suggested model: it has been argued that current methods for real-time affect/mood detection can only be used as diagnostic means (Moridis & Economides, 2009b, 2009a). The assessment of learners' affects, intentions, and beliefs claims heterogeneous sensing and grading modalities that analyze text, eye gaze, speech, handwriting, gesture, as well as neurophysiological markers (Blikstein & Worsley, 2016); these tasks, although would provide high quality and granularity information about learners' treatment of the testing items, however, they are highly complex, are usually measured with

the use of specialized equipment (e.g., facial expressions recognition software), and are not commonly employed in simple computer-based testing procedures.

4.3. Research model and hypotheses

Driven by previous research results, the research hypotheses on the causal relationships between the considered factors are outlined below:

4.3.1. The aggregated response-times

Time to answer correctly (TTAC) and Time to answer wrongly (TTAW) are defined as the total time that learners spend on viewing the assessment items and submitting the correct and wrong answers respectively. By definition, they indicate the respective response-time the learners constantly aggregate on answering the test items, according to the correctness of the response (Papamitsiou et al., 2014). Based on previous results (Papamitsiou et al., 2014, 2016, 2018; Shih et al., 2008; Wang & Hanson, 2005), we assume that learners who answer correctly many items will aggregate more time to answer correctly and are more likely to score higher. Conversely, learners who answer incorrectly many items will aggregate more time to the respective factor and are more likely to score lower, both in fixed or adaptive tests. Thus, we hypothesized that:

***H1a:** Time to answer correctly will have a positive effect on performance.*

***H1b:** Time to answer wrongly will have a negative effect on performance.*

4.3.2. Goal expectancy

Goal-expectancy (GE) reflects the learners' dispositions regarding their achievement expectations in the assessment (Terzis & Economides, 2011), and has two dimensions: (a) how satisfied they are with their preparation for the assessment, and (b) their desirable level of success. Before taking the assessment, the learners report on how prepared they are to take the test, and set a goal regarding the percentage of correct answers that they perceive as satisfying performance. In other words, goal-expectancy is a measure of learners' motivation and interest to take the test. The decision to employ goal-expectancy instead of other constructs of goal-orientation (e.g., the 2x2 achievement goals framework (Elliot & McGregor, 2001)) was based on the fact that this construct was particularized and validated strictly in assessment-oriented procedures, and as such, it better fits the context of this study. In line with previously reported findings (Papamitsiou et al., 2014; McMillan & Hearn, 2008; Zimmerman, 2002), we believe that, in both fixed and adaptive testing contexts, well-prepared and motivated learners (i.e. who score high in goal-expectancy) will answer more items correctly; these learners will have fewer wrong answers, and consequently, they will accumulate less time on items that finally will answer wrong. Therefore:

***H2a:** Goal expectancy will have a positive effect on time to answer correctly.*

***H2b:** Goal expectancy will have a negative effect on time to answer wrongly.*

In addition, the highly motivated learners are expected to devote more time-spent on items of increased complexity or difficulty, trying to master them in fixed procedures. This is hypothesized

because the learners who are well prepared and want to succeed in the test, are more likely to find the hard items challenging for their abilities, and they will probably engage more on their solution. Similarly, in adaptive testing, answering correctly on the administered items implies advanced knowledge mastery, sufficient preparation and increased goal expectations, and results in higher performance. In adaptive tests, the higher the learners' performance is, the higher the complexity of the next item administered to the learner. As such:

H2c: Goal expectancy will have a positive effect on time to answer hard.

It should be noted that no hypotheses can be stated concerning the other levels of item difficulty: most of the learners are expected to try to answer (all or most of) items of medium and low difficulty, despite their preparation and achievement expectations.

4.3.3. Self-efficacy

Self-efficacy (SE) has been defined as one's belief in one's ability to succeed in specific situations or accomplish a task (Bandura, 2006; Zimmerman, Bandura, & Martinez-Pons, 1992). It is a determinant of motivation and performance and a measure of self-confidence (Schunk, 1991, 1995). Perceived self-efficacy plays a major role in how the subject approaches goals, tasks, and challenges. Learners who score high in self-efficacy – that is, those who believe they can perform well – are more likely to view difficult tasks as something to be mastered rather than something to be avoided, and are more likely to make efforts to complete these tasks, and to persist longer in their efforts. According to previous findings (Puzziferro, 2008; Tabak et al., 2009), the self-confident learners who believe that they can efficiently complete tasks, are more likely to be more motivated and have higher goal-expectations from themselves, regarding the test results. These highly self-confident learners are more likely to submit more correct answers, aggregating more time on TTAC and less time on TTAW. On the contrary, learners who doubt their efficiency are more possible to answer wrongly on most items. Therefore, we assumed that:

H3a: Self-efficacy will have a positive effect on goal-expectancy.

H3b: Self-efficacy will have a positive effect on time to answer correctly.

H3c: Self-efficacy will have a negative effect on time to answer wrongly.

4.3.4. Effort

Effort is “the motivational state commonly understood to mean trying hard or being involved in a task. Effort is increased when the subject tries harder, when there are incentives to perform well, or when the task is important or difficult” (Humphreys & Revelle, 1984, p. 158). Thus, effort is about how much engaged the learners are in completing the tasks. Previous research revealed that the learners' performance during testing is strongly associated with their perception of task difficulty and on-task mental effort (Papamitsiou & Economides, 2015). Accordingly, we believe that the more engaged the learner remains (higher effort), the more possible it is to achieve a higher score in both fixed and adaptive testing procedures. Moreover, it is reasonable to assume that dealing with difficult

and demanding items requires more effort to process them and formulate/choose an answer. Therefore:

H4a: *Effort will have a positive effect on performance.*

H4b: *Effort will have a positive effect on time to answer hard.*

In addition, and according to previous research (Papamitsiou & Economides, 2015; Capa et al., 2008; Setzer et al., 2013; Wise & Kong, 2005), we believe that motivated learners will exhibit higher effort on achieving a higher score compared to less prepared ones, because their goal-orientation and/or self-efficacy is higher. Low-prepared or less self-confident learners are expected to persist less in their efforts, both in fixed and in adaptive conditions. Thus:

H2d: *Goal expectancy will have a positive effect on effort.*

H3d: *Self-efficacy will have a positive effect on effort.*

4.3.5. Perceptions about the clarity of test items

Clarity of content was proposed as a determinant of learner satisfaction (Kurucay & Inan, 2017; Wang, 2003). Learners' perceived clarity of the test items was explored regarding its effect on test result: since the test items are designed to measure knowledge acquisition, their validity and clarity are critical for learners' response strategies (Doll & Torzadeh, 1988; Fitzpatrick, 1983; Sun et al., 2008). If learners understand the items, they are more likely to be successful (Puzziferro, 2008). Research findings indicate a positive effect of clarity of content on learning performance (Papamitsiou et al., 2018). This subjective perception of how well the test items meet the learners' expectations for learning and support success (Lo, 2010) might help refine our insight on the test result. Accordingly, we assume that learners' beliefs about the clarity of the (fixed or adaptive) test items, i.e., whether they considered these items to be concise and coherent to the course's content, will be reflected on the score.

H5: *Perception of item clarity will have a positive effect on performance.*

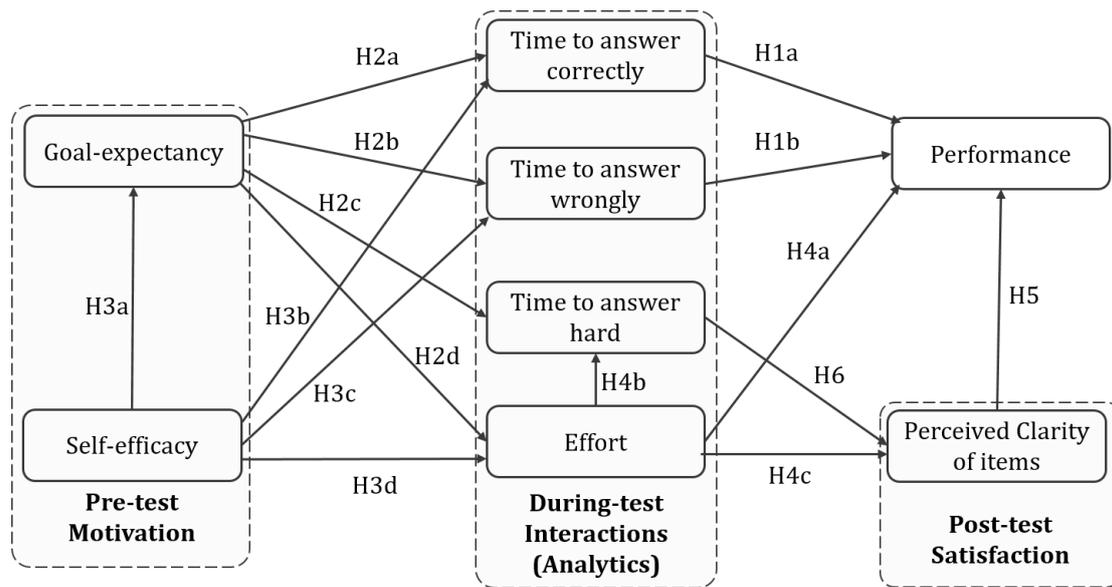
Moreover, when the learners exhibit high effort to answer the items, either these items are more likely to be perceived as perfectly clear and understandable, but their complexity requires increased effort exertion (i.e., hypothesis H4b), or to be perceived as less clear. Thus, the difficult items is possible to be perceived as comprehensible (despite their increased demands), whilst, effort exertion is also expected to be indicative of the learners' perceived clarity of the items' content: the higher the effort, the less understandable the item. As such:

H6: *Time to answer hard will have a positive effect on perceived clarity of items.*

H4c: *Effort will have a negative effect on perceived clarity of items.*

Figure 4-1 illustrates the causal relationships among the factors considered and hypothesized in testing analytics.

Figure 4-1. Overall research model and factor relationships with hypotheses in summative fixed and adaptive tests



4.4. Methodology

4.4.1. Research participants and study design

Data were collected from a total of 259 undergraduate students (108 males [41.7%] and 151 females [58.3%], aged 20-27 years old [$M=22.6$, $SD=1.933$, $N=259$]) at a European University. The students took the computer-based mid-term assessment test for the Computers II course (related to introduction to databases, information systems and e-commerce) at the University computer lab, for 60 minutes. The assessment test consisted of two phases: a fixed (30 minutes maximum) and next an adaptive test (no participant exceeded 30 minutes).

For the purposes of the fixed phase, 34 multiple choice items were used in total. For the needs of the adaptive phase, an item-bank of 52 pre-calibrated multiple choice items was constructed. The adaptation algorithm is a configuration of the Measurement Decision Theory (Rudner, 2003), and is briefly described in Appendix B. A minimum of 8 and a maximum of 12 items were used to classify the students based on their diagnosed mastery level. In both phases, each item had four possible answers, but only one was the correct. During the design of the tests, two instructors agreed on the difficulty (easy, medium, hard) of all 86 items'– instead of using, for example, the 3PL IRT's β parameter (Birnbaum, 1968), that would unnecessarily increase the complexity of the research model. For the score computation, only the correct answers were considered, without penalizing the incorrect answers (i.e., without negative scores). Further, each item was weighted based on its difficulty level, and contributed differently to the overall test score, ranging from 0.8 points (easy) to 1.2 points (medium) to 2 points (hard). In the case that students chose not to submit an answer to an item (fixed phase), they received a score of zero for this one.

Before taking the exam and right after the completion of the procedure, the students had to answer to the pre-test and post-test questionnaires that measure their goal-expectancy, and self-

efficacy, and their perceived clarity of the test items' respectively. Perceived clarity of items was measured twice, once at the end of each phase. The participation to the procedure was optional. As an additional motivation to ensure students' overall effort, their average score on this two-phase test was set to contribute to up to 30% of their final course grade. Prior to the test, all students signed an informed consent form that explained to them the procedure and was giving the right to researchers to use the data collected for research purposes. Students were aware that their answers were being tracked, but not their response-times, because we wanted them to act spontaneously.

4.4.2. Data collection and measurements

Data were collected with an web-based self-assessment environment (Appendix A). According to the criteria defined in Section 2.1 and the research hypotheses stated in Section 3, measures commonly used in the field of learning analytics (e.g., response-times) (Papamitsiou et al., 2018; Joksimović, Gašević, Loughin, Kovanović, & Hatala, 2015; Kizilcec, Pérez-Sanagustín, & Maldonado, 2017; Kovanović et al., 2015) were computed from the logged interactions trace data. Table 4-1 summarizes these factors, as well as a short description their type and value range.

Table 4-1. List of factors considered in testing analytics

Factor	Full Name	Description	Type	Value
Pre-test				
SE	Self-efficacy	Confidence in student's own ability to complete tasks and achieve results	Latent – measured via questionnaire	1-7
GE	Goal-expectancy	Student's perception of preparation and motivation to succeed	Latent – measured via questionnaire	1-7
During test				
TTAC	Time to answer correctly	The response-time a student aggregates on submitting correct answers	Simple – computed from actual data	≥0 (msec)
TTAW	Time to answer wrongly	The response-time a student aggregates on submitting wrong answers	Simple – computed from actual data	≥0 (msec)
TTAH	Time to answer hard items	The response-time a student aggregates on higher difficulty items	Simple – computed from actual data	≥0 (msec)
RTE	Response Time Effort	When s student exhibits solution behavior – a measure of engagement	Composite – computed from actual data	0-1
Post-test				
CONT	Perceived clarity of the test items	Student's considerations about the comprehensibility of test items	Latent – measured via questionnaire	1-7
TS	Performance (Test score)	The test result – the score the student gets	Computed	1-10

For the effort calculation, the Response Time Effort (RTE) measures the proportion of items which the students try to solve (solution behavior) instead of guessing the answers (Wise & Kong, 2005) (Appendix B). For measuring students' goal expectancy (GE), self-efficacy (SE) and to determine their perceived clarity of the items (CONT), we adopted items from formerly validates instruments (i.e., Terzis and Economides (2011), Bandura (2006), Appendix C). All

items were measured in a 7 point Likert-like scale (1 = strongly disagree to 7 = strongly agree).

Each students' score (TS) is calculated as: $TS = \sum_{i=1}^N d_i z_i$, according to the correctness of the student's answer on item i , $z_i \in \{0,1\}$ and the difficulty of the item, d_i .

4.4.3. Data analysis

For addressing the research question, the construction of a path diagram that contains the structural and measurement model was conducted with the Partial least-squares (PLS) analysis technique (Chin, 1998; Tenenhaus, Vinzi, Chatelin, & Lauro, 2005). Estimating and directly testing theoretically proposed chains of cause and effect (i.e., causal dependencies) between latent variables, is between the method's core objectives. Our decision to use PLS instead of an ordinary least-squares regression method (e.g. Hierarchical Linear Modelling) was based on our aim to reduce the complex constructs to a smaller set of uncorrelated components and perform least-squares regression on them, instead of on the original data. In PLS the sample size has to be a) 10 times larger than the number of items for the most complex construct, and b) 10 times the largest number of independent variables impact a dependent variable (Chin, 1998). In our model, all complex predictors have three items (see Appendix C), and the largest number of independent variables impacting a dependent variable is four (TTAC, TTAW, RTE, CONT to TS). Thus, our sample (259) surpasses the required value of 40.

4.4.4. Measures and Evaluation Criteria

In PLS, the items' factor loadings on the corresponded constructs have to be higher than 0.7 (Chin, 1998). The construct validity is confirmed by obtaining convergent – discriminant validity. Convergent validity is carried out by Average Variance Extracted (AVE) and has to be higher than 0.5 and the AVE's squared root of each variable has to be higher than its correlations with the other constructs (Barclay, Higgins, & Thompson, 1995; Fornell & Larcker, 1981). Cronbach's α and composite reliability are used to confirm reliability of the measurement model, and they both have to be higher than 0.7 (Tenenhaus et al., 2005).

The structural model evaluates the relationship between exogenous and endogenous latent variables by examining the variance measured (R^2). R^2 values of 0.67, 0.33, and 0.19 are substantial, moderate, and weak, respectively (Chin, 1998). The quality of path model can be evaluated by the Stone-Geisser's Q^2 value (Geisser, 1974; Stone, 1974), an evaluation criterion for the cross-validated predictive relevance of the PLS path model. The Q^2 statistic measures the predictive relevance of the model by reproducing the observed values by the model itself. A Q^2 greater than 0 means the model has predictive relevance; Q^2 statistic less than 0 mean that the model lacks predictive relevance. Finally, a bootstrap procedure evaluates the significance of the path coefficients (β value) and total effects, by calculating t-values. For the measurement and the structural model we used SmartPLS 3.2.

4.5. Results

4.5.1. Convergent validity - Discriminant validity

The results support the measurement model. Table 4-2 displays the construct items' reliabilities (Cronbach's α , composite reliability), AVE and factor loadings and confirms convergent validity for the latent constructs.

Tables 4-3 and 4-4 present the variables' correlation matrix for the fixed and the adaptive phase respectively. In these tables, the diagonal elements are the square root of the AVE of a construct. Discriminant validity is also confirmed according to the Fornell-Larcker criterion (Fornell & Larcker, 1981), i.e., the AVE of each construct is higher than the construct's highest squared correlation with any other construct.

Table 4-2. Results for the Latent Constructs of the Measurement Model

Construct Items	Factor Loadings (>0.7)^a	Cronbach's α (>0.7)^a	Composite Reliability (>0.7)^a	Average Variance Extracted (>0.5)^a
GE		0.832	0.900	0.751
GE1	0.802			
GE2	0.896			
GE3	0.898			
SE		0.718	0.842	0.640
SE1	0.836			
SE2	0.828			
SE3	0.732			
CONT(f)^b		0.771	0.867	0.686
CONT1(f)	0.852			
CONT2(f)	0.870			
CONT3(f)	0.759			
CONT(a)^b		0.704	0.834	0.627
CONT1(a)	0.848			
CONT2(a)	0.810			
CONT3(a)	0.721			

^a Indicates an acceptable level of reliability and validity, ^b Indicates fixed/adaptive procedure
GE: Goal-expectancy, **SE:** Self-efficacy, **CONT:** Perceived clarity of the test items

Table 4-3. Measurement Model – fixed testing

	1	2	3	4	5	6	7	8
1. GE	0.87							
2. SE	0.57	0.80						
3. TTAC	0.38	0.32	1.00					
4. TTAW	-0.36	-0.41	-0.10	1.00				
5. TTAH	0.25	0.19	0.81	0.27	1.00			
6. RTE	0.27	0.21	0.27	-0.23	0.26	1.00		
7. CONT	0.26	0.14	0.17	-0.24	0.13	0.11	0.83	
8. TS	0.57	0.51	0.47	-0.73	0.28	0.37	0.36	1.00

GE: Goal-expectancy, **SE:** Self-efficacy, **TTAC:** Time to answer correctly, **TTAW:** Time to answer wrongly, **TTAH:** Time to answer hard items, **RTE:** Response Time Effort, **CONT:** Perceived clarity of test items, **TS:** Performance (test score)

Table 4-4. Measurement Model – adaptive testing

	1	2	3	4	5	6	7	8
1. GE	0.87							
2. SE	0.56	0.80						
3. TTAC	0.37	0.23	1.00					
4. TTAW	-0.21	-0.07	-0.08	1.00				
5. TTAH	0.14	0.12	0.68	0.26	1.00			
6. RTE	0.41	0.25	0.67	-0.57	0.15	1.00		
7. CONT	0.39	0.22	0.28	-0.15	-0.02	0.35	0.79	
8. TS	0.60	0.32	0.57	-0.59	0.02	0.72	0.44	1.00

GE: Goal-expectancy, SE: Self-efficacy, TTAC: Time to answer correctly, TTAW: Time to answer wrongly, TTAH: Time to answer hard items, RTE: Response Time Effort, CONT: Perceived clarity of test items, TS: Performance (test score)

4.5.2. Testing hypotheses

A bootstrap procedure with 3000 resamples was used to test the statistical significance of the path coefficients (β value) in the model. Table 4-5 summarizes the results for the hypotheses in both contexts.

Table 4-5. Hypothesis testing results

Hypothesis	Fixed				Adaptive			
	Path	β	t	P	Path	β	t	P
H1a	TTAC → TS	0.351*	9.749	0.000	TTAC → TS	0.343*	5.816	0.000
H1b	TTAW → TS	-0.639*	21.031	0.000	TTAW → TS	-0.425*	7.751	0.000
H2a	GE → TTAC	0.286*	4.032	0.000	GE → TTAC	0.346*	5.752	0.000
H2b	GE → TTAW	-0.197*	2.805	0.005	GE → TTAW	-0.245*	3.163	0.002
H2c	GE → TTAH	0.194*	3.214	0.001	GE → TTAH	0.191*	2.522	0.003
H2d	GE → RTE	0.215*	2.978	0.003	GE → RTE	0.397*	5.913	0.000
H3a	SE → GE	0.568*	12.403	0.000	SE → GE	0.563*	10.800	0.000
H3b	SE → TTAC	0.158*	2.361	0.019	SE → TTAC	0.038	0.549	0.583
H3c	SE → TTAW	-0.297*	4.362	0.000	SE → TTAW	0.071	1.031	0.303
H3d	SE → RTE	0.091	1.253	0.211	SE → RTE	0.026	0.398	0.691
H4a	RTE → TS	0.120*	3.354	0.001	RTE → TS	0.179*	2.404	0.017
H4b	RTE → TTAH	0.207*	3.003	0.003	RTE → TTAH	0.109	1.351	0.177
H4c	RTE → CONT	0.080	1.347	0.179	RTE → CONT	0.361*	6.611	0.000
H5	CONT → TS	0.135*	3.575	0.000	CONT → TS	0.211*	5.637	0.000
H6	TTAH → CONT	0.114	1.677	0.094	TTAH → CONT	-0.073	1.368	0.172

* $p < 0.05$.

GE: Goal-expectancy, SE: Self-efficacy, TTAC: Time to answer correctly, TTAW: Time to answer wrongly, TTAH: Time to answer hard items, RTE: Response Time Effort, CONT: Perceived clarity of test items, TS: Performance (test score)

As seen from table 4-5, hypotheses **H1a**, **H1b**, **H4a** and **H5** regarding the factors that directly affect test score (i.e., response-times and effort, both corresponding to the learners' active engagement with the testing items during the examination), are strongly supported in both testing contexts. Hypotheses **H2** and **H3** about the motivational factors' effects on response-times

and effort, are partially supported in both settings. Specifically, **H2a**, **H2b** and **H2d** are strongly confirmed in both procedures, but **H2c** (concerning the effect of goal-expectancy on the time allocated on answering hard items) is supported only in the fixed testing conditions. From hypothesis **H3**, only **H3a** is strongly confirmed in both set-ups. The remaining sub-hypothesis regarding the impact of self-efficacy on response-times (**H3b**, **H3c**) are supported only in the fixed testing procedure, whereas the hypothesized effect of this factor on effort (**H3d**) was not supported. In addition, the results supported the assumption for hypothesis **H4b** (weakly in adaptive testing), however, the opposite results for **H4c** (strongly in adaptive testing and weakly in fixed testing) was confirmed. Finally, for hypothesis **H6**, the cause-effect relationship was found statistically insignificant and no concrete conclusion can be drawn for this factor.

4.5.3. Overall Model Fit

According to these results, the suggested model explains almost the 73% of the variance in fixed test score and 67% of the variance in adaptive test score, which is slightly lower than the first one, but still statistically significant. The cross-validated predictive relevance of the model was confirmed in both cases ($Q_{fixed}^2=0.708$ and $Q_{adaptive}^2=0.635$). Table 4-6 synopsizes the total effects of the selected factors, as well as the variance (R^2) and cross-validated predictive relevance (Q^2) explained by the proposed model.

From the analysis indirect effects were discovered as well. In table 4-6 it is shown that goal expectancy has strong indirect effects on the test score in fixed test ($\beta=0.258$, $t=5.164$) and in adaptive test ($\beta=0.322$, $t=5.368$) respectively. Similarly, self-efficacy indirectly and strongly affects both test scores in fixed ($\beta=0.403$, $t=9.373$) as well as in adaptive test ($\beta=0.171$, $t=4.097$).

Table 4-6. R^2 , Q^2 and Direct, Indirect and Total effects

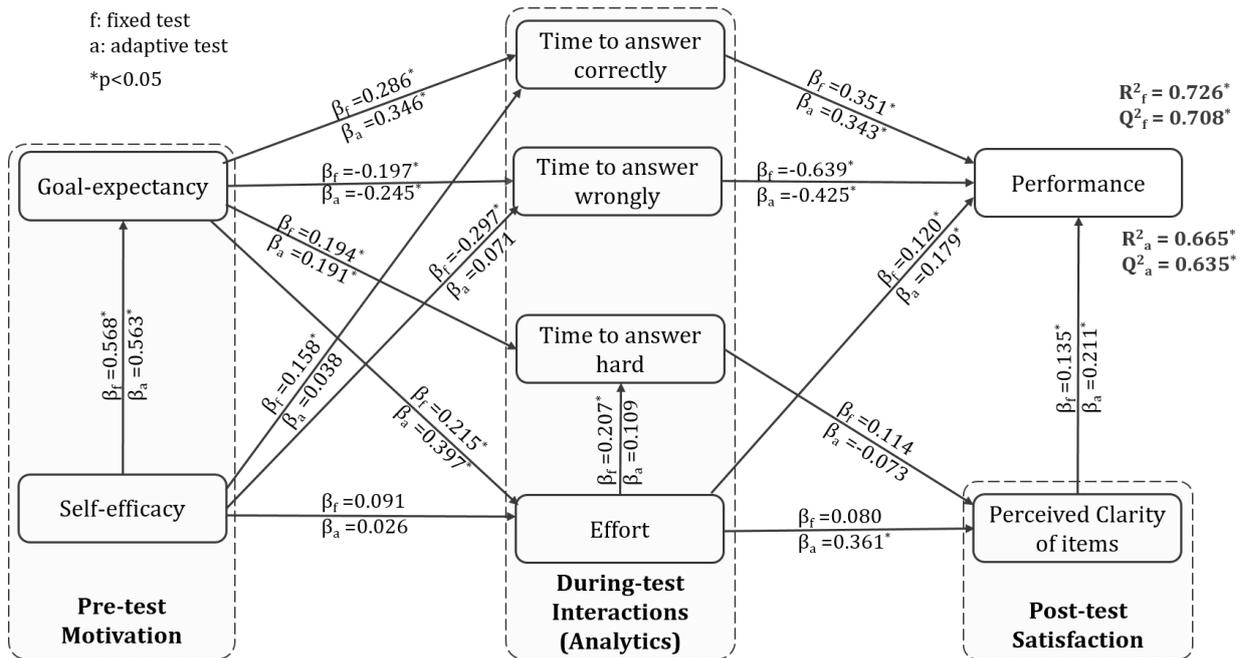
	Endogenous	R^2	Q^2	Exogenous	Direct effect	Indirect effect	t	Total effect
Fixed test	TS	0.726	0.708	TTAC	0.351		9.749	0.351*
				TTAW	-0.639		21.031	-0.639*
				TTAH		0.015	1.320	0.015
				RTE	0.120	0.014	3.599	0.134*
				GE		0.258	5.164	0.258*
				SE		0.403	9.373	0.403*
				CONT	0.135		3.575	0.135*
Adaptive Test	TS	0.665	0.635	TTAC	0.343		5.816	0.343*
				TTAW	-0.425		7.751	-0.425*
				TTAH		-0.015	1.259	-0.015
				RTE	0.179	0.074	3.478	0.253*
				GE		0.322	5.368	0.322*
				SE		0.171	4.097	0.171*
				CONT	0.211		5.637	0.211*

* $p < 0.05$

GE: Goal-expectancy, **SE:** Self-efficacy, **TTAC:** Time to answer correctly, **TTAW:** Time to answer wrongly, **TTAH:** Time to answer hard items, **RTE:** Response Time Effort, **CONT:** Perceived clarity of test items, **TS:** Performance (test score)

The measurement results are summarized in Figure 4-2.

Figure 4-2. Path coefficients of the research model, overall variance explained (R^2) for test score and cross-validated predictive relevance (Q^2).



4.6. Discussion: exploring the premise of testing analytics

The main objective of this study – expressed as the RQ – was to dive into the collected learner-generated raw data and extract a set of variables that can be transformed into the most informative factors for reasoning learners' actions and scores in formal summative computer-based testing settings in higher education. The demonstrated approach adopted a data-driven analytics perspective to shed light to the causation of learners' actions, to successfully diagnose their test score, and to develop the corresponding causal model, accordingly. The results outlined in Table 4-6 support that the overall prediction precision and the cross-validated prediction relevance of testing analytics are higher than 72% in fixed tests and higher than 66% in adaptive procedures, and are statistically significant. The data analysis revealed some interesting findings.

First, both accumulated time-spent on answering correctly and wrongly are strong direct determinants of learners' test scores. As seen from Table 4-5, response-time aggregated to answer correctly and to answer wrongly have significant direct positive effect ($\beta = 0.351$, $t = 9.749$) and significant direct negative effect ($\beta = -0.639$, $t = 21.031$) on fixed test score respectively. Similarly, the effects of time-spent aggregated to correct ($\beta = 0.343$, $t = 5.816$) and to wrong answers ($\beta = -0.425$, $t = 7.751$) on adaptive test score are also strong, supporting hypothesis **H1**. The statistically significant effects of these factors provide empirical evidence that contradicts Chang's et al. (2011) claims about the absence of relationship between response-time and score, and reinforce Wang's and Hanson's (2005) case that response-times should be treated as fixed predictors rather than random variables in the measurement models. Monitoring changes in these factors can timely indicate the learners' progress on the test at its early stages.

Second, the motivational factors measured prior to taking the test, i.e., goal expectancy and self-efficacy, are strong indirect determinants of learners' score in both testing conditions (Table 4-6). This finding aligns with Liu et al. (2015), adds evidence to prior claims regarding the effect of motivational factors on academic achievement (Schunk, 1991, 1995) and explains *why* motivated learners are more likely to score higher in summative tests: they are well-prepared and self-confident. Learners who trust their own ability to efficiently complete tasks, tend to set higher achievement expectations and to exhibit higher goal-oriented behavior (Terzis & Economides, 2011). This is verified by the strong direct positive effect that self-efficacy has on goal expectancy ($\beta=0.568$, $t=12.403$ in fixed test; $\beta=0.563$, $t=10.800$ in adaptive test), confirming hypothesis **H3a**. These findings facilitate the explanation of the variance in test score in terms of motivational factors.

Additionally, current results also revealed that learners' expectations regarding their satisfying score achievement (i.e., goal-expectancy) is a strong direct determinant of all the aggregated response-time factors in both testing settings. This finding is in line with previous results (Papamitsiou et al., 2014), confirms that goal-expectancy is strongly correlated with response-times (supporting hypotheses **H2a**, **H2b**, **H2c**), and explains *how* learners' preparation to take the test is projected on their response-times; well-prepared learners will aggregate more time to answer correctly, devoting considerable time to answer items of increased difficulty/demands, and they will accumulate less time to submitting wrong answers. The opposite time allocation characterizes the poorly prepared learners and it reasons their test score, as previously stated.

On the contrary, the direct effects of the motivational self-efficacy factor on accumulated response-times to answer correctly and answer wrongly were verified only in fixed test conditions (Table 4-5). However, in the adaptive setting, self-efficacy has strong *indirect* effects on the response-time factors (Table 4-6), confirming hypothesis **H3b**, **H3c**. Further exploration is required on the direct cause-effect relation of self-perceptions of efficiency on time allocation on items in adaptive testing.

Another interesting finding of this study was that, both in fixed and in adaptive tests, the effect of self-efficacy on effort exertion was not statistically significant, not supporting hypothesis **H3d**. Although it was expected that more self-confident learners who believe that they can perform well, would be more careful when they answer the test items, and persist longer in their efforts to successfully accomplish the tasks, compared to less self-reliant learners, the findings in this study are weak and do not support prior claims (Tabak et al., 2009).

Moreover, effort was a strong determinant of test score (Table 4-6); the total effect of effort exertion during items' manipulation on test score is found to be significant and positive in adaptive ($\beta=0.253$, $t=3.478$) as well as in fixed ($\beta=0.134$, $t=3.599$) test conditions (supporting hypothesis **H4a**). The time-driven measure for effort used in this study (i.e., RTE; Wise & Kong,

2005), incorporates guessing and expresses learners' on-task engagement. Guessing is considered as an important factor in IRT's 3PL model (Birnbaum, 1968) since it can cause a confusion regarding the validity of the result of learners' actual knowledge measurement. The current results (i.e., the higher the effort, the higher the test score) provide additional proof to prior claims (Setzer et al., 2013) that taking the test seriously is reflected on on-task engagement.

However, the effect of effort on learners' time allocation on difficult items was found to be statistically significant, and to support hypothesis **H4b**, only in fixed test conditions. In adaptive testing, no strong relationship between the two factors was identified, although they were weakly positively associated (Table 4-5). This result might be due to the nature of adaptive test itself, which tailors the next administered item to the estimated learner's mastery level. As such, difficult items are delivered only to more proficient or struggling learners. Thus, no or low effort is considered for the poorly performing learners, resulting to a decrease in the overall effect of effort on time-allocation on hard items.

Furthermore, cheating might give the students some extra points, but at the end, it does not guarantee the final grade success, unless the students actually try harder to answer. Current findings advance our understanding on *why* the less engaged learners, who guess the correct answers instead of trying to figure it out, score lower compared to those who exert higher devotion on items: those who usually cheat are low goal-orientation students (**H2d**). The results make it apparent that goal-expectancy has a significant positive effect on on-task effort ($\beta=0.215$, $t=2.978$ in fixed test; $\beta=0.397$, $t=5.913$ in adaptive test). This means that well-prepared, highly motivated learners try harder to achieve their goals and engage more, exhibiting solution behavior, rather than cheating and guessing the answers.

In addition, learners' overall satisfaction from the test, measured after its completion in terms of perceived clarity of the test items, was a strong direct determinant of the test score in both test settings (Table 4-6). This result confirms hypothesis **H5a** and implies that the learners' satisfaction from the comprehensibility of the test items (Wang, 2003) is directly reflected on the test score ($\beta=0.135$, $t=3.575$ in fixed test; $\beta=0.211$, $t=5.637$ in adaptive test). In a sense, this finding is intuitive since the validity and clarity of the test items are critical for learners' response strategies; besides, the items are designed to measure knowledge acquisition, (Doll & Torkzadeh, 1988; Fitzpatrick, 1983; Sun et al., 2008) and thus, the more understandable they are, the more possible for the learners to successfully answer them. Even more, in the adaptive testing conditions, the adaptation mechanism decides upon the next most suitable item to deliver according to the learners' previous answers on the basis of information to individually match the ability level of each learner (Finkelman, Kim, Weissman, & Cook, 2014) and their quality is critical to this decision. Thus, it is straightforward that learners' self-reflection on their own understanding of the test content will influence their score as well.

On the contrary, the results regarding the effects of effort on perceived clarity of the test items did not support the initial hypothesis, neither in the fixed nor in the adaptive test (hypothesis **H4c**). Specifically, a weak positive and a strong positive impact of exhibiting solution behavior on how much the learners understood the items as clear were detected in the two test set-ups respectively ($\beta=0.080$, $t=1.347$ in fixed test; $\beta=0.361$, $t=6.611$ in adaptive test). This finding is interesting because it implies that due to the fact that the items were clear and comprehensible, the learners engaged more with them, trying to solve them, instead of guessing the answer. As such, it extends previous claims that if learners understand the items, not only they are more likely to be successful (Puzziferro, 2008), but also they are more likely to exhibit effort exertion and engage in solution behavior.

Lastly, the results of this study regarding the effect of time to answer the harder items on the learners' perceived clarity about these items (hypothesis **H6**) were inconclusive: in the fixed test, this effect was weakly positive, whereas in the adaptive test, the effect was weakly negative (Table 4-5). This hypothesis requires additional empirical explorations.

4.7. Implications and conclusions

Inspired from previous research works – originating mostly from the well-established Psychometrics and learning analytics domains – a set of well-defined cognitive and motivational factors were selected to explain the variance in learners' test score; their between cause-effect relationships were statistically explored as well. The developed causal and measurement models diagnose satisfactorily the obtained test score and contribute sufficiently to reasoning learners' actions in formal computer-based summative fixed and adaptive testing conditions in higher education. Consequently, the arising question is: *how we could exploit these results towards developing credible testing systems or services?*

4.7.1. Towards exploiting testing analytics

Practical Implications: Based on the findings from the current study, testing analytics could be infused into a testing system in order to facilitate the interpretation of the achieved scores. Monitoring the identified cognitive and motivational factors, one can be more aware about what to expect from the learners at the end of the exam; this information can be proven really useful for the interpretation of the test score and it can also explain why learners act the way they do during the exam. For example, the well-prepared learners exhibit higher self-efficacy, tend to avoid cheating, are satisfied from the test items, perceiving them as challenging and coherent to the course's content, and they usually get higher scores at the end of the tests. In a similar manner, in fixed tests that include the same items for all learners, it is more complicated for poorly prepared learners to deeply understand the requirements of the harder items, leading them to a blur judgement regarding the comprehensibility of the items, as opposed to adaptive tests, where the items are delivered according to the individual's detected expertise, and thus, the learners' perceptions about the clarity of the items are more straightforward. These poorly prepared

learners do not allocate much time on hard items, they usually are not self-reliant and they tend to believe that their low scores are due to the lack of clarity of the test items. In a sense, the cause-effect relationships between the included factors tell a “story” about learners’ manipulation of the items and their final answering decisions.

Going a step further from summative testing procedures, the findings of this study could potentially be beneficial for formative assessment tests, as well. For example, building systems that would allow learners to monitor their own progress, and help them evaluate and adjust their strategies to increase goal achievement is a request. By providing learners the opportunity to monitor their progress on a test and, at the same time, the autonomy to choose between appropriately suggested items to solve next, we would allow them to self-regulate and determine a unique, personal “path” to successfully complete their goals. For example, simple visualizations of the learners’ time-spent on each item compared to the average time-spent of the rest of the class, or compared to the aggregated TTAC and TTAW of the class, could increase the learners’ awareness regarding their own time-management, and affect their understandings regarding the difficulty of the item. However, not all learners are competitive with the class. In that case, learners might prefer to see their own variables compared, for example, with their own past answers, or their performance on a highly scored item.

Methodological Implications: An analytics-driven approach like the one suggested in this study, could help instructors to adjust their examination strategy. It is commonly agreed that there is a need to gain insight into learners’ perceptions and discover how they behave when dealing with test items that have different requirements. When instructors mark an item as easy, medium or hard, they should be aware of the learners’ comprehensions, their ability level and an estimation of effort needed to accomplish the task. Development of high-quality examination systems is important for instructors who do not have direct access to the advanced commercial methods used by educational testing companies. An analytics-enhanced testing system could increase instructors’ awareness regarding the progress of the class during testing and inform them about the items that might confuse the learners. For example, it could create diagnostics on the progress of each learner, the frequency of selection of each item and which items should be changed, reconsidered or removed, and propose respective adjustments.

4.7.2. Conclusions and Future work

Existing testing methods have provided well-established formats for the measurement of knowledge. However, they have undergone comprehensive criticism, mostly due to a superficial approach of assessing learners’ skills and knowledge that fails to look deeper in the causation of learners’ responses. The search in literature revealed a number of factors that have been acknowledged for their capabilities to explain the obtained test score. Yet, a model that holistically explains the cause-effect relationships between these factors and the contribution of each one of them on the interpretation of the test score was missing. Moreover, none of the prior studies has

directly infused the testing procedures with critical information to boost the instructors' awareness beyond the test score, to the interpretation of this result.

The innovation and contribution of the present study is that it followed an analytics-driven methodology and introduced *testing analytics* as a context-aware process that monitors and records detailed data related to the online testing context, interprets and maps the real state of the data, organizes them, uses them, and continues by predicting their future state. From the demonstrated findings, it becomes apparent that testing analytics associate learners' manipulations of the items to their motivation, effort and to the items' difficulty, and reason their choices in terms of response-times, guessing, self-efficacy and satisfaction. Test score is sufficiently and directly explained by accumulated response-times on correctly or wrongly answered test items, by time-spent on items with higher difficulty level, by learners' response-time effort, and by their satisfaction from the items' comprehensibility. The indirect effects of learners' goal-expectancy and their perceived self-efficacy on test score were also statistically significant, allowing for test score interpretation and advancing our understanding on the motivational factors behind learners' actions.

A testing analytics infused system could inform instructors about learners' progress during testing (e.g. via visualizations or detailed diagnostics) and boost their awareness regarding the expected achievement, to make sure that learners do not have to "*wait to fail*" before they might be eligible for help. For example, total time to answer could indicate learners' disengagement from the test, and could be explored as a possible factor of a "*gaming the system*" behaviour (Baker et al., 2004). "*Gaming the System*" was introduced to explore disengagement from tutors and detect the factors that are associated with this behaviour. Thus, total time to answer could be analyzed into "useful"/"wasted" time and might be examined in relation to learners' emotional state during testing. The appropriate integration of emotional factors should be investigated as well.

The approach suggested in this study was applied on a dataset collected during testing procedures in the context of midterm exams. The nature of the data collected (time-based parameters) and the general-purpose methodology followed for the analysis of these data, render this approach replicable and transferable to other assessment contexts, as well. Further, the proposed ideas should be combined with additional features like emotional states, and/or a time-driven formulation of items' actual difficulty and learners' motivation to achieve on low-stakes tests. Finally, since the mechanisms for tracking temporal data are cost-effective, consume low computational resources, and can be easily implemented in any testing system, the possibility of adopting similar methodologies even for larger scaled high-stakes examination procedures should also be explored.

Chapter 5 : From prediction of performance to learner models – the role of personality

*“The same boiling water that softens the potato, hardens the egg-
It’s about what you’re made of, not the circumstances”*

Anonymous

Exhibiting achievement behavior during Computer-based testing: what temporal trace data and personality traits tell us?

5.1. Introduction

The introduction of digital technologies in education has already opened up new opportunities for tailored, immediate and engaging Computer Based Assessment (CBA) experiences (Bennett, 1998; Chatzopoulou & Economides, 2010). CBA is the use of information technologies (e.g. desktop computers, mobiles, web-based, etc.) to automate and facilitate assessment and feedback processes. Computerized assessment allows for monitoring and tracking data related to the context, interpreting and mapping the real current state of these data, organizing them, using them and predicting the future state of these data (Leony et al., 2013; Papamitsiou & Economides, 2016; Triantafillou et al., 2008). On the contrary, traditional offline assessment render these facilities unattainable. However, differences in learners’ behavior during CBA have a deep impact on their educational performance and their level of achievement. Compiling learners’ behavior in CBA processes and creating the corresponding behavioral models is a primary educational research objective (e.g., Abdous et al., 2012; Blikstein, 2011; Shih et al., 2008).

Learner behavioral modelling can be defined as the process of information extraction from different data sources into a profile representation of learner’s knowledge level, cognitive and affective states, and meta-cognitive skills on a specific domain or topic (McCalla, 1992; Thomson & Mitrovic, 2009). A learner model is a synopsis of multiple learner’s characteristics – either static (e.g., age, gender, etc.), or dynamic. Performance, goals, achievements, prior and acquired domain knowledge (Self, 1990), as well as learning strategies, preferences and styles (Peña-Ayala, 2014) are among the most popular dynamic characteristics. Decisions making abilities, critical and analytical thinking, communication and collaboration skills (Mitrovic & Martin, 2002), motivation, emotions/feelings, self-regulation and self-explanation (Peña, Kayashima, Mizoguchi, & Dominguez, 2011) are also commonly used to complement the learner’s profile.

More recently, the time dimension has been explored for modelling learner behavior. For example, Shih, Koedinger and Scheines (2008) used worked examples and logged response times to model the students’ time-spent in terms of “thinking about a hint” and “reflecting on a hint”. Other studies examined the effect of student’s response times on prediction of their achievement

level (Papamitsiou et al., 2016; Xiong et al., 2011), explored the relationships between study-time and motivation (Nonis & Hudson, 2006), and proposed what should be adapted in the Computerized Adaptive Testing (CAT) context regarding orientation to time (Economides, 2005).

Efficient use of time is widely assumed to be a key skill for students (Claessens, van Eerde, Rutte, & Roe, 2007; Kelly & Johnson, 2005; MacCann, Fogarty, & Roberts, 2012), and it is summarized under the term "*time management behavior*". Claessens et al. (2007, p. 36) defined time management behavior as "behaviors that aim at achieving an effective use of time while performing certain goal-directed activities". However, the results from empirical evidence on the relationship between students' time-management and level of achievement converge to an unclear landscape (Claessens et al., 2007; Hamdan, Nasir, Rozainee, & Sulaiman, 2013; Trueman & Hartley, 1996).

5.2. Related work & Motivation of the research

Explaining students' time-management according to behavioral models enhanced with personality aspects is expected to provide additional evidence towards better understanding when they actually exhibit achievement behavior. According to Pervin and John (Pervin & John, 2001, p. 10), "personality represents those characteristics of the person that account for consistent patterns of feeling, thinking, and behaving". In a sense, personality could be defined as the set of the individuals' characteristics and behaviors that guide them to make decisions and act accordingly under specific conditions (Chamorro-Premuzic & Furnham, 2005). Researchers have concluded to five factors that describe personality traits (Costa & McCrae, 1992; John & Srivastava, 1999). According to the Big Five model, these factors are: a) agreeableness, b) extraversion, c) conscientiousness, d) neuroticism, and e) openness to experience.

A search in literature revealed that there is limited evidence that agreeableness is relevant to time management behavior (Claessens et al., 2007; for conflicting evidence see McCann et al. 2012). Moreover, researchers found that extraverts showed faster response times than introverts (Dickman & Meyer, 1988; Robinson & Zahn, 1988), while others reported no overall differences between groups (Casal, Caballo, Cueto, & Cubos, 1990). Yet, in a study of undergraduate students, it was found that highly conscientious students use their time more efficiently (Kelly & Johnson, 2005). It was also found that conscientiousness was a significant predictor of test performance, and time-on-task fully mediated the conscientiousness-performance relationship (Biderman, Nguyen, & Sebren, 2008). Van Hoya & Lootens (2013) found that highly neurotic individuals is less likely to use time management strategies, while, individuals high on openness find it difficult to manage their time effectively to complete tasks.

From the above derives that the experimental results regarding the relationships between personality traits and time-management skills are inconclusive. Thus, additional research is required, and different research approaches should be considered. Recent advances in the field of assessment analytics, triggered our interest on exploiting analytic methods in this

case as an alternative research methodology. Assessment analytics concern applying fine-grained analytic methods on multiple types of data, aiming to support teachers and students during the assessment processes. This is a repetitive procedure that continues by making practical use of detailed student-generated data captured by CBA systems, and providing personalized feedback accordingly (Ellis, 2013).

Moreover, when it comes to Computer-Based Testing (CBT) procedures – which is a typical, popular and widespread method of online assessment – it would be worthwhile to have in-depth knowledge of students' behavior in the testing environments, and understand how this affects their achievement level. In turn, this insight will contribute to the improvement of the testing services at a larger scale. This is the first study – to the best of our knowledge – that exploits assessment analytics methods for associating personality traits with response-times for modelling examinees' achievement behavior during CBT.

Despite the criticism on interpreting students' logged data into actual learning behaviors, a large body of literature has provided empirical evidence of strong correlation between them (Jo, Kim, & Yoon, 2015; Romero, López, Luna, & Ventura, 2013). In our approach, the choice of the accumulated response times to code time-management behavior is justified because these variables could facilitate multiple purposes: providing analytics related to time-management for increasing students' awareness on how they progress on each item compared to the rest of the class during testing, identifying the actual difficulty of an item for further adapting the test to examinee's abilities on-the-fly, making possible the detection of unwanted examinee behavioral patterns (such as guessing or slipping) via process mining methodologies, to name a few. Moreover, the mechanisms for tracking temporal data are cost-effective, consume low computational resources, and can be easily implemented in any CBA system.

5.2.1. Objectives and research questions

This study's objective is to carry out an experimental study in order to contribute towards exploiting assessment analytics methods for deeper understanding the examinee's time-spent behavior during CBT according to the five personality traits. The main focus of this study is on exploring the use of time-driven assessment analytics with the Big Five Inventory (BFI - John & Srivastava, 1999) to explain achievement behavior in terms of personality and response times on task-solving. This is expected to further improve student models for guiding personalization of testing services. As such, we also aim to investigate assessment analytics capabilities on classifying students, and contribute to creating enhanced student models. Thus, the research questions are twofold:

RQ1: *Which is the effect of the five personality factors on time-spent behavior during CBT?*

RQ2: *How accurately can we classify the students during testing according to their personality traits and behavior expressed in terms of response-times?*

5.3. Temporal learning analytics for prediction of performance

The Temporal Learning Analytics (TLA) have been proposed as a predictive model of achievement level in order to interpret students' participation and engagement in assessment activities in terms of "time-spent". Previous studies (Papamitsiou & Economides, 2014b; Papamitsiou et al., 2014) structured a measurement model consisting of temporal (response-times) and other latent factors (e.g. goal-expectancy, level of certainty) in order to predict students' score during CBT.

More precisely, these studies explored the effects of total time to answer correctly (TTAC), total time to answer wrongly (TTAW), goal-expectancy (GE) and level of certainty (CERT) on test score (Actual Performance - AP) during CBT. Preliminary results highlighted a detected trend that TTAC and TTAW have a direct positive and a direct negative effect on AP respectively, while GE was found to be a statistically significant indirect determinant of AP (Papamitsiou et al., 2014). Furthermore, level of certainty (CERT) – i.e. the students' cautiousness and confidence during testing in terms of time-spent on answering the quiz – explains satisfactorily the students' AP during low-stakes CBT procedures as well. In addition, CERT has direct positive and negative effects on TTAC and TTAW respectively. That is because more confident students (i.e. with higher level of certainty) will spend more time on correctly answering the questions, while unconfident students (i.e. with lower level of certainty) will spend more time and finally will submit the wrong answers (Papamitsiou & Economides, 2014b). In a sense, certainty seems to increase students' effort to answer the quiz. The suggested TLA model explains almost the 63% of the variance in AP. These findings are illustrated and synopsized in Figure 5-1.

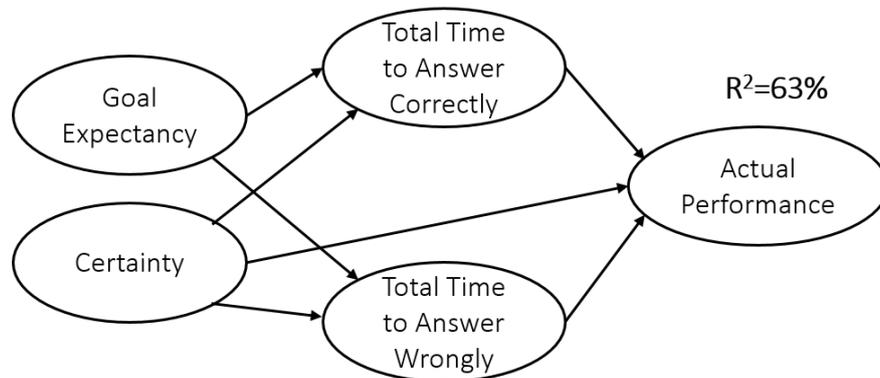


Figure 5-1. TLA for predicting performance during CBT (Papamitsiou & Economides, 2014b).

Moreover, Papamitsiou and Economides (2014d) explored the effect of extroversion (E) and conscientiousness (C) on students' time-spent behavior during CBT in a case study with 96 secondary education students. Preliminary results from this study showcased that E is positively related to GE and C positively affects the students' CERT. Finally, results from former studies revealed that response-times have satisfactory discrimination ability regarding students' behavior and are appropriate for modeling student behavior in learning activities (Papamitsiou et al., 2016).

5.4. Research Model and Hypothesis – Concepts of student models

As stated in the previous chapters, goal-expectancy (GE) is a variable which measures goal orientation regarding the use of a CBA (Terzis & Economides, 2011). Further, level of certainty (CERT) is a time-dependent measure of cautiousness during the assessment. This study goes a step further by correlating these factors to personality traits. The goal is to develop and explore a causal model to determine and explore the effect of personality traits on time-spent behavior and achievement level during CBT. In Figure 5-2, the dashed arrows represent formerly explored hypotheses that will not be re-examined here. The rest of the arrows depict the relations between variables that will formulate our research hypotheses.

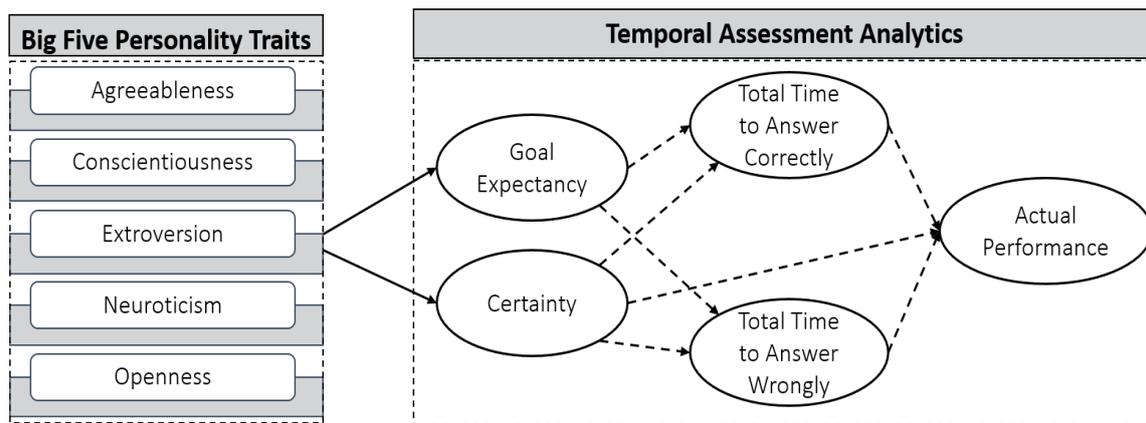


Figure 5-2. Overall research model and variables relationships.

5.4.1. Personality traits and hypothetical relationships

Agreeableness (A): Agreeableness refers to the humane aspects of people, such as altruism, being helpful, sympathetic and emotionally supportive towards others (Digman, 1990). The behavioral tendencies typically associated with this factor include being kind, considerate, co-operative, and tolerant (Graziano & Eisenberg, 1997). Agreeable students usually comply with teacher instructions, tend to exert effort and stay focused on learning tasks (Vermetten, Lodewijks, & Vermunt, 2001). This trait was positively correlated with learning goal orientation (Bipp, Steinmayr, & Spinath, 2008), mostly in collaborative learning contexts. Although CBT is not a typical collaborative process, agreeable students are expected to exceed higher goal expectancy and cautiousness. Thus, we hypothesize that:

H1: *Agreeableness will have a positive effect on goal-expectancy*

H2: *Agreeableness will have a positive effect on certainty*

Extraversion (E): Extraversion implies an energetic personality and includes traits such as sociability, activity, assertiveness, and optimism (Watson & Clark, 1997). This trait is related to leadership (John & Srivastava, 1999) and was significantly correlated to motivational concepts such as goal-setting and self-efficacy (Judge & Ilies, 2002). Because extraverts tend to set high achievement goals and attain them, they are likely to set active skill/knowledge acquisition goals. However, research has shown that extraversion correlates negatively with caution and

carefulness. It means that the less extrovert the person is, the more careful the person will be (Boroujeni, Roohani, & Hasanimanesh, 2015). The above imply that extrovert students are more likely to have higher expectations from their preparation, but lower cautiousness due to their impulsive and spontaneous behavior. Thus, we hypothesized that:

H3: *Extroversion will have a positive effect on goal-expectancy*

H4: *Extroversion will have a negative effect on certainty*

Conscientiousness (C): Conscientiousness describes impulse control that facilitates task- and goal-oriented behavior, such as thinking before acting, delaying gratification, planning, organizing, and prioritizing tasks. It is a personality trait used to describe persons being careful, responsible and with a strong sense of purpose and will (Devaraj, Easley, & Crant, 2008; John & Srivastava, 1999). Studies have shown that conscientiousness was very strongly correlated with an achieving style and modestly correlated with a deep style (Furnham, Christopher, Garwood, & Martin, 2008). Conscientious students are described as achievement oriented (John & Srivastava, 1999). Conscientiousness has been found to be a strong predictor of goal-setting, achievement expectancy, and self-efficacy motivation (Judge & Ilies, 2002). These imply that conscientious students are more likely to be cautious during assessment, and exhibit higher goal expectations. Thus we hypothesized that:

H5: *Conscientiousness will have a positive effect on goal-expectancy*

H6: *Conscientiousness will have a positive effect on certainty*

Neuroticism (N): Neuroticism represents individual differences in distress and refers to degree of emotional stability, impulse control, and anxiety (McCrae & John, 1992). With respect to neuroticism and self-regulation, Kanfer and Heggstad's (1997) model predicts that anxiety leads to poor self-regulation because anxious individuals are not able to control the emotions necessary to maintain on-task attention. Previous results indicated a negative relation between neuroticism and goal-setting motivation, expectancy motivation, and self-efficacy motivation (Judge & Ilies, 2002). Neurotic students are expected to face CBT as a stressful procedure, and they are likely to find it difficult to relax, concentrate and stay focused during the assessment. Their general negativity will probably have a negative effect on their goal expectancy and level of certainty during CBT. Thus, we hypothesized:

H7: *Neuroticism will have a negative effect on goal-expectancy*

H8: *Neuroticism will have a negative effect on certainty*

Openness to Experience: Openness to experience is reflected in a strong intellectual curiosity and a preference for novelty and variety. Individuals who score high on openness to experience are creative, flexible, curious, unconventional, search for new experiences and knowledge, and display an eager to learn (McCrae, 1996). This trait has been positively correlated with learning motivation (Tempelaar, Gijsselaers, van der Loeff, & Nijhuis, 2007) and critical thinking (Bidjerano & Dai, 2007). These characteristics lead researchers to link openness with

engaging in learning experiences (Barrick, Mount, & Judge, 2001), and associate it with deep learning (Chamorro-Premuzic, Furnham, & Lewis, 2007). This mean that they are more likely to inquire knowledge and make considerations rather than maintain their level of certainty. Moreover, individuals with a learning goal orientation demonstrate behaviors and hold beliefs that are consistent with those who are high in openness to experience (Zweig & Webster, 2004). Thus, we hypothesized:

H9: *Openness to experience will have a positive effect on goal-expectancy*

H10: *Openness to experience will have a negative effect on certainty*

The research model and hypotheses are illustrated in Figure 5-3.

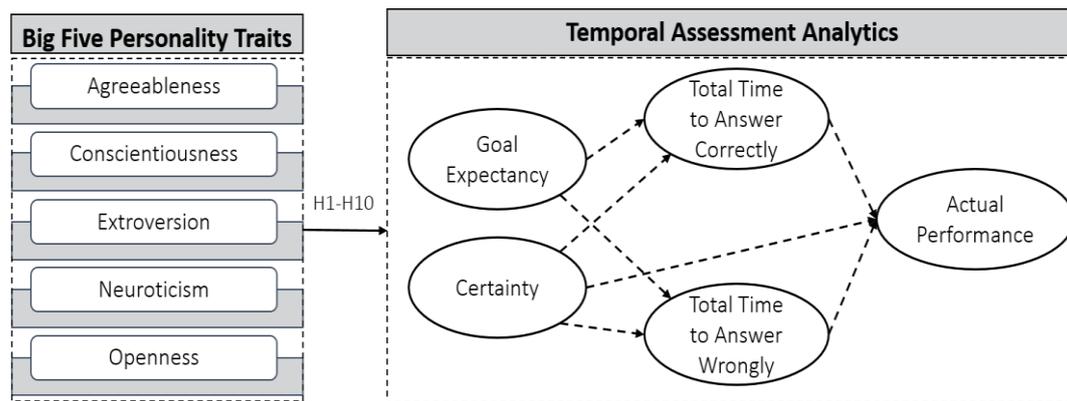


Figure 5-3. Research model and hypothesis.

We should mention that investigating hypotheses **H2**, **H4**, **H6**, **H8** and **H10** – which are all related to the time-driven level of certainty (CERT) variable – are feasible only in CBT contexts, and could not be explored in traditional offline testing conditions.

5.4.2. Conceptual classification of examinees

Supervised classification is the task of identifying to which group (label) a new observation is categorized, according to a training set of data containing observations whose group membership is known (Duda, Hart, & Stork, 2000). In other words, supervised classification is about learning a target function f to map the input feature space x to one of the discrete, predefined class labels y . In our study, the exploratory variables (i.e., the feature space) include the response-times variables (i.e., TTAC, TTAW), the behavioral variables (i.e., GE, CERT) and the personality traits (i.e., A, E, C, N, O). The class to be predicted is one of the different levels of achievement in the CBT. Five levels of achievement during the CBT were identified: the “low achiever”, the “careless achiever”, the “neutral achiever”, the “struggling achiever” and the “high achiever”. We used these terms to name the groups and make sense of the results. We discretized the target variable, i.e., the different levels of achievement, for multiple reasons. Firstly, many machine learning algorithms are known to produce better models by discretizing continuous attributes (Kotsiantis & Kanellopoulos, 2006). Secondly, some models (e.g. Naive Bayes, used in this study, and Decision Trees) do not function with continuous features, but require discrete ones. Even more, mining of association rules with continuous attributes is a major research issue,

and discretizing continuous attributes is necessary in this case (Srikant & Agrawal, 1996). Thirdly, it is more convenient computationally to represent information as a finite set states and more meaningful to elaborate on a handful of cases. Lastly, a “reasonable” number of partitions during discretization has been acknowledged to tackle data overfitting issues in machine learning and data mining domains. The behavioral patterns which are assumed to be relevant to each level of achievement, contain all of the selected features and aim to represent how students actually behave during CBT.

As seen from previous studies, response times to answer correctly have a positive impact on AP and time-spent on wrongly answered questions has a negative effect on AP (Papamitsiou et al., 2014). In this study, we also wanted to consider response times as a core feature of the achievement class the student belongs to. Thus, we assumed that high TTAC is a characteristic of high achievers, while high TTAW better suits the class of low achievers. Struggling achievers might have high TTAC, but they also aggregate non negligible amounts of time to TTAW. Conversely, although careless achievers are marked with higher TTAW, they also gather appreciable TTAC.

Similarly, we assumed that high and struggling achievers usually score high in GE, while for low and careless achievers a lower GE is expected. Regarding CERT, high achievers are foreseen to exhibit higher levels of certainty, but this feature should be somewhat lower for struggling achievers, who nevertheless demonstrate a trend to increase their certainty. On the other hand, low and careless achievers are supposed to be less confident students, expressing lower levels of certainty during CBT.

Furthermore, personality traits are also key features of the student models. Previous results on the relations between personality traits and achievement behavior (e.g., Chamorro-Premuzic et al., 2007; Furnham et al., 2008; Judge & Ilies, 2002) allow for the following assumptions: high and struggling achievers are expected to score higher in extroversion, agreeableness, conscientiousness and openness, while low and careless achievers will demonstrate amplified neuroticism.

Table 5-1. Description of achievers’ classes and their characteristics

<i>C1: Low Achiever</i>	<i>C2: Careless Achiever</i>	<i>C3: Neutral Achiever</i>	<i>C4: Struggling Achiever</i>	<i>C5: High Achiever</i>
TTAC (--)	<i>TTAC (-)</i>	<i>TTAC (+-)</i>	<i>TTAC (+)</i>	<i>TTAC (++)</i>
TTAW (++)	<i>TTAW (+)</i>	<i>TTAW (-+)</i>	<i>TTAW (-)</i>	<i>TTAW (--)</i>
GE (--)	<i>GE (-)</i>	<i>GE (+-)</i>	<i>GE (+)</i>	<i>GE (++)</i>
CERT (--)	<i>CERT (-)</i>	<i>CERT (-+)</i>	<i>CERT (+)</i>	<i>CERT (++)</i>
E (--)	<i>E (-)</i>	<i>E (+-)</i>	<i>E (+)</i>	<i>E (++)</i>
A (--)	<i>A (-)</i>	<i>A (+-)</i>	<i>A (+)</i>	<i>A (++)</i>
C (--)	<i>C (-)</i>	<i>C (+-)</i>	<i>C (+)</i>	<i>C (++)</i>
N (++)	<i>N (+)</i>	<i>N (-+)</i>	<i>N (-)</i>	<i>N (--)</i>
O (--)	<i>O (-)</i>	<i>O (+-)</i>	<i>O (+)</i>	<i>O (++)</i>

Table 5-1 synthesizes the descriptions of the five classes of achievers during CBT according to the research hypotheses and assumptions. This table provides a summary of the features per achievement category, using the signs “+” and “-“ for indicating dominant or absent occurrence of the respective feature. In this study, we want to observe if the selected features are equally suitable for the configuration of students’ classes, and how the assumptions on behavioral patterns are related to students’ final score.

5.5. Methodology

5.5.1. Research participants and data collection

One hundred and twelve (112) undergraduate students (48 males [42.9%] and 64 females [57.1%], aged 19-26 years old (M=20.7, SD=1.887, N=112)) from the Department of Economics at University of Macedonia, Thessaloniki, Greece, were enrolled in the experimental procedure. Five (5) randomly generated groups of 20 to 25 students attended the midterm exams of the Management Information Systems II course (related to databases, telecommunications and e-commerce), for 50 minutes each group, at the University computer lab.

For the purposes of the examination, we used 25 questions in total, distributed in the 5 equivalent tests of 9 multiple choice questions each (some of the questions were shared in more than two tests). Each question had two to four possible answers, but only one was the correct. The questions were delivered to the participants in predetermined order. The fixed-testing module of the LAERS environment allowed students to temporarily save their answers on the items, to review them, to alter their initial choices, and to save new answers. Students could also skip an item and answer it (or not) later. They submitted the quiz answers only once, whenever they estimated that they were ready to do so, within the duration of the test.

During the design of the testing procedure, we asked two experts to rate all 25 questions regarding their difficulty (easy, medium, hard). The two experts agreed on the questions’ difficulty. All questions used in the current study correspond to the first five levels of the factual, conceptual and procedural domains of the knowledge dimension according to the revised Bloom’s taxonomy (Anderson (Ed.) et al., 2001) for reasons of holistically assessing knowledge acquisition within the available quiz time.

For the score computation, only the correct answers were considered, without penalizing the incorrect answers (i.e., without negative scores). Further, each question’s participation on the score was according to its difficulty level, varying from 0.75 points (easy) to 1.25 points (medium) to 1.625 points (hard). In case students chose not to submit an answer to an item, they received zero points for this one.

Before taking the tests and right after the completion of the procedure, each participant had to answer to the pre-test and post-test questionnaires that measure each student’s goal expectancy and personality traits respectively. For the needs of the study, goal-expectancy was measured by utilizing three items from the Computer Based Assessment Acceptance Model

(Terzis & Economides, 2011). In order to extract the students' personality traits the BFI was used. BFI has 44 items: eight items for extraversion (E) and neuroticism (N), nine items for agreeableness (A) and conscientiousness (C), and ten items for openness to experience (O). The five point Likert-type scale with 1 = strongly disagree to 5 = strongly agree was used to measure each of these items (Appendix C). We selected BFI, because it has been known for its reliability, validity and clear factor structure (Srivastava, John, Gosling, & Potter, 2003).

The participation to the midterm exams procedure was optional. Students were aware that their answers were being tracked, but not that their time-spent was being measured, because we wanted them to act spontaneously. All participants signed an informed consent form prior to their participation. The informed consent explained to the participants the procedure and it gave the right to researchers to use the data collected during the CBT for research purposes. As external motivation to increase students' overall effort, we set that their score would participate up to 30% of their final grade. It should be noted that the samples of 112 participants and 25 questions are limited (compared to the large scale tests implemented by the testing organizations) and thus, they are very likely biased.

5.5.2. Data analysis for the structural and measurement model

In this study, for addressing RQ1, the construction of a path diagram that contains the structural and measurement model was conducted with the Partial least-squares (PLS) analysis technique (Chin, 1998). PLS is suitable for studies that have small samples. In PLS the sample size has to be a) 10 times larger than the number of items for the most complex construct, and b) 10 times the largest number of independent variables impact a dependent variable (Chin, 1998). In our model, the most complex predictor is O with ten items (see section 5.4.1), and the largest number of independent variables impacting a dependent variable is three (TTAC, TTAW and CERT to AP). Thus, our sample (112) is fair enough, since it is above the required value of 100.

In PLS, the items' factor loadings on the corresponded constructs have to be higher than 0.7 (Chin, 1998). The construct validity is confirmed by obtaining convergent – discriminant validity. Convergent validity is carried out by Average Variance Extracted (AVE) and has to be higher than 0.5 and the AVE's squared root of each variable has to be higher than its correlations with the other constructs (Barclay et al., 1995; Fornell & Larcker, 1981). Cronbach's α and composite reliability (CR) are used to confirm reliability of the measurement model, and they both have to be higher than 0.7 (Tenenhaus, Vinzi, Chatelin & Lauro, 2005).

Structural model evaluates the relationship between exogenous and endogenous latent variables by examining the variance measured (R^2) (Chin, 1998). R^2 values equal to 0.02, 0.13 and 0.26 are considered as small, medium and large respectively (Cohen, 1988). Moreover, a bootstrapping procedure is used to evaluate the significance of the path coefficients (β value) and total effects, by calculating t-values. Finally, in PLS the quality of path model can be evaluated by the Stone-Geisser's Q^2 value (Geisser, 1974; Stone, 1974), which represents an evaluation

criterion for the cross-validated predictive relevance of the PLS path model. The Q^2 statistic measures the predictive relevance of the model by reproducing the observed values by the model itself. A Q^2 greater than 0 means the model has predictive relevance; whereas Q^2 statistic less than 0 mean that the model lacks predictive relevance. For the measurement and the structural model we used SmartPLS 3.0.

5.5.3. Data analysis for supervised classification

Towards addressing RQ2, our next step was to classify students according to their personality and time-spent behavior during the CBT. The task was to determine to which of the predefined classes a new observation belongs, on the basis of a training set of correctly identified observations. These predefined classes contain instances with measurements on different variables (predictors) whose class membership (labels) is known. In this study, we used as predictors the students' time-based characteristics (i.e., TTAC, TTAW, CERT), and their self-reported characteristics (i.e., GE and personality traits – A, E, C, N, O) and as class labels their level of achievement (AP). We explored Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF) and classification based on association rules (or class-association rules – CARs, and in particular the JCBA algorithm) for classifying students. These advanced supervised learning techniques are among the most common approaches explored with a plurality of different attributes in the learning analytics and educational data mining research domain.

- *Support Vector Machines (SVM)* is a supervised learning method for linear modelling. For classification purposes, nonlinear kernel functions are often used to transform the data into a feature space of a higher dimension than that of the input before attempting to separate them using a linear discriminator (Cortes & Vapnik, 1995). In this work, a third degree polynomial kernel function was employed.
- *Naïve Bayes (NB)* are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the predictors in each class. The method estimates the parameters of a probability distribution, computes the posterior probability of that sample belonging to each class, and classifies the test data accordingly (Tan et al., 2005).
- *Random Forests (RF)* are ensembles of decision trees. The training algorithm for RF applies the general technique of bagging: repeatedly selects a random sample with replacement of the training set, fits trees to these samples, and uses these replicates as new learning sets. At each candidate split in the learning process, RF select the best among a subset of predictors (subset of the features) randomly chosen at that node (Breiman, 2001).
- *Classification rule mining* aims to discover a small set of rules in the dataset to form an accurate classifier. Classification Based on Association rules is an integration of classification rule mining and association rule mining (Liu, Hsu, & Ma, 1998). The integration is done by focusing on mining association rules, and the set of rules that are selected as candidate rules, satisfy certain support and confidence thresholds. They are called the classification

association rules (CARs), they have only a particular attribute in the consequent, and can be used to build a model or classifier. When predicting the class label for an example, the best rule (with the highest confidence) whose body is satisfied by the instance is chosen for prediction.

The performance of a classification model is expressed in terms of its *error rate*, which is given as the proportion of wrong prediction to the total predictions (Alpaydin, 2010; Tan et al., 2005). The errors committed by a classifier are generally divided into resubstitution errors (training errors) and test errors (generalization errors). The resubstitution error is the proportion of misclassified observations on the training set, whereas the test error is the expected prediction error on an independent set. A good model must have low resubstitution error as well as low test error (Mitchell, 1997). Further, a method commonly used to evaluate the performance of a classifier is cross-validation. The k-fold cross-validation method segments the data into *k* equal-sized partitions. This procedure is repeated *n* times so that each partition is used the same number of times for training and exactly once for testing. We used a stratified *k*=10-fold cross-validation with *n*=100 iterations for estimating the misclassification (test) error (Alpaydin, 2010; Mitchell, 1997). Yet, the Kappa statistic measures the agreement of prediction with the true class. A value of Kappa equals to 1.0 signifies complete agreement. Moreover, sensitivity analysis is a method for identifying the “cause-and-effect” relationship between the inputs and outputs of a prediction model. This method is often followed to rank the variables in terms of their importance. Finally, F-score is a measure of a test’s accuracy, and considers the precision and the recall of the test. In simple terms, high precision means that an algorithm returned substantially more relevant than irrelevant results, while high recall means that an algorithm returned most of the relevant results (Alpaydin, 2010; Mitchell, 1997). The F-score can be interpreted as a weighted average of the precision and recall. An F-score reaches its best value at 1 and worst score at 0 (Tan et al., 2005). We implemented the classification techniques in Weka 3.8.

5.6. Results

5.6.1. Structural and measurement model – Hypothesis testing

The results support the measurement model. Table 5-2 displays the items’ reliabilities (Cronbach’s alpha, C.R), AVE and factor loadings and confirms convergent validity for the latent constructs.

Table 5-3 presents the variables’ correlation matrix. In this table, the diagonal elements are the square root of the AVE of a construct. According to the Fornell-Larcker, the AVE of each latent construct should be higher than the construct’s highest squared correlation with any other latent construct. Thus, discriminant validity is also confirmed.

Table 5-2. Results for the Latent Constructs of the Measurement Model

Construct Items	Factor Loadings (>0.7) ^a	Cronbach's a (>0.7) ^a	C.R. (>0.7) ^a	AVE (>0.5) ^a
GE		0.83	0.89	0.74
GE1	0.855			
GE2	0.874			
GE3	0.842			
CERT		0.78	0.89	0.81
TCV	0.954			
TTV	0.840			
E		0.86	0.88	0.54
E1	0.613			
E2	0.707			
E3	0.865			
E4	0.658			
E5	0.725			
E6	0.634			
E7	0.823			
E8	0.608			
A		0.88	0.89	0.51
A1	0.701			
A2	0.762			
A3	0.700			
A4	0.564			
A5	0.771			
A6	0.731			
A7	0.705			
A8	0.744			
A9	0.675			
C		0.87	0.90	0.51
C1	0.737			
C2	0.645			
C3	0.781			
C4	0.782			
C5	0.620			
C6	0.692			
C7	0.763			
C8	0.674			
C9	0.648			
N		0.86	0.88	0.52
N1	0.727			
N2	0.683			
N3	0.713			
N4	0.724			
N5	0.667			
N6	0.740			
N7	0.629			
N8	0.770			
O		0.89	0.91	0.53
O1	0.688			
O2	0.787			
O3	0.686			
O4	0.655			
O5	0.791			
O6	0.651			
O7	0.552			
O8	0.791			
O9	0.766			
O10	0.713			
TTAC	1.000	1.00	1.00	1.00
TTAW	1.000	1.00	1.00	1.00
LP	1.000	1.00	1.00	1.00

^a Indicates an acceptable level of reliability and validity

GE: Goal-expectancy, **CERT:** Level of certainty, **A:** Agreeableness, **C:** Conscientiousness, **N:** Neuroticism, **O:** Openness, **TTAC:** Time to Answer Correctly, **TTAW:** Time to Answer Wrongly, **TCV:** Total Check Views, **TIT:** Idle Time

Table 5-3. Discriminant Validity for the Measurement Model

	1	2	3	4	5	6	7	8	9	10
1. GE	0.857									
2. CERT	0.252	0.901								
3. TTAC	0.390	0.240	1.000							
4. TTAW	-0.415	-0.177	-0.302	1.000						
5. E	0.512	0.155	0.227	-0.428	0.735					
6. A	0.364	0.161	0.042	-0.097	0.355	0.714				
7. C	0.407	0.342	0.385	-0.364	0.345	0.151	0.714			
8. N	-0.134	-0.216	-0.144	0.065	0.018	-0.008	-0.115	0.721		
9. O	0.245	-0.069	0.217	-0.236	0.553	0.237	0.275	-0.050	0.728	
10. LP	0.645	0.340	0.773	-0.561	0.394	0.152	0.552	-0.114	0.257	1.000

GE: Goal-expectancy, CERT: Level of certainty, A: Agreeableness, C: Conscientiousness, N: Neuroticism, O: Openness, TTAC: Time to Answer Correctly, TTAW: Time to Answer Wrongly, TCV: Total Check Views, TIT: Idle Time

A bootstrap procedure with 3000 resamples was used to test the statistical significance (*t-value*) of the path coefficients (β) in the model. Table 5-4 summarizes the results for the hypotheses.

Table 5-4. Hypothesis testing results

<i>Hypothesis</i>	<i>Path</i>	β	<i>t</i>	<i>P</i>	<i>Result</i>
H1	A→GE	0.203*	2.635	0.008	Support
H2	A→CERT	0.120	1.205	0.228	Not Support
H3	E→GE	0.415*	4.390	0.000	Support
H4	E→CERT	0.162	1.512	0.131	Not Support
H5	C→GE	0.249*	3.385	0.001	Support
H6	C→CERT	0.324*	3.659	0.000	Support
H7	N→GE	-0.116	1.303	0.193	Not Support
H8	N→CERT	-0.195*	2.107	0.035	Support
H9	O→GE	-0.107	0.967	0.333	Not Support
H10	O→CERT	-0.286*	2.210	0.027	Support

* $p < 0.05$.

GE: Goal-expectancy, CERT: Level of certainty, A: Agreeableness, C: Conscientiousness, N: Neuroticism, O: Openness

As seen from Table 5-4, extroversion (E) and agreeableness (A) have a significant direct positive effect on goal-expectancy (GE); conscientiousness (C) has a significant direct positive effect on both goal-expectancy (GE) and certainty (CERT); neuroticism (N) and openness (O) have a significant direct negative effect on certainty (CERT). Thus, six out of the ten initial hypotheses are supported.

The overall variance (R^2) and cross-validated predictive relevance (Q^2) explained by the proposed model for actual performance during testing (LP) are depicted in Table 5-5. According to these results, the suggested model explains almost the 73% of the variance in AP.

Table 5-5. R², Q² and Direct, Indirect and Total effects

<i>Dep. Variable</i>	<i>R²</i>	<i>Q²</i>	<i>Indep. Variables</i>	<i>Dir. effect</i>	<i>Indir. effect</i>	<i>Total effect</i>	<i>t-value</i>	<i>P-value</i>
LP	0.730	0.709	TTAC	0.638		0.639*	12.398	0.000
			TTAW	-0.346		-0.346*	5.669	0.000
			GE		0.361	0.361*	5.922	0.000
			CERT	0.125	0.124	0.249*	3.023	0.003
			A		0.103	0.103*	2.715	0.007
			E		0.190	0.190*	3.889	0.000
			C		0.171	0.171*	3.686	0.000
			N		-0.090	-0.090	2.027	0.043
TTAC	0.152	0.137	GE	0.351		0.351*	4.551	0.000
			CERT	0.152		0.152	1.636	0.102
			A		0.090	0.090*	2.586	0.010
			E		0.170	0.170*	3.767	0.000
			C		0.137	0.137*	3.128	0.002
			N		-0.070	-0.070	1.772	0.077
			O		-0.081	-0.081	1.509	0.131
			TTAW	0.173	0.160	GE	-0.396	
CERT	-0.077					-0.077	0.761	0.447
A		-0.090				-0.090*	2.442	0.015
E		-0.177				-0.177*	3.232	0.001
C		-0.124				-0.124*	2.629	0.009
N		0.061				0.061	1.415	0.157
O		0.064				0.064	1.168	0.243

* p<0.05

GE: Goal-expectancy, **CERT:** Level of certainty, **A:** Agreeableness, **C:** Conscientiousness, **N:** Neuroticism, **O:** Openness, **TTAC:** Time to Answer Correctly, **TTAW:** Time to Answer Wrongly

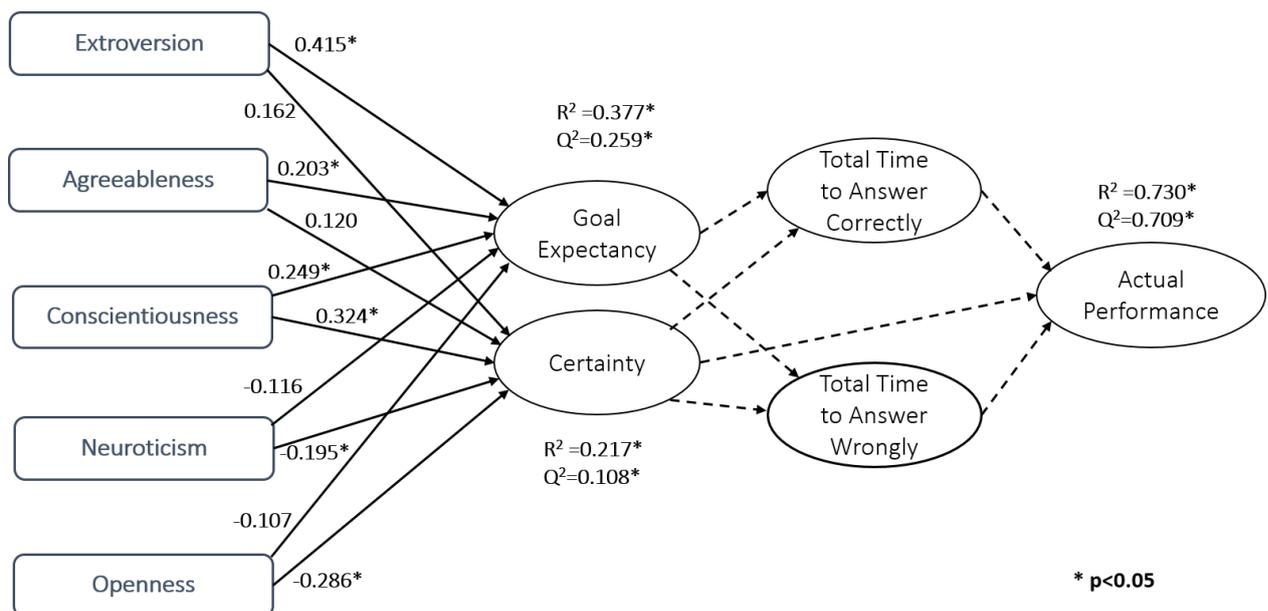


Figure 5-4. Path coefficients of the research model and overall variance (R²).

Moreover, and since GE and CERT have been found to directly impact total time to answer correctly (TTAC) and total time to answer wrongly (TTAW), Table 5-5 also displays the indirect effects of personality traits on the time-based variables (TTAC, TTAW), due to their relation to GE and CERT. These results are also summarized in Figure 5-4. This figure illustrates the path coefficients for the initial hypotheses of the research model.

5.6.2. Classification Results

Table 5-6 outlines the SLA methods that we applied on the input data, the number of classes being predicted (i.e., the different categories of students' performance results), the overall accuracy of the prediction (for training and testing respectively) together with the respective sample sizes (90% for training and 10% for testing for all SLA methods), and the tool used during the analysis.

Table 5-6. A summary of the classification approach

SLA used	# of classes predicted	Sample size	Accuracy of prediction	Simulation tool used
SVM, NB, RF	5-class	112 samples in total 101 for training 11 for testing	100% for training 80% for testing	Weka 3.8

The initial raw log file contained a sample of the 9 features to be used in this study (i.e., TTAC, TTAW, GE, CERT, A, E, C, N, O). The structural and measurement model evaluation conducted in the previous stage showed that some of these features were not statistically significant for prediction purposes. These features were O and N, and therefore, we removed these attributes. Moreover, prior to rejecting them, we confirmed that they were “noisy” by using feature subset selection. Performing feature selection reduces overfitting, improves accuracy, and reduces training time (Guyon & Elisseeff, 2003). In this process, algorithms search for a subset of predictors that optimally model measured responses, based on constraints such as required or excluded features and the size of the subset. In this study, we ranked the 9 attributes from most to least informative using the Attribute Selection method of Weka: a) the attribute evaluator assesses the attribute subsets, and b) the search method searches the space of possible subsets.

Figures 5-5a, 5-5b, 5-5c illustrates the results from the exploratory analysis of the initial dataset. In particular, Figure 5-5a displays the time-management variables (i.e. TTAC vs. TTAW), while Figure 5-5b shows GE vs. TTAC and Figure 5-5c represents CERT vs. TTAC for each target class (C1, C2, C3, C4 and C5).

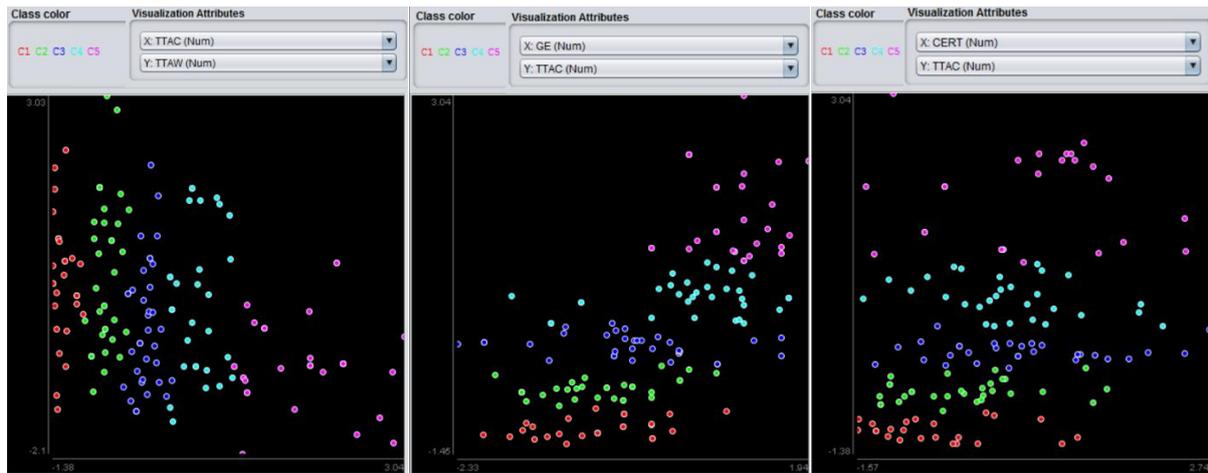


Figure 5-5. Graphical exploratory analysis on classes' characteristics: (a) the five classes according to their time-spent, (b) the five classes according to goal-expectancy, and (c) the five classes according to their level of certainty

Table 5-7 presents the performance results (resubstitution error, true test error, Kappa statistic, sensitivity, and F-score) for the four methods used to develop a classification model in this study with seven features and with testing sample size 10% of the initial dataset.

Table 5-7. Performance metrics for cross-validation 10% with seven features

<i>Test Set Size</i>	<i> cvpartition = 10% (k-fold=10)</i>			
Classifier	SVM	NB	RF	JCBA
Resub Error	0.29	0.30	0.22	0.34
True Test Error*	0.22	0.24	0.20	0.26
Kappa Statistic	0.65	0.63	0.68	0.45
Sensitivity	0.83	0.82	0.87	0.63
F-score	0.78	0.75	0.80	0.71

*True test error=cross-validation error

These results demonstrate that all methods achieve high classification performance, since the true test error varies from 0.20 (RF method) to 0.26 (JCBA method). Further to that, the sensitivity measure, the F-score and the Kappa statistic are also high (0.63-0.87, 0.71-0.80 and 0.45-0.68 respectively). Moreover, from this table it becomes apparent that the RF method provides better classification results compared to the other methods, while the SVM method also achieves satisfactory results.

5.7. Discussion

5.7.1. Which is the effect of the five personality factors on time-spent behavior during CBT (Hypotheses H1 to H10)?

A timeless research question regarding learners' behavior in different learning contexts, concerns the impact of personality aspects (traits or facets) on time-management and achievement. However, the search in literature yielded inconclusive results regarding the effects of personality traits on how students use their time during learning activities, and how efficiently they allocate their time in relation to the learning outcomes and performance. The first aim of this

study – expressed in RQ1 – was to explore the use of time-driven assessment analytics methodology with BFI towards explaining achievement behavior during CBT in terms of personality and response times on task-solving. The innovation and contribution of our approach is that it exploits assessment analytics capabilities for shedding light into examinees' interactions during testing. In particular, we adopted the data-driven TLA methodology, which is about gaining insight into students' goal expectations and carefulness during assessment, as well as explaining how they behave during the activity based on their response times (Papamitsiou et al., 2014). Previous results had provided strong indications that the temporal interpretation of students' engagement in activity could be used for predicting their progress. As shown in Table 5-7, the overall prediction accuracy of the suggested approach in this study is 80%, which is statistically significant. The data analysis revealed some interesting findings.

First, the effect of agreeableness on goal expectancy (i.e., on student's goal orientation and perception of preparation) is strong ($\beta=0.203$, $t=2.635$, $p=0.008$), and confirms our first hypothesis ($H1$). This means that agreeable students tend to stay focused on their assessment orientation. This finding is also in line with Bipp, Steinmayr and Spinath (2008), and adds evidence to prior claims by Terzis, Moridis and Economides (2012) that agreeableness would have a positive effect on goal-expectancy, who, however, did not verify that hypothesis. Moreover, agreeableness was found to be a strong indirect determinant of both types of response-times ($\beta=0.090$, $t=2.586$, $p=0.010$ on TTAC, and $\beta= -0.090$, $t=2.442$, $p=0.015$ on TTAW respectively). This finding indicates that agreeable examinees exert effort (in terms of time-spent) on dealing with the assessment tasks and constitutes additional evidence towards clarifying the “vague” relation of this personality trait with time-management (Claessens et al., 2007). In addition, agreeableness is also associated with social desirability (Digman, 1997), which has also been shown to be negatively correlated with performance ratings, as assessment becomes more learning-orientated and less socially-influenced (Murphy & Cleveland, 1995). However, in our study, agreeableness was found to be a strong positive indirect determinant of actual performance ($\beta=0.103$, $t=2.715$, $p=0.007$). Yet, our findings also verified that agreeableness has a positive effect on the student's level of certainty (i.e., how certain the student wants to be when answering a question), but the effect was not statistically significant and the second hypothesis ($H2$) was not supported.

Moreover, although prior studies (Terzis et al., 2012) did not verify that extroversion has a positive effect on goal expectancy, in our case, this hypothesis ($H3$) was also confirmed ($\beta=0.415$, $t=4.390$, $p=0.000$). This finding indicates that extrovert students tend to set active skill/knowledge acquisition goals and believe that they are prepared enough to achieve them. This also complies with previous results that demonstrated that extraversion is significantly related to motivational concepts such as goal-setting and self-efficacy (Judge & Ilies, 2002). Going a step beyond, this finding could suggest that students with an extrovert behavioral aspect designate their goal orientations more precisely. As a result, they seem to be more self-aware regarding their perceptions of preparation. Reinforcing de Raad's and Schouwenburg's (1996)

findings that highly extrovert students will perform better academically – because of a positive attitude leading to their desire to learn and understand – our results also correlated strongly and positively extraversion with actual performance ($\beta=0.190$, $t=3.889$, $p=0.000$). Furthermore, extraversion was found a strong positive indirect determinant of response times on correctly answered questions (TTAC, $\beta=0.170$, $t=3.767$, $p=0.000$) and a strong negative indirect determinant of time-spent on wrongly answered questions (TTAW, $\beta=-0.177$, $t=3.232$, $p=0.001$). This mean that, due to their increased perception of preparation, extrovert students are more likely to answer correctly and allocate time on TTAC. In addition, regarding the impact of extroversion on students' level of certainty, we initially assumed that extroverts are expected to act more impulsively and spontaneously, without straggling for gaining high level of certainty. This assumption derived from prior research results (Boroujeni et al., 2015) However our hypothesis on the negative correlation between extroversion and certainty ($H4$) was not supported. On the contrary, a positive effect was detected, although it was not statistically significant ($\beta=0.162$, $t=1.512$, $p=0.131$).

Another finding was that conscientiousness has a strong direct positive impact on both goal expectancy and level of certainty ($\beta=0.249$, $t=3.385$, $p=0.001$ and $\beta=0.324$, $t=3.659$, $p=0.000$ respectively). Conscientiousness is related to responsibility towards goal achievement and describes students that think before acting. Consequently, we assumed that this trait is expected to have a positive effect on both behavioral parameters (GE and CERT). In fact, by definition, level of certainty reflects the level of student's cautiousness when dealing with assessment tasks. As such, the strong relationship of conscientiousness with certainty was a priori valid. Moreover, research has also linked conscientiousness to goal commitment and self-set goal setting (Gellatly, 1996). In our study, both hypotheses ($H5$ and $H6$) were supported from the analysis on the collected data. This finding suggests that conscientious students will spent more time to view the questions again and again before saving an answer, trying to assure that they will submit the correct answer. In addition, due to their strong sense of purpose, conscientious students demonstrate a deeper engagement with the assessment activity in terms of time. Moreover, the impact of certainty on the response times variables is also strong ($\beta=0.137$, $t=3.128$, $p=0.002$ on TTAC, and $\beta=-0.124$, $t=2.629$, $p=0.009$ on TTAW). This finding confirms once more that time-on-task fully mediates the conscientiousness–performance relationship (Biderman et al., 2008; Tabak et al., 2009). Another interpretation of this finding is that conscientious students manage their time more efficiently and aggregate more time on correctly answered questions. Moreover, our results are in line with Conard (2006), who correlated this characteristic to school and college grades. Precisely, the data analysis shown a strong positive effect of conscientiousness on actual performance ($\beta=0.171$, $t=3.686$, $p=0.000$).

On the contrary, our results indicate that neuroticism only marginally is correlated with actual performance ($\beta=-0.090$, $t=2.027$, $p=0.043$). One would expect this negative relationship because of neurotics' overall negative dispositions, anxiety during the exams and poor self-

regulation (Kanfer & Heggstad, 1997). Furthermore, according to Van Hoye & Lootens (2013) highly neurotic individuals is less likely to use time management strategies. This is also reflected in the negative effect of neuroticism on the total response time on correctly answered questions ($\beta=-0.070$, $t=1.772$, $p=0.077$) and its positive impact on aggregated time on wrongly answered questions ($\beta=0.061$, $t=1.415$, $p=0.157$), although these relationships were not found to be statistically significant. The only strong correlation detected between neuroticism and the explored variables was that of the level of certainty. More specifically, neuroticism has a strong negative effect on certainty ($\beta=-0.195$, $t=2.107$, $p=0.035$). This result confirms our hypothesis regarding this relationship (*H8*) and is in line with Kanfer & Heggstad (1997). Moreover, neuroticism affects negatively a student's goal expectancy (Judge & Ilies, 2002), but in our study, this hypothesis (*H7*) was not strongly supported ($\beta=-0.116$, $t=1.303$, $p=0.193$).

Finally, openness to experience did not relate to goal orientation ($\beta=-0.107$, $t=0.967$, $p=0.333$). In addition, in contrast to our initial assumptions on a positive association between these two variables, a negative relation came up. Perhaps this hypothesis (*H9*) was not supported because, in the CBT context used in this study, students high on openness to experience did not perceive the task-related assessment to be creatively stimulating. On the other hand, Chamorro-Premuzic, Furnham, and Lewis (2007) suggested that highly open minded students are more likely to inquire knowledge and make considerations rather than maintain their level of certainty. This claim is explored under hypothesis *H10*, and is supported by our data analysis ($\beta=-0.286$, $t=2.210$, $p=0.027$). Likewise, our findings indicate weak correlations of openness to experience with both response times variables ($\beta=-0.081$, $t=1.509$, $p=0.131$ on TTAC, and $\beta=0.064$, $t=1.168$, $p=0.243$ on TTAW). These findings align with Van Hoye & Lootens's (2013) claim that individuals high on openness to experience find it difficult to manage their time effectively to complete tasks. Consequently, it is expected that such a personality will exhibit moderate achievement behavior in time-limited, task-oriented testing activities, although the advanced critical thinking and deep learning skills. This is reflected on our finding that openness to experience has statistically insignificant effect on actual performance ($\beta=-0.110$, $t=1.850$, $p=0.064$).

5.7.2. How accurately can we classify the students during testing according to their personality traits and behavior expressed in terms of response-times?

Differences in learners' behavior during assessment have a deep impact on their level of achievement. Compiling learners' behavior in CBA processes and creating the corresponding behavioral models is a primary educational research objective. The emergence of assessment analytics along with the recent trend to exploit students' time-spent habits, urged our interest on associating personality traits with response-times for modelling examinees' behavior during CBT. The second goal of this study – stated as RQ2 – was to explore student-generated temporal trace data and personality aspects for modelling students' behavior during CBT according to the students' test score. Our goal was to seamlessly identify the students' time-spent behavioral patterns in order to dynamically shape the respective models. The motivation for our experimentation was based on significant results reported in previous studies that analysed

temporal parameters for user modelling (Papamitsiou et al., 2016, 2014; Shih et al., 2008; Xiong et al., 2011).

Our findings verify formerly reported results (Belk et al., 2014; Shih et al., 2008) regarding the capability of temporal data to represent, describe and model the students' behavior. In particular, our findings indicate that TTAC and TTAW in combination with goal expectancy and level of certainty could satisfactorily be used for classification of students during CBT. The low misclassification rates are indicative of the accuracy of the proposed method (True Test Error: 0.20-0.24). Further to that, from table 5-9 it becomes apparent that the ensemble Random Forest method provided the most accurate classification results compared to the other methods.

The TTAC and TTAW variables seem to be highly related to achievement. In this case, students in classes C5 and C4 ("high achievers" and "struggling achievers", respectively) obtain the best final marks and exhibit higher time-based commitment to the task-solving activity. These students are classified as highly goal-oriented and with high levels of certainty. In particular C5 members are marked with the higher response-times on TTAC and the lower time-spent on TTAW. A bit lower is the range of TTAC values for C4 members, who however, appear to exhibit higher total time to review the questions (which is a factor loading on the level of certainty). For both classes, GE is reported as high. The major difference between these two classes is identified in the TTAW factor, which for C4 members appears to be higher. As such, this variable could be used for distinguishing the two classes.

Similarly, students in classes C1 and C2 ("low achievers" and "careless achievers", respectively) are identified by their medium-low achievement, and exhibit minimum engagement with the testing items in terms of time-spent, denoting low goal-orientation and low levels of confidence. More precisely, students in C1 aggregate the higher response times on TTAW, with the lower levels of goal expectancy. Moreover, members of C2 score high in TTAW as well, but the value range for TTAC is a bit higher than the respective for C1 students. In this case, TTAC is the factor that could be used to distinguish low achievers from careless achievers. Nevertheless, according to their scores, totally unconcerned students seem to belong to C1 class, while in C2 are categorized the students that try a bit more, but still are careless and disengaged. For C1 students, level of certainty gets its lower values, and for C2 participants it is also very low.

Regarding their personality factors, students from both C4 and C5 classes are categorized as extroverts, conscientious and agreeable. Minor difference between these two classes are detected in the other two personality traits (i.e. neuroticism and openness), with the C4 class students to appear as more neurotic and more open to experience compared to those in C5. However, as stated in section 5.2, these two features were considered only during exploratory analysis, and excluded from the classification process because of their limited prediction accuracy.

Conversely, the results for C1 and C2 classes concerning the dominant personality traits of the less achieving students were as expected: both classes appear to have introvert members,

who are less cautious and more disagreeable. Students from both classes also appear to be more neurotic, but regarding the openness to experience factor, the result from the exploratory analysis was inconclusive.

Finally, the members of class C3 (“neutral achievers”) exhibit the most unclear behavior regarding all variables. The aggregated response times on TTAC and TTAW are similar (this is an expected attribute of this class), their goal-expectancy varies from high to low, and the same stands for their level of certainty as well. As such, these factors are only moderate predictors for medium achieving students. The personality factors for the members of this class also present mixed results. This is probably the reason that increases the misclassification error during their assignment to one of the classes. However, even in this case the misclassification rates remain low for all classifiers explored in this study.

According to these findings, most of the initial assumptions summarized in Table 5-3 (please see section 5.4.2), are confirmed. However, the assumptions that were not confirmed are reconsidered and synopsised as follows (Table 5-8).

Table 5-8. Achievers’ classes and their characteristics (reconsideration)

C1: Low Achiever	C2: Careless Achiever	C3: Neutral Achiever	C4: Struggling Achiever	C5: High Achiever
TTAC (--)	TTAC (-)	TTAC (+-)	TTAC (+)	TTAC (++)
TTAW (++)	TTAW (+)	TTAW (-+)	TTAW (-)	TTAW (--)
GE (--)	GE (-)	GE (+-)	GE (+)	GE (++)
CERT (--)	CERT (-)	CERT (-+)	CERT (+)	CERT (++)
E (--)	E (-)	E (+-)	E (+)	E (++)
A (--)	A (-)	A (+-)	A (+)	A (++)
C (--)	C (-)	C (+-)	C (+)	C (++)
N (++)	N (+)	N (-+)	N (--)	N (-)
O (?)	O (?)	O (+-)	O (++)	O (+)

In this table, modifications on the assumptions (compared to Table 5-1) are marked with bold fonts (in the shaded cells), while the inconclusive results for classes C1 and C2 regarding the personality trait of openness to experience are indicated with the question mark sign.

5.8. Implications

The findings presented in this chapter, are interesting in two different senses: a) personality factors are significant predictors in the temporal estimation of students’ performance, and b) the temporal factors that imply students’ engagement in activities should be further explored regarding their added value towards modelling test-takers and dynamically reshaping the respective models.

Consequently, the arising question is: how we could exploit and utilize these findings towards developing credible assessment systems, applications or services? In this section we discuss about possible implications of the findings.

5.8.1. Reclaiming personality factors: Implications for examinees

Development of automated, data-driven, adaptive CBA environment is expected to provide students with opportunities to demonstrate their developing abilities, support self-regulated learning and help them evaluate and adjust their assessment strategies to improve performance.

Our findings revealed that extroverts seem to be more self-aware regarding their perceptions of preparation (*H3*), and that agreeable students tend to stay focused on their assessment orientation (*H1*). A possible implication of these two findings would be to appropriately scaffold the agreeable and extrovert students during CBA through a real-time visualization (for example) that associates time-spent with goal-achievement. Similarly, conscientious students demonstrate a deeper engagement with the assessment activity (*H5*). For these students, the CBA environment could provide analytics on how they progress on each assessment item (or task) compared to the rest of the class or compared to their own previous states. Yet, conscientious students will spend more time to view the questions again and again before saving an answer, trying to assure that they will submit the correct answer. This means that conscientious students try to increase their level of certainty (*H6*). For this purpose, an adaptive (or intelligent) CBA environment could timely prompt a hint to the cautious students, when the system detects that these students are struggling to gain their confidence regarding the correct answer. Furthermore, another finding was that neurotics' overall negative dispositions, anxiety during the exams and poor self-regulation affects negatively their certainty and performance (*H7*). In this case, the CBA environment could supply the neurotic students with suitable emotional feedback in order to balance the negative feelings that the assessment itself causes to them, and to increase their self-confidence and certainty. The form of the emotional feedback is an open issue to be further explored. Yet, individuals high on openness to experience find it difficult to manage their time effectively to complete tasks (*H10*). That is probably happening because they did not perceive the task-related exam to be creatively stimulating. For these students, different forms of assessment tasks should be made available by the CBA environment. For example, time-spent could be tracked to measure the duration of solving/ implementing sub-activities or sub-tasks in the context of project-based learning, or the duration of studying and exercising with learning modules during inquiry-based learning, etc. In that way, the open to experience students could improve their time-management skills and their overall performance.

5.8.2. Enhancing student models: Implications for systems developers

It is generally acknowledged that it is important for systems developers to identify the behavioral parameters that could be used for fully adapting the CBA system, application or service (in general, environment) to the learners' level of ability/expertise or for providing personalized feedback during the assessment process.

Based on the findings, we suggest that one can identify a set of functional temporal (and/or behavioral) factors that could constitute the core components of a CBA system's architecture. For example, TTAC, TTAW, GE, CERT and personality traits (i.e., E, A, and C) are only indicative variables that could be embedded into a testing system in order to model the test-

takers and to guide adaptation and personalization of test. Systems like that would aim at personalizing the deliverable service according to their user's model. For example, such a service could be the recommendation of the next most appropriate task according to the student's model and detected level of expertise (based on the corresponding timely predicted performance). In this case, the system should be "trained" in order to "recognize" and model its current users based on their temporal and behavioral data. Then, it should "choose" the appropriate task (among the collection of tasks from an item bank) that best corresponds to the needs and meets the abilities of the user, in order to improve the expected outcome. Finally, the system should inform the users about their progress and either suggest the selected task (as a CAT system) or allow the users to make their own choice of the next task (as a CBT system).

5.9. Conclusions and Future Work

The present study attempted to shed light to the "vague" landscape of the impact of personality traits on time-management during testing. The purpose of this study was to contribute towards exploiting time-driven assessment analytics methods with the Big Five Inventory for deeper understanding the examinees' time-spent behavior on task-solving during CBT according to the five personality traits and their achievement level. A second goal was to investigate the assessment analytics capabilities on classifying students and contribute to generating student models enhanced with temporal behavior attributes to guide personalization of testing services. Thus, the research questions were twofold:

RQ1: *Which is the effect of the five personality factors on time-spent behavior during CBT?*

RQ2: *How accurately can we classify the students during testing according to their personality traits and behavior expressed in terms of response-times?*

In order to answer on these research questions (RQ1, RQ2) we formed 10 hypotheses related to the personality traits and examined their relationships to the other temporal and/or behavioral factors of the TLA model. Moreover, 5 additional assumptions were developed regarding the configuration of the student models to explore for classification purposes. Towards estimating the validity of our hypotheses, we carried out a case study with a modified version of the LAERS assessment environment. One hundred and twelve (112) undergraduate students from a Greek University enrolled in a CBT experimental procedure. Partial Least Squares (PLS) was used to explore the relationships between the included factors and evaluate the structural and measurement model, and three Supervised Learning Classification algorithms were used to compare the obtained classification results based on students' performance, i.e. using as class labels the students' performance score classes.

Regarding the first research question (RQ1), results from this study are encouraging and provide strong indications that the collected real-time actual data (TTAC, TTAW, CERT, AP) and the self-reported perceptions (GE, personality traits) are strongly correlated. More precisely, it was found that examinees' extraversion, agreeableness and conscientiousness indirectly and positively affect examinees' total time to answer correctly and negatively affect their total time to

answer wrongly respectively. These factors were also significant indirect predictors of actual performance as well. Moreover, it was found that extraversion and agreeableness have a direct strong positive impact on goal-expectancy, conscientiousness directly and positively affects examinee goal-expectancy and level of certainty, and examinees' neuroticism and openness have a direct negative effect on level of certainty.

Regarding the second research question (RQ2), it was also found that all methods explored here (i.e. SVM, NB, RandomForest and JCBA) provide significant classification results, but the ensemble RandomForest algorithm classifies examinees according to their time-spent more accurately. This finding confirms and complies with previous research results that suggest the use of time-dependent factors for enhancing student models. Moreover, this study goes one step beyond by introducing the characteristics of each one of the five identified classes.

The approach suggested in this chapter was applied on a dataset collected during a testing procedure in the context of mid-term exams. The nature of the data collected (time-based parameters) and the general-purpose methodology followed for the analysis of these data, render this approach replicable and/or transferable to other contexts, and eliminate the restriction of using it only during testing. The temporal factors are not contextualized to the LAERS assessment environment, but a similar tracker could be embedded in any adaptive learning system. For example, time-related parameters (time-spent) could be tracked to measure the duration of solving/implementing sub-activities or sub-tasks in the context of project-based learning, or to measure the duration of studying and exercising with learning modules during inquiry-based learning, etc., along with the number of repeating the intermediate, facilitating steps (e.g. watch educational videos, open/use educational resources, participate in discussions, etc.).

However, these findings need to be validated by additional experimentation and bigger participant samples. Further investigation regarding the inconclusive personality traits (neuroticism and openness) is also required. In addition, other personal factors, such as gender or learning styles, should be examined. Regarding the investigation of the further improvement of the classification accuracy due to the inclusion of these features and whether they contribute to providing better classification results, it is an open future research question to be addressed, and it is beyond the goals of the present study. For this purpose, additional data (not available in the current study – e.g., prior grades, learning preferences, socio-demographic characteristics, etc. – yet extensively studied for purposes of modelling students' achievement behavior) should be treated as the alternative feature space. As a next step, we envisage creating the learner model simultaneously, while the student takes the test, in a stream mining fashion, which would enrich the profile modeling with a notion of dynamics, allowing for adaptive question sequencing.

Chapter 6 : “Current-awareness” – Enhancing the learner models with temporal dynamics

“The Afternoon knows what the Morning never suspected.”

Robert Lee Frost

Towards currently-aware learner models

6.1. Introduction

Nowadays, most of the human learning occurs online, whereas – even in traditional classrooms – learning practices of the past tend to become obsolete (Moore, Dickson-Deane, & Galyen, 2011; Song & Lee, 2014). In order to support individuals to become better learners in this context, it is necessary to provide them quality online learning services, personalized and adapted to the individual’s learning needs and facilitating the individual’s learning goals (Brusilovsky & Peylo, 2003; Economides, 2009a). Therefore, compiling learners’ interactions in online learning/assessment processes and creating robust, accurate and easy-to-use learner models accordingly, is a prerequisite (Brusilovsky et al., 2016).

Learning in online conditions has a strong inherent temporal dimension: it is time-framed and it carries information about the current state of the learners, at the time that learning occurs (Azevedo, 2014; Xie, Zheng, & Zhang, 2018). Online learning environments capture the learning experience in real-time, and current learner modeling approaches diagnose the learner states in real-time. However, most of these systems do not exploit the changes detected in learner data to refit the parameters of the learner models in run-time as well, based on the new data, due to several constraints. For example, in Bayesian Knowledge Tracing (Corbett & Anderson, 1994), the model parameters fitting is typically done offline due to the volume of real-data that is computationally demanding and impose a single pass only. In another example, in order to construct and refit a new model, Bayesian Networks (Conati, Gertner, & VanLehn, 2002) require an update that considers all observed data, yet performing this step for each additional data point that arrives at each time-point is not feasible. Moreover, a constraint in Performance Factors Analysis (Pavlik, Cen, & Koedinger, 2009) logistic modeling approach is that all observations are treated as equally important, regardless of when they happened.

Despite these obstacles, the tremendous increase in online learning has signaled the momentum of processing the learner data and detecting changes in run-time, and has enforced the demand for refitting the parameters of the learner models on-the-fly, accordingly. In order to be timely updated and valid (a) the learner models should be continuously aware (at any time-point) of the learners’ traits and of the *changes* in their learning states, as they evolve over time; (b) all the information from the learners’ interactions arrives continuously over time and should be processed with satisfactory efficiency and scalability, before being integrated into the learner models: what is detected in the data should be compared and aligned with the measurable

learning performance *at that time*. These issues highlight a research gap in the area of model construction and run-time utilization, which the typical artificial intelligence approaches cannot address, due to the need to cope with *challenges of real-time computing*.

In view of the above, the present contribution aims at opening the discussion towards shaping and refitting “*currently-aware*” learner models based on the continuously arriving data in run-time and elaborating on the importance of integrating the temporal dimension in learner modeling. “*Currently-aware*” means that the models are time-framed, they carry information about the learners’ current learning state, they allow for refitting the parameters and re-defining the learning state in run-time (i.e., when changes occur), and they allow for re-estimating the learners’ states, as well.

More precisely, the problem this study attempts to address is *how* to incorporate the temporality of learning into the learner models, and to refit their parameters during an online learning procedure; the models need to be frequently revised with timely gathered and configured information (*when*), generated from the learners’ interactions (*what*), in a way that they are useful and easy-to-use (*why*). The core idea is to treat the arriving learners’ interactions data as an *adaptive data-stream*, and to classify learners in run-time, according to a set of time-varying predictors (observational data). The data-stream perspective takes what has been a static view of a problem, and adds a strong temporal dimension to it (Babcock, Babu, Datar, Motwani, & Widom, 2002; Kleinberg, 2016): the learners shall be continuously re-classified to a changing (re-defined) mastery level for each one of the multiple skills being measured, according to the progress achieved for that skill. The innovation of this work is that it considers the changes in data at each time-point, and suggests a method that efficiently identifies in run-time the changes in learners’ traits, and refits the models’ parameters accordingly, to reflect the progress in multiple skill/knowledge mastery; the method successively incorporates this information in the learner models according to the learners’ changes in skill/knowledge acquisition, over time.

The research described in this chapter was carried out as an exploratory study to evaluate the effectiveness of the proposed approach, and is contextualized in web-based self-assessment conditions; self-assessment trains learners to self-regulate their motivation and behavior, and fosters reflection on their knowledge/skills progress, resulting in understanding themselves as learners (Nicol & Macfarlane-Dick, 2006; Sluijsmans, Dochy, & Moerkerke, 1998). Diagnosing learners’ knowledge and adapting the self-assessment to the individual’s needs by considering their learning state in run-time is essential.

6.2. Learner modeling: Related work - Motivation of the research & research question

6.2.1. A brief overview of learner modeling issues and approaches

Learner modeling has attracted increased interest as a research topic because it is fundamental in all adaptive learning systems. In general, learner modeling can be defined as the process of information extraction from different data sources and its compilation into profile

representations of learners' knowledge level (on a specific domain or topic), affective states, cognitive and meta-cognitive skills (McCalla, 1992). Essentially, the learner model is an *estimation* of the current state of the learner, based on the available observational data from her interactions with a learning environment. It summarizes multiple learner's traits, either static (e.g., gender) or dynamic. The later ones are inferred from the learner's performance data during the learning activities, and are used for decision making (e.g., intervening adaptations). The most popular dynamic traits include prior and acquired domain knowledge (Corbett & Anderson, 1994; Self, 1990), learning strategies, preferences and styles (Peña-Ayala, 2014), emotions and affective states (Brawner & Gonzalez, 2016; Calvo & D'Mello, 2010; Conati & Maclaren, 2009; Moridis & Economides, 2009b; Munshi et al., 2018), meta-cognitive skills (Aleven, McLaren, Roll, & Koedinger, 2006; Segedy, Kinnebrew, & Biswas, 2011), as well as critical and analytical thinking (Mitrovic & Martin, 2002). In a serious review, Desmarais & Baker (2012) presented a generalized view of the landscape on learner modeling by identifying key areas: knowledge, affect, motivation, disengagement, self-regulation, meta-cognition, group modeling and long-term modeling. A detailed survey on overall learner modeling approaches can be found in Chrysafiadi and Virvou (2013b); the review elaborates on the learners' traits considered in the selected learner modeling approaches, and how the learner models can be used for adaptation and personalisation purposes.

6.2.2. Modeling learners' knowledge using knowledge tracing and logistic models

The approach proposed in the present chapter targets at modeling learners' knowledge/skills. Therefore, the review of relevant research is bounded in this domain. The basic goal of skills/knowledge modeling is to estimate the learners' current knowledge state and to predict future performance based on data about past performance. In a comprehensive review of the key learner modeling methods that have contributed to the establishment and the success of Intelligent Tutoring Systems, Pavlik, Brawner, Olney, & Mitrovic (2013) highlighted the role of knowledge tracing (e.g., Baker, Corbett, & Aleven, 2008; Corbett & Anderson, 1994; Käser, Klingler, Schwing, & Gross, 2017; Yudelso, Koedinger, & Gordon, 2013) constraint-based models (Mitrovic, 2012; Mitrovic, Ohlsson, & Barrow, 2013; Ohlsson, 1994) and knowledge space models (Craig et al., 2013; X. Hu et al., 2012) on modeling learners' knowledge/skill mastery. It has been argued that generalized and universal learner models are not only very hard to construct, but also they are out-of-scope, because learner modeling is a highly context-aware procedure: contextual information (e.g., the type of knowledge being modeled, the available input data, the purpose of the model) have to be considered (Pelánek, 2017). Overall, two are the prevailing approaches adopted for modeling learner knowledge: (a) knowledge tracing, and (b) logistic models.

The topic of knowledge tracing has been heavily studied within the intelligent tutoring community. Bayesian Knowledge Tracing (BKT) is a standard for modeling learners' knowledge, skills and learning over time, assuming that learner knowledge is represented as a set of binary variables; the core assumption is that the learner either possesses or does not possess a skill

(Corbett & Anderson, 1994). The classic BKT model is in fact a Hidden Markov Model, where learner knowledge is a hidden variable and learner performance is an observed variable. Classic BKT uses skill-oriented, learner-specific parameters to infer the learner's knowledge state from performance, and to estimate the probability of the learner to have reached mastery of a concept/skill. There are two types of parameters: transition probabilities and emission probabilities. The emission probabilities include the probability of responding to a task incorrectly by mistakenly applying a known skill (p_S), and the probability of responding correctly by applying an unknown skill required in that task (p_G). The transition probabilities include the probability of a skill transitioning from the unlearned to learned state after an opportunity to apply it (p_T), and the probability of the learner knowing the skill beforehand (p_{L_0}). The probabilities across each of these binary variables are updated as a learner answers tasks of a given concept correctly or wrongly. Parameter fitting for the learner parameters (p_{L_0} , p_T , p_S , p_G) is typically done (offline) using expectation-maximization, a stochastic gradient descent or brute force grid search.

Overall, BKT allows for efficient parameter learning and accurate inference. Many extensions and variants to BKT have been developed and incorporate individual characteristics (Pardos & Heffernan, 2010; Yudelson et al., 2013), forgetting mechanisms (i.e., a function of the activation of an item for restudying) (Khajah, Lindsey, & Mozer, 2016; Nedungadi & Remya, 2015), item difficulty (Gowda, Rowe, de Baker, Chi, & Koedinger, 2011; Pardos & Heffernan, 2011), time between attempts (Qiu, Qi, Lu, Pardos, & Heffernan, 2011), response-times and instructional interventions (Lin et al., 2016), etc. BKT is a popular and useful method, yet researchers enumerate some problems, synopsisized as follows:

1. The binary response data used to model the discrete transition from unlearned to learned state might be appropriate for modeling understanding and sense making processes, yet only for fine-grained skills and concepts (i.e., that can be broken-down in simpler skills and concepts), thus imposing a limit on the kinds of learning tasks that can be modeled.
2. Estimation of skill-oriented knowledge only is an incomplete assumption (Yudelson et al., 2013). To resolve this, it was suggested to estimate an individualized weight for each learner and then adjust the model's generated parameters accordingly. However, a weight can be estimated only off-line, after all data is obtained, making the approach a no run-time solution (Piech et al., 2015).
3. There may be multiple sets of parameters that fit the data equally well (Beck & Chang, 2007), making interpretation difficult. Optimization techniques used for fitting the model parameters (e.g., expectation-maximization, brute force grid search) suffer from this hitch. The learned parameters may also produce a model that fits the data well, but violates the assumptions of the approach (i.e., the transition from unknown to known), resulting in inappropriate pedagogical decisions if used in a real system (Baker et al., 2008). Hawkins, Heffernan, & Baker (2014) employed heuristics to estimate when students learn skills, and

then used these estimates to refit the four BKT parameters, lacking, however, in precision compared to the previous optimization techniques.

4. The classic BKT model is designed per skill, uses simple transition models to describe how each concept state evolves, and cannot capture the relationship between different concepts. Therefore, when multiple skills are required to solve a problem, it becomes difficult to decide to which skill a particular observation should belong to. As a result, although BKT can output the learners' mastery level of some predefined concepts, it lacks the ability to extract undefined concepts and model complex concept state transitions. To overcome this problem and capture more complex knowledge representations, dynamic Bayesian networks, ensembles of models, recurrent neural networks and feature-aware knowledge tracing have been proposed to model the dependencies between the different skills, and to jointly represent multiple skills in a single model (Gonzalez-Brenes, Huang, & Brusilovsky, 2014; Käser et al., 2017; Khajah et al., 2016; Pardos, Gowda, Baker, & Heffernan, 2012; Piech et al., 2015). These models achieve good predictive accuracy, but lack interpretability.
5. BKT miss out significant non-cognitive factors (e.g., motivation, goal-setting, self-efficacy, emotions) which influence the learners' behavior during learning and assessment, and whose impact is reflected in the results of learners' outcome (Gutman & Schoon, 2013).

Another major "family" of methods for forecasting learners' mastery of knowledge, estimating proficiency and modeling learning accordingly, is a class of logistic models. In this case, a continuous variable is employed to model a skill, and gradual change is used to model learning. These models typically include an item difficulty parameter and they use a logistic function for mapping a difference between a skill and an item difficulty into the probability of a correct answer. The one parameter logistic model (Rasch model) is a typical example (Verhelst & Glas, 1995). Such models are intensively used in Item Response Theory (IRT); IRT is widely used to model the response of each learner of a given ability to each item in a test (W. J. van der Linden et al., 2010). IRT amounts to structured logistic regression, estimating latent quantities corresponding to the learner's ability and the items' attributes like difficulty (Johns, Mahadevan, & Woolf, 2006; Wilson, Karklin, Han, & Ekanadham, 2016). It is typically used in testing procedures and, unlike BKT, it is not considered as an appropriate method for modeling learning, since (usually) skills are not expected to change during a test. Moreover, although IRT models are attractive because they have a small number of parameters, however, these models are *static* (Johns et al., 2006), in terms that the employed item-variables (i.e., the difficulty of the items, the discrimination ability of the items, and the pseudo-guessing parameter for each item) are not dynamically refitted over time. Recent methods attempt to integrate IRT and BKT (e.g., Gonzalez-Brenes et al., 2014; Khajah, Huang, González-Brenes, Mozer, & Brusilovsky, 2014) and IRT extensions that allow for a dynamic change of skill have also been proposed (X. Wang, Berger, & Burdick, 2013).

Other typical logistic models employed in learner modeling include the Performance Factor Analysis (PFA) (Pavlik et al., 2009), the Additive Factors Model (AFM) (Cen, Koedinger, & Junker, 2006), Instructional Factors Analysis (IFA) (Chi, Koedinger, Gordon, Jordan, & Van Lehn, 2011), and the Elo rating system (Pelánek, 2016). These models infer changes in knowledge from sequential data and have been found to outperform BKT (Gong, Beck, & Heffernan, 2011). BKT and logistic models differ in their basic assumptions about learning and in their ability to handle multiple skills at the same time. And whilst in BKT the core assumption is the transition of unlearned to learned state, in logistic models, the assumptions are more diverse. For example, in AFM, one assumption is that all students learn the same concepts in the same manner, the difficulty of the concepts is the same for all students, there is no effect of past practice opportunities, and the model ignores the correctness of the previous responses (Cen et al., 2006). Unlike AFM, PFA (an AFM logistic regression variant) considers the correctness of individual responses. In PFA there are two fundamental parameters of prior practice, namely, success and failure, which contribute differently to future performance, resulting in better fitting (Pavlik et al., 2009). PFA uses a standard logistic regression model in which the learner performance is the dependent variable, predicted by the item difficulty, the effect of prior successes and the effect of prior failures for each skill. Extending PFA, IFA incorporates an additional parameter reflecting the gain from previous answers, further improving predictive performance (Chi et al., 2011). Parameter fitting for the logistic models can be done easily using standard logistic regression. The Q-matrix (i.e., a binary matrix that maps specific skills/rules to particular questions or types of questions) (Barnes, 2005) is specified either manually, or by using techniques like matrix factorization.

Although logistic methods seem to overcome some of the drawbacks of BKT (e.g., the handling of multiple skills), however, they are less appropriate for understanding and sense making processes for fine-grained concepts. Another potential problem is that PFA treats all observations as equally important, regardless of when they happened. The approach also ignores the potential impact of other skills and only considers skills tagged by the human expert as being relevant to the task to be solved.

More recently, the Elo rating system (Elo, 1978) was suggested for the interpretation of a learner's answer to an item as a match between the learner and the item (Pelánek, 2016). The Elo rating system is a formulation that is closely related to the Rasch model used in IRT, yet it differs from IRT's one parameter model in the use of heuristic equations instead of the maximum likelihood, as well as in the core learning assumption that learners' skills change over time. Pelánek (2016) used the Elo rating system to asymmetrically update the models in the case of correct/incorrect answers, by using time from previous attempt. Although the Elo rating system can model changing skills, it does not make fixed assumptions about learning and therefore, it lacks in granularity and it should not be expected to achieve optimal performance for a particular situation.

6.2.3. Motivation of the research: Why we need currently-aware learner models?

As explained in the introduction, the increased adoption of online learning environments has resulted in the need for accurate real-time adaptive services/systems, which in turn require robust, real-time learner models. From the literature review it became apparent that although the existing methods construct the learner models in real-time, however, in many cases, the human expertise is often required to label the different concepts, yet, in order to improve predictive accuracy, some approaches result in models that lack interpretability. Foremost, refitting the models' parameters is a time-intensive process which usually takes place offline. In a recent review of learner modeling, Pelánek (2017) highlighted the need for continuously updating the models right after the arrival of an observation, in real-time.

In order to holistically address these issues, the key role of the temporal dimension of learning should be considered. The reason is that when learning occurs in real-time, changes in learners' interactions data may vary in unforeseen ways (because learning behavior is a sensitive-to-change quantity that reforms over time), and there is a need to automatically align what is detected in the data with the actual learner state at that time-point, in a meaningful way. Refitting the models' parameters in run-time is necessary and requires enhancing the models' with a notion of "*temporal dynamics*" (Kleinberg, 2016; Whiting, 2015). Doing so allows to (a) predict the points in time when critical events are expected to cause/trigger significant changes in learners' behaviour; and (b) identify which was the learners' state exactly at that point in time and how this state changed, signalling the need for adaptation (what to adapt and when to adapt it). Moreover, temporal dynamics make it feasible to identify the intervals that the learners' states stabilize, indicating periods of more permanent skill/knowledge acquisition. The challenge towards "*currently-aware*" learner models is to refit the models parameters on-the-fly in order to capture the complexity of the learning process in run-time and to keep up-to-date the progress of the learners' multiple skills/knowledge acquisition in a single, adaptive, meaningful, fine-grained representation, and predict the future learning states from the current ones.

6.2.4. Research question and suggested approach

The review of the literature revealed that there is sparse practical evidence on dynamically refitting the parameters of the learner models in run-time, obtaining "*current awareness*". Moreover, handling issues of reconsidering the description of the predicted class the learner is categorized to from the changes in the values of the learners' attributes (i.e., *class drifting*), while maintaining the models' granularity, remains an open topic. Furthermore, due to the high requirements of real-time computing, none of the previous attempts for learner modeling can be characterized as "*memory-less*", restricting data tracking and long-term data storing. To this end, additional research is required and different methods should be considered for shaping and updating dynamic "*currently-aware*" learner models in run-time. The present study considers the identified issues, builds upon the well-established logistic and BKT models, benefits from advances on adaptive stream mining, and introduces and evaluates a method for

revising the temporal information in the learner models, with respect to the learners' changing level of multiple skill/knowledge mastery. Thus, the emerging research question is:

RQ: *How feasible is it to maintain “currently-aware” learner models in real-time, by refitting their parameters in run-time, and how accurately can these models approximate the learners' next cognitive states from the current ones?*

6.3. The learner modeling method: A “why”-“what”-“when”-“how” approach

The success of a learning/assessment environment significantly depends on how efficiently it identifies and models its users. The primary purpose of learner modeling is to capture and represent the learners' progress in skill/knowledge acquisition during the learning/assessment process (*why*). Learner knowledge is a latent construct and the goal is to quantify it using the available observational data (*what*). In each case, the first step is to determine an appropriate set of parameters that sufficiently capture and explain the variation in learners' behavior and attainments, as well as efficiently discriminate and categorize the learners. Next, these parameters need to be refitted periodically (*when*) to reflect the changes in knowledge/skill mastery and to better correspond to the current learner state.

6.3.1. The features of the learner models

For the representation of learners' progress, the set of parameters that reflect the acquisition of each skill/knowledge mastery (*what*) was selected according to the following criteria: (1) *The parameter is a good estimator of learners' mastery of knowledge:* it has been previously found to provide statistically significant explanation of the variance of learners' mastery level in learning/assessment activities; (2) *The parameter reflects the impact of non-cognitive factors:* the effect of non-cognitive factors on knowledge/skill acquisition has been widely approved in literature, and thus, the learner models should “carry” this type of information; (3) *The parameter will enhance the models with a notion of “temporal dynamics”:* temporal dynamics are patterns and trends in information streams that allow for temporal predictions (Whiting, 2015). This capacity is beneficial towards gaining “current awareness”, because it contributes to aligning what is detected in the stream (i.e., the temporal pattern), with what actually occurs. From this perspective, the parameters in the learner models might be time-varying factors (i.e., rapidly grown data that evolve over time) or temporal measures that will signal changes in learners' knowledge and facilitate the detection of concept drifting (Kleinberg, 2016; Whiting, 2015).

Previous research has acknowledged the temporal dimension of learning and has systematically proposed considering response time for learner modeling purposes (Klinkenberg, Straatemeier, & van der Maas, 2011; Lin, Shen, & Chi, 2016; Papamitsiou et al., 2016; Papamitsiou & Economides, 2017; Pelánek & Jarušek, 2015; Van Der Linden, 2009). The reason for doing so is mostly for improving the accuracy in predictions of performance. However, it is not only about making more accurate predictions, but these predictions should also make sense, be useful, and be easy-to-convert to adaptation decisions. Integrating time in the learner models is expected to

contribute to better explaining the state of the learners, at the moment that learning occurs, and predicting when changes in the learning states will occur, as well. Thus, response-time is one of the core parameters of the learner models.

According to the previous studies, the basic information used for learner modeling is the correctness of answers, and, usually, it is only it. From the literature review it was shown that item-oriented, skill-oriented and learner-oriented parameters are considered in the learner models to efficiently capture the learner states in a holistic manner. Such parameters include the items' difficulty, their discrimination ability, the learners' guessing and slipping probability, the time from previous attempt, the use of hints, a forgetting rate, the easiness of a skill, etc. Based on these findings, as well as on the above defined criteria, in the present approach, we consider the learners' prior skill mastery, and their accumulated response-times to answer correctly/wrongly for discriminating their ability, their time-spent according to the difficulty of the items, their Response Time Effort (Wise & Kong, 2005) that incorporates guessing.

Regarding the non-cognitive factors, the ones employed in the learner models are self-efficacy and goal-expectancy. That is because, on the one hand, perceived self-efficacy plays a major role in how the subject approaches tasks, and challenge, whereas, on the other hand, goal expectancy reflects the learners' achievement expectations (for additional evidence, see Bandura, 2006; Gutman & Schoon, 2013; Hsiao, Bakalov, Brusilovsky, & König-Ries, 2011; Papamitsiou & Economides 2017; Terzis & Economides, 2011). Other non-cognitive factors (e.g., emotions) could be investigated for improving the models.

Table 6-1 summarizes the parameters of the learner models during each stage of the proposed modeling process, as well as a short description of their roles, their type and value range.

Table 6-1. Features in learner models

Variable	Full Name	Description	Type	Value	
1st stage: Non-Cognitive					
Self reported	SE	Self-efficacy	Confidence in student's own ability to complete tasks and achieve results	Latent – questionnaire	1-5
	GE	Goal-expectancy	Student's perception of preparation and desired level of success	Latent – questionnaire	1-5
2nd Stage: Cognitive					
Time-varying	PSM	Prior Skill Mastery	Student's prior knowledge to successfully complete tasks that correspond to the specific skill	Composite – computed from actual data	{1,.., 5}
Temporal	TTAC	Time to answer correctly	The response-time a student aggregates on submitting correct answers	Simple – computed from actual data	≥0 (msec)
	TTAW	Time to answer wrongly	The response-time a student aggregates on submitting the wrong answers	Simple – computed from actual data	≥0 (msec)

	TTAE	Time to answer easy items	The response-time a student aggregates on lower difficulty items	Simple – computed from actual data	≥0 (msec)
	TTAM	Time to answer medium items	The response-time a student aggregates on medium difficulty items	Simple – computed from actual data	≥0 (msec)
	TTAH	Time to answer hard items	The response-time a student aggregates on higher difficulty items	Simple – computed from actual data	≥0 (msec)
Time-driven	CERT	Level of certainty	How certain the student wants to be (cautiousness)	Latent – questionnaire	1-5
	RTE	Response Time Effort	When a student exhibited solution behavior (engagement)	Latent – questionnaire	1-5

6.3.2. The stages of learner modeling

6.3.2.1. Overall approach

For constructing the models, a set of data analysis methods are applied on these features (*how*). Fig. 6-1 illustrates the overall learner modeling method, from the data gathering phase to configuring the complete models (*when*). In brief, during the first stage, learners' non-cognitive, self-reported goal-expectations and self-efficacy are used for clustering similar learners into groups (1). Next, a set of time-varying features (i.e., prior skill mastery, accumulated response-times, effort, level of certainty) is adaptively mined as a data-stream for configuring and updating the learner models in real-time (2).

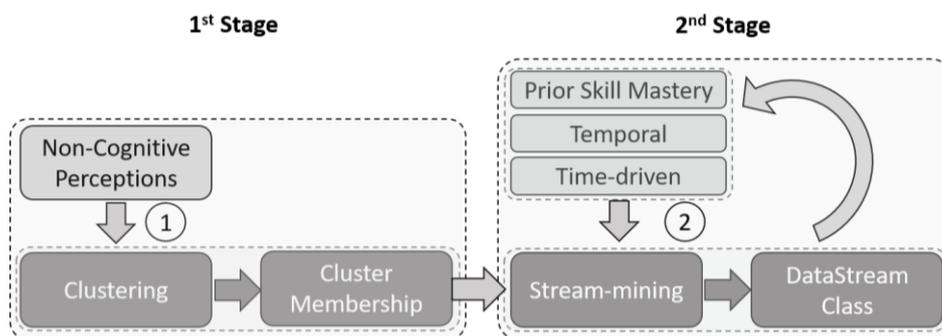


Figure 6-1. The overall learner modeling method.

Further elaborating on the use of clustering for categorizing the learners at the first stage of the method, previous research has shown that the heterogeneity of learners in a learning activity can be an important factor as different types of learners may be strong in different knowledge concepts (Beck & Gong, 2013; Streeter, 2015). And, although for fine-grained skills learners should be differentiated by the skill estimates, however, clustering is an appropriate method for discriminating learners using their non-cognitive characteristics. For example, clustering has been applied for gaining insight into learner behavior (Streeter, 2015). It was also argued that detected learner clusters can be applied to improve predictions by using different models or parameter values for each cluster (Gong, Beck, & Ruiz, 2012; Trivedi, Pardos, & Heffernan, 2011). Therefore, in the present approach, clustering is used for two reasons: (a) to

initially categorizing the learners according to their non-cognitive characteristics that are expected to influence their knowledge mastery; and (b) the output of the clustering process is expected to contribute to improving the prediction performance at the second stage of modeling.

6.3.2.2. Description of the stages

At the first stage, clustering is used for segregating homogeneous groups of learners according to their non-cognitive perceptions; learners who are assigned in the same cluster have similar traits of self-awareness – reflecting what they have yet to learn (Kaufman & Rousseeuw, 1990) – corresponding to similar expected level of skill/knowledge mastery. In the context of understanding data, clusters are potential classes, i.e., conceptually meaningful groups of objects that share common characteristics (Tan et al., 2005). For making sense of the results, each cluster is mapped to a class with a descriptive, human-readable label, according to the expected level of skill/knowledge mastery it is associated with. Five clusters representing the learners' perception of skill/knowledge mastery are shaped: $K=\{missed, scattered, insufficiently\ gained, gained, acquired\}$. The Cluster Membership (CM) – i.e., an id associated with each cluster label – is infused into the models at the next stage, to associate what is detected within the data-stream with a non-cognitive causation of the learners' mastery.

For the estimation of learners' current knowledge state and for their categorization in run-time during the second stage, an adaptive stream-classification approach is used due to the continuous supply of the small, yet fixed number of rapidly grown, time-varying features (Aggarwal, 2006). The student-generated training observations arrive in a stream and are inspected a single time only (i.e., process the entire dataset in one-pass); next, they must be discarded to make room for subsequent observations. The mechanism for configuring the learners' current knowledge state during this stage, processes the arriving data as follows: Let us consider a data-stream S as a sequence of labeled examples $s^t = (x^t, y^t)$ for $t = 1, 2, \dots, T$, where x^t is a vector of feature values (parameters of the model), and y^t is a discrete class label ($y^t \in K$). Every example is generated with a probability distribution $P^t(x, y)$. For each learner j , up-to- n vectors (i.e., (x_j, y_j)) are stored in the learner model (m_j) , $m_j=(s_{j1}, \dots, s_{jn})$, one per skill being assessed (for simplicity reasons, we follow the revised Bloom's taxonomy (Anderson (Ed.) et al., 2001), and use up-to-six vectors, yet other categorizations of skills could be used, as well). At each point time t , each vector contains the values s_j^t of each one of the eight learner's time-varying parameters (see Table 1), plus the CM value (the id), estimated in the previous phase, plus the class id the learner is classified to up-to-this moment for the particular skill. The classes are labeled to reflect the level of mastery of the specific skill/knowledge (i.e., $K=\{missed, scattered, insufficiently\ gained, gained, acquired\}$) according to a threshold of correctly answered items. The threshold is a parameter that can be determined manually by the instructor before the learning activity. In the beginning of the process (t_0), the values of all features are initialized to zero (except for CM that contains the value assigned in the previous step). Each time a learner interacts with an item that

assesses a skill (i.e., after a time interval Δ), the learner's time-varying features are updated into the respective vector, and the learner is classified to one of the five possible classes. The general classification task is to learn from the past training set of examples the relationship between the set of parameters and the class id. This relationship corresponds to a classifier that determines the class for the coming example $s^{t+\Delta}$; the classifier will provide its prediction *at any time* based on what it has learned from the examples $\{s^1, \dots, s^t\}$ seen so far.

However, for the same classes being trained, the distribution of the parameters' values (may) change with time, causing unpredictable class drifting (Bifet & Gavaldà, 2009). If for two distinct points in time, t and $t + \Delta$, there exists x such that $P^t(x, y) \neq P^{t+\Delta}(x, y)$, then class drift has occurred. When class drift occurs, either one or both of the following changes: (a) prior probabilities of classes $P(y)$, (b) class conditional probabilities $P(x/y)$. Drift detection is not a trivial task, yet fast drift detection is required to quickly replace the outdated model. The goal is the data-stream classifier to adapt to the changing class by forgetting outdated data, while learning new class descriptions. In the suggested approach, a decision tree is induced from the data-stream, and each example s^t is inspected in the stream one time only: it is kept in the nodes only if it is considered as relevant to guide the adaptation to class drift. When class drift is detected, an alternate decision tree is induced according to the new class description. For making this decision, statistics about the example are used, and each node can decide which of the last instances are relevant. For that purpose, instances of estimators of frequency statistics are placed in the nodes (the estimator can be an Exponential Weight Moving Average (Perry et al., 2010), for example). The statistics include the Hoeffding bound (Hoeffding, 1963), classification error, and other heuristics. Let us consider an instance A_{ikc} of an estimator, which has the value k in its i^{th} feature, and class c . The estimator is updated at the leaf and an alternate tree is also started at the root, induced from an arriving example, if $|A_{ikc} - P_{ikc}| \leq \sqrt{\frac{\ln(1/\delta')\Delta}{t(t+\Delta)}}$, where, P_{ikc} is the probability that the arriving example at the node has value k in its i^{th} feature, and class c . δ' should be the *Bonferroni correction* of δ (Dunn, 1961) to account for many tests performed, and all of them need to be simultaneously correct with probability $1 - \delta$, whereas δ is the desired probability that the correct feature is chosen at every point in the tree.

Overall, at a randomly selected time-slot t_1 , a learner's vector contains the values for each one of the learner's time-varying parameters, the learner's CM, as well as the class the learner is categorized to up-to this moment. At a later time-slot t_2 , the vector contains the revised values for all the previous features, as well as the new class the learner has been assigned to (Fig. 6-2).

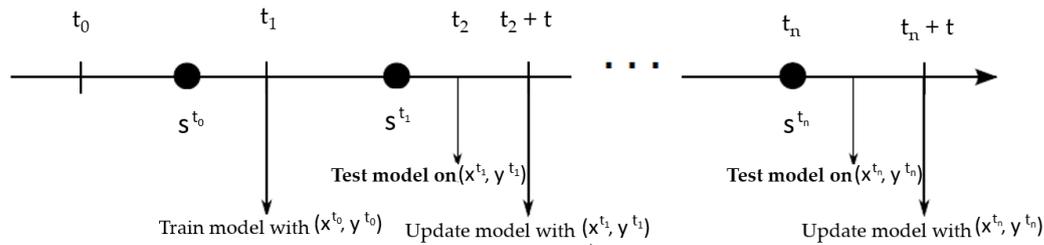


Figure 6-2. Outline of real-time model update (testing and training).

The difference between the two classes reflects the progress made by the learner, by considering the learner's CM, as well. The values are finalized when the learner completes all items that assess the specific skill, and the class the learner is finally categorized to regarding this skill, is also stored in the vector. The result of this process (DSC – Data-StreamClass) is a vector containing up-to-six values, one for each of the up-to-six skills being assessed, and corresponding to the class the learner is finally categorized to for the respective skills. The second stage of the learner modeling process might be repeated up-to-n times, depending on how many skills are being assessed, allowing for representing multiple skills in a joint profile.

6.4. Experimental evaluation

In order to ensure that the suggested modeling method is sound and that its contribution is significant, an exploratory study with non-experimental setup was designed and conducted. The first layer of the Layered Evaluation Framework -LEF (Brusilovsky, Karagiannidis, & Sampson, 2004) was applied for assessing the validity of the conclusions drawn by the learner models and answering whether the learners' traits are being successfully detected and stored in the learner models. The stream mining was implemented in MOA (Release 2016.04) (Bifet, Holmes, Kirkby, & Pfahringer, 2010).

6.4.1. Participants and experimental setup

Data were collected with the LAERS environment (Papamitsiou & Economides, 2013) at a European University during a self-assessment process with 503 undergraduate students (231 males [45.9%] and 272 females [54.1%], aged 19-28 years old (M=20.21, SD=1.483, N=503)). Five randomly generated groups of 100-102 students attended the procedure for the Microeconomics II course (related to monopolistic competition, competitive strategy, equilibrium theory) at the University computer lab.

For the self-assessment, 45 multiple choice items were used in total, distributed in 5 equivalent tests of 25 items each (some of the items were shared in more than two tests). Each item had 2-4 possible answers, but only one was the correct. The items were delivered to the participants in predetermined order. The set of available interactions to students was as follows: the students could view the items without answering them, they were allowed to temporarily save their answers, to review them, to alter their initial answer choices and save new answers, as well as skip an item and answer it (or not) later. They finalized and submitted their answers only

once, whenever they were ready to do so. Each item corresponded to one of the first five levels of the factual, conceptual and procedural domains of the knowledge dimension according to the revised Bloom’s taxonomy (see Table 6-3). We chose to deliver to students items of these five levels only, for two reasons: (a) the inherent complexity of the sixth level, and (b) we wanted to explore the accuracy of the learner models with the maximum possible number of skills being assessed. During the design of the tests, two instructors agreed on the items’ difficulty (easy, medium, hard) – instead of using, for example, the 3PL IRT’s β parameter (Birnbaum, 1968), that would unnecessarily increase the complexity of the models at this stage of the models’ formation.

Before taking the self-assessment, each participant had to answer to a pre-test questionnaire that measures each learner’s goal expectancy and self-efficacy (Appendix C). Table 6-2 synthesizes the exploratory study, i.e., the description of the sample and the self-assessment process.

Table 6-2. Synopsis of the exploratory study

Sample		Self-assessment	
Overall Size	Groups Sizes	Pre-test	Quiz (25 questions/group)
503	G1: 100	Goal Expectancy, Self-Efficacy	(5 items – L1: remembering,
	G2: 100		5 items – L2: understanding,
	G3: 100		5 items – L3: applying,
	G4: 101		5 items – L4: analyzing,
	G5:102		5 items – L5: evaluating)

The participation to the procedure was optional; it was offered to facilitate the learners’ self-preparation before the final exams. All participants signed an informed consent form prior to their participation. The informed consent explained to them the procedure and was giving the right to researchers to use the data collected for research purposes. Learners were aware that their interactions were tracked and anonymized prior to being analyzed, and that the collected data would be stored for 3 years.

6.4.2. Data analysis methods

At the first stage of the method, for clustering similar learners according to their non-cognitive, self-reported goal-expectancy and self-efficacy, the popular *k-means* algorithm was used. The purpose of *k-means* is to group the data by segregating a set of observations into *k* clusters according to a similarity measure. (Tan et al., 2005). The method initially selects *k* points as the centers of each cluster and iteratively reassigns points to their nearest cluster centroid, followed by recalculating the cluster centroids. The value of *k* is usually defined using the Silhouette measure of cohesion and separation for the interpretation and validation of consistency within clusters of data (de Amorim & Hennig, 2015). The minimum required sample size, given the desired statistical power level (≥ 0.8) and the probability level ($p < 0.05$), is 85. In this study, the sample size is 503, and thus, it is large enough.

Next, the learner-generated time-varying data were adaptively mined as a data-stream for estimating the learners' current state and predicting the class of skill possession each learner should belong to. For this purpose, the *HoeffdingAdaptiveTree (HAT)* classifier was employed (Bifet & Gavaldà, 2009). HAT incrementally induces a decision tree from a data-stream, inspecting each example in the stream one time only (for description of HAT algorithm, see Appendix). In our study, due to the limited size of the dataset, compared to the big data generated in other contexts, two parameters of HAT were lowered from their default values: (a) the grace period (i.e., how many examples should be seen in a leaf since the last evaluation, before revisiting the decision); we experimented with different, commonly used values (50, 100, 200), and (b) the sample frequency was set to 200 instances between samples. The other parameters of HAT were used with their default values.

HAT was selected for effectively resolving the issues of unpredictable class drifting and reduced classification accuracy, caused by changes in distribution of the features' values over time. It has been acknowledged to be as accurate as other well-known methods for tree induction on data-streams with drift (e.g., Concept-adapting Very Fast Decision Trees (Hulten, Spencer, & Domingos, 2001)) and, in some cases, it has substantially lower error (Bifet & Gavaldà, 2009). HAT does not need a fixed size of sliding window (for managing statistics at the nodes), the size of which is a parameter that is very difficult to guess, and it also performs better in terms of time compared to bagging methods (e.g., ASHT Bagging (Bifet, Holmes, Pfahringer, Kirkby, & Gavaldà, 2009)).

6.4.3. Evaluation metrics

For the evaluation of the learner modeling method proposed in this study, the first layer of the Layered Evaluation Framework (LEF; Brusilovsky et al., 2004) was employed for two reasons: (a) "*this (i.e., the first) evaluation layer does not assume that the adaptation decision making component has already been developed*", allowing for model evaluation at early stages of adaptive systems' development, prior to building the full adaptation mechanism, and (b) according to the LEF, a critical question for assessing the validity of the learner models is: "*Is the user's actual knowledge of the subject being successfully captured?*", allowing for every sound methodology to be applied for addressing this question.

Measuring data-stream classification performance involves space (the available memory is usually fixed), learning time and accuracy. The most popular evaluation method for stream classifiers is the *Predictive Sequential (prequential)* error estimation (Gama, Sebastião, & Rodrigues, 2009). It allows to monitor the evolution of the streaming classifiers and evaluates the performance of the models by testing each example and then using it for training in sequence. It is achieved with a forgetting mechanism (like a time window or a fading factor) and the accuracy is incrementally updated. Furthermore, the *Kappa statistic* and the *Temporal Kappa Statistic* (K_{temp}) measure performance of streaming classifiers (Bifet, Read, Žliobaite, Pfahringer, & Holmes,

2013). The *Kappa statistic* measures the agreement of the prediction with the true class; a value of *Kappa* equals to 1.0 signifies complete agreement. K_{temp} values ranges between (1, $-\infty$) and equals 1.0 if the classifier is accurate.

6.5. Results

A brief description of the collected dataset is outlined in Table 6-3.

Table 6-3. Synopsis of the collected dataset

Thresholds of correct answers for skill acquisition:									
<i>0=missed, 1=scattered, 2=insufficiently gained, 4=gained, 5=acquired</i>									
Interactions: view (v), re-view (r), save answer (s), change answer (c), skip (l)									
Dataset Size	Observations (Interactions) per skill	Average Observations (Interactions) per learner per skill	Average Interactions per item per skill	Average Frequency of Interactions per skill					
				v	r	s	c	l	
2.3MB	12834	24.6	4.84	1.1	1.8	2.0	1.1	0.7	
	12922	25.2	5.04	1.1	2.1	2.1	1.1	0.5	
	12946	26.4	5.32	2.1	2.6	2.6	1.2	1.1	
	12758	25.8	4.34	2.7	2.9	2.6	1.7	1.6	
	12882	25.2	4.76	2.8	2.9	2.7	1.8	2.0	

6.5.1. Exploratory data analysis

The GE and SE factors are latent variables, measured via questionnaires. For evaluating the validity of these latent constructs, the items' factor loadings on the corresponded constructs have to be higher than 0.7, Cronbach's α and composite reliability (CR) have to be higher than 0.7, and Average Variance Extracted (AVE) has to be higher than 0.5. Table 4 confirms validity for the latent constructs.

Table 6-4. Results for Validity of the Latent Constructs

Construct Items	Factor Loadings (>0.7) ^a	Cronbach's α (>0.7) ^a	Composite Reliability (>0.7) ^a	Average Variance Extracted (>0.5) ^a
GE		0.83	0.89	0.75
GE1	0.833			
GE2	0.876			
GE3	0.883			
SE		0.72	0.78	0.54
SE1	0.782			
SE2	0.722			
SE3	0.712			

^a Indicates an acceptable level of reliability and validity

6.5.2. Clustering results

In this study we used $k=5$ clusters. The value of k was determined according to two criteria: (a) the average Silhouette=0.6, and (b) the value of five clusters is meaningful and easy to interpret pedagogically and to reflect granularity in knowledge mastery. Other values of k were

explored as well ($k=3$, $k=4$, $k=6$), but in these cases some of the clusters were under-represented. With the selected value, k -means clustering divided the students into five groups, with 122 (24.3%), 68 (13.4%), 94 (18.7%), 122 (24.3%) and 97 (19.3%) records each, respectively, illustrated in Figure 6-3.

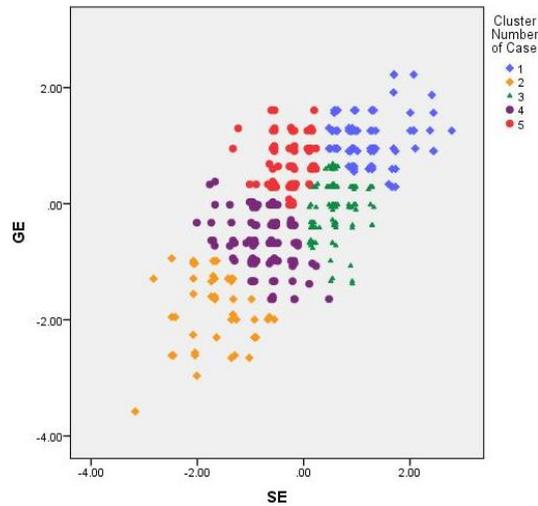


Figure 6-3. Assignment of students to clusters according to goal-expectancy and self-efficacy.

As apparent from this figure, the clusters are clearly distinguished from each other. This qualitative observation is further confirmed from the clustering analysis: as seen from k -means ANOVA output (Table 6-5), GE and SE contribute the most to the clusters, since the F -values are large (p -values < 0.05) for both variables, and their means are significantly different between clusters (Fig. 6-4).

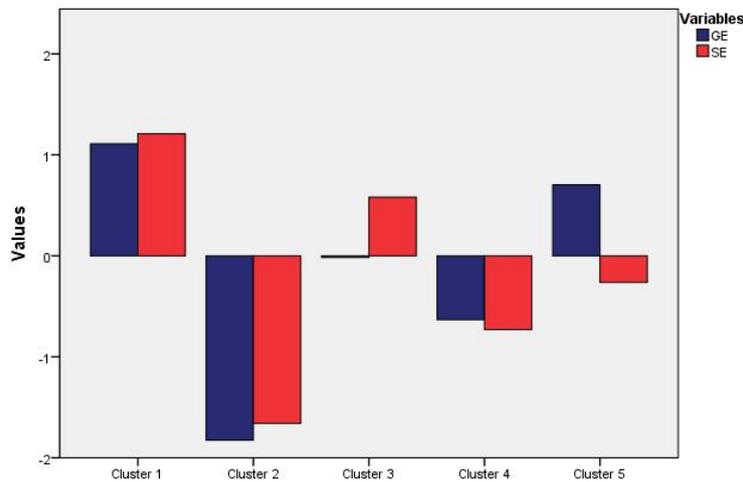


Figure 6-4. Final cluster centers.

Table 6-5. k -means ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
GE	87.29	4	.23	498	382.26	.000
SE	91.12	4	.19	498	470.42	.000

However, the clusters have been chosen to maximize the differences among cases in different clusters, and thus, the *F-test* should be used only for descriptive purposes. The observed significance levels cannot be interpreted as tests of the hypothesis that the cluster means are equal. Post-hoc Bonferroni test revealed that, statistically, these five clusters are significantly different from one another, i.e., the means of their within points are not equal ($p < 0.05$ in all pairwise comparisons).

6.5.3. Data Stream Classification results

For monitoring how learners' multiple cognitive skill/knowledge self-assessment progresses in real-time, and for adaptively classifying them on-the-fly accordingly, the HAT classifier was applied on a stream of the eight time-varying features (Table 6-1), and the CM, as explained in section 3.2. Table 6-6 summarizes the accuracy and time results for the HAT algorithm. They are estimated for the three cases explored for the grace period parameter (i.e., 50, 100, 200). The table includes both the overall converging results, as well as the results during the intermediate states of the self-assessment – evaluating the acquisition of each skill.

Table 6-6. Performance Metrics for the HAT Classifier

Classifier: HAT	Performance Metrics											
	Accuracy (prequential error)			Kappa Statistic			K _{temp}			Time (in sec)		
	50	100	200	50	100	200	50	100	200	50	100	200
Grace period	50	100	200	50	100	200	50	100	200	50	100	200
Skill1: remembering	66.5	75.2	76.2	53.15	57.22	61.15	43.59	51.53	57.38	1.36	1.08	0.97
Skill2: understanding	68.5	77.2	79.5	55.73	58.19	64.73	-203.1	-136.7	-125.2	1.19	1.28	1.47
Skill3: applying	72.4	81.6	84.6	59.91	61.46	65.91	57.50	57.78	63.06	1.30	1.30	1.25
Skill4: analyzing	70.9	80.0	83.7	58.72	60.11	63.72	11.34	48.21	57.76	1.08	1.03	1.41
Skill5: evaluating	76.4	84.5	88.0	62.50	64.65	69.40	-23.08	-11.9	4.76	0.75	0.91	0.92
Overall	70.9	79.7	82.4	58.00	60.33	64.98				5.68	5.60	6.02

As seen from this table, a grace period of 200 was found to generate the highest level of accuracy for the (almost) 13000 examples per skill. HAT achieves a satisfactory classification accuracy: for each skill, it is higher than 76%, and overall, it is approximately 82.4% (Kappa=0.65). Moreover, Fig. 6-5 plots the accuracy over time (as the number of arriving examples gradually increases) for each skill. The blue line represents the accuracy vs. the number of examples trained/tested, and the red line is the trendline.

The classifier appears to provide stable and accurate predictions for all skills, throughout the self-assessment process. After approximately 2500 builds the prediction accuracy begins to stabilize.

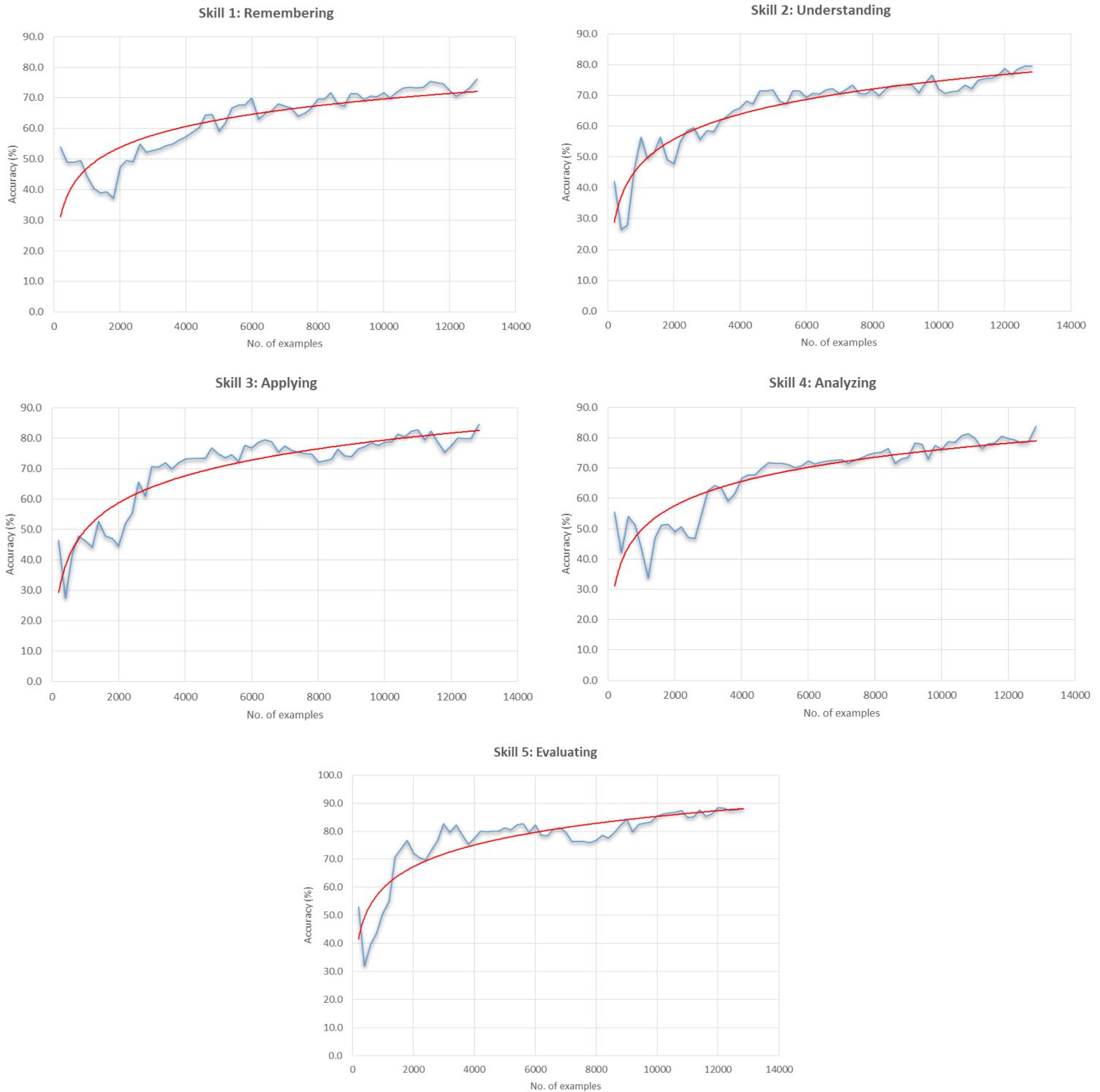


Figure 6-5. Accuracy of the HAT algorithm for student models' configuration in real-time, for each skill being assessed.

6.5.4. Overall evaluation results

According to the obtained results, two significant facets become apparent: (a) the accuracy of learners' skill/knowledge acquisition progress being estimated by refitting the models parameters in run-time is higher than 82% (the final achievement levels are not yet available), (b) the overall classification accuracy begins to stabilize after a small number of builds

(approximately 2500 examples). Moreover, the accuracy in the estimation of the learners' progress for each skill, guided by the eight time-varying features, gradually improves (first skill: 76.2% - last skill: 88.0%). The values of the respective parameters are stored in the learner models (tree nodes) and they are refitted throughout the process. As apparent, the overall results can be considered statistically significant, and the learners' current knowledge/skill mastery is sufficiently and successfully captured by the proposed models.

6.6. Discussion

The main objective of the present study – expressed as the RQ – was to investigate whether currently-aware learner models make it feasible to efficiently shed light onto learners' current learning states and predict future ones, by refitting the parameters of the models in run-time, according to the continuously changing learner interaction data. The goal was to explore whether enhancing the models with temporal information improves the accuracy of temporal predictions, and more importantly, whether the constructed models are easy-to-understand and interpret. As seen in Table 6-7, the overall correct observations in the models reach 82%. The data analysis revealed some interesting findings, discussed in this section.

6.6.1. Initializing the student models

The first critical issue that should be resolved, concerned the initialization of the models' states, given that no prior information about the learners' knowledge states was available. Other modeling methods employ, for example, probabilities of prior knowledge (Pardos & Heffernan, 2010) or learners' background knowledge in other domains (Chrysafiadi & Virvou, 2013a). In the present approach, the idea was to directly measure learners' motivational perceptions, previously found to be good estimators of learners' mastery of knowledge (Papamitsiou et al., 2014; Zimmerman, Bandura, & Martinez-Pons, 1992), and categorize them accordingly, in classes of potential skill/knowledge mastery.

At the first stage of learner modeling, the *k-means* divided the learners into five groups, which were significantly different from one another (Bonferroni test, $p < 0.05$ for all pairwise comparisons). The selected $k=5$ is a meaningful solution since the clustering convergence to zero was achieved in less than ten iterations. The cluster sizes were balanced; none of the clusters was underrepresented, indicating efficient (small number of clusters) and effective (comparable numbers of cases) clustering.

In order to make sense of the results, a preliminary look at the clusters outlines the non-cognitive traits of the learners assigned to each of them (Figure 6-3): learners in cluster 1 (C1) score high in both non-cognitive factors, whereas learners in cluster 2 (C2) score low in these features; learners in cluster 4 (C4) exhibit moderate motivation; learners in the other two clusters either score high in goal-expectancy but low in self-efficacy (cluster 5 – C5), or they score high in self-efficacy and low in goal-expectancy (cluster 3– C3).

Motivational concepts such as goal-setting and self-efficacy have been found to strongly affect the learners' progress and educational achievements (Zimmerman et al., 1992), and the level of skill/ knowledge mastery can vary accordingly. High motivation is correlated to better learning/assessment results, and as goal-expectations and self-efficacy decrease, it indicates learners' lack of self-confidence and reduced self-esteem, leading to lower educational attainments. Interpreting the clustering results accordingly, leads to the following observations: learners in C1 and C2 seem to be more self-aware regarding their skill mastery and potential abilities and expertise – unless they are very optimistic or very pessimistic, which, at a later stage, causes a decrease in classification accuracy. As such, C1 is associated with the class “*acquired*” skill/knowledge (class A) and C2 with the class “*missed*” skill/ knowledge (class E). Learners from C3 and C5 either count on their preparation, although they doubt their self-efficacy, or they believe in their abilities to complete tasks, although they are not prepared enough to successfully deal with them. Thus, C3 is associated with the class “*gained*” skill/ knowledge (class B), while C5 with the “*scattered*” skill/ knowledge class (class D). Finally, learners in C4, behave more or less neutrally; they think that they are prepared, but not prepared enough, and that they are efficient, but not efficient enough. Thus, they either possess a skill insufficiently, or they straggle to gain it, leading to associating C with the “*insufficiently gained*” skill/knowledge class (class C). The information about the cluster membership is stored in the CM variable, and is fed to the models for aligning perceptions with actual interaction data and for updating and finalizing the learner's next states.

6.6.2. Integrating temporal information in the models

The second major issue to be addressed was what information should be included in the learner models, to contribute sufficiently to determining the learners' current knowledge states, and to allow for temporal predictions, with respect to the learners' continuously changing level of skill/knowledge mastery. The next step concerned *how* to constantly integrate this information in the models and update them in run-time throughout the self-assessment process, maintaining “*current awareness*”. In other words, it was essential to define (a) the learners' parameters that could capture the complexity of the information required to model the progress made over time, and, (b) the method to accurately update these parameters in run-time and re-classify learners accordingly.

Well-established methods like logistic models and BKT learner models in real-time, however, in many cases, the human expertise is often required to label the different concepts, yet, in order to improve predictive accuracy, some approaches result in models that lack interpretability. Foremost, refitting the models' parameters is a time-intensive process which usually takes place offline. The models are mostly based on the correctness of the answer/solution, and consider item-oriented, skill-oriented and learner-oriented parameter such

as the items' difficulty, their discrimination ability, the learners' guessing and slipping probability, the time from previous attempt, a forgetting rate, etc.

The core idea behind the suggested approach was to "borrow" a set of essential parameters from BKT and logistics models as concepts, and to map these concepts to time-varying learner observational data. More precisely, instead of employing probabilities for modeling learners' knowledge state, all measures used to describe the learners' progress are associated to or expressed in terms of continuously tracked and aggregated response-times that arrive as an evolving data-stream – except for prior knowledge that is instantly calculated according to the correctness of the answer and a pre-defined threshold. The stream is then subjected to adaptive classification. This innovation is introduced in favor of incorporating time in the models, refitting the parameters in run-time, and enhancing the models with a notion of "*temporal dynamics*"; signaling changes in learners' state at any time-slot of the procedure.

The empirical results from the data-stream classification at the second stage indicate that the aggregated response-times on answering the self-assessment items, in combination with the time-spent according to the items' level of difficulty, the students' time-varying level of certainty, their effort and their evolving mastery of skill (prior knowledge) could efficiently be used for identifying the current state of the learners in real-time. Treating the time-varying variables in an adaptive data-streaming fashion provides an average classification accuracy that is higher than 82%; the learning time of HAT is only a few seconds, with a grace period of 200 examples seen in a leaf since the last evaluation, before revisiting the decision. As more examples are trained, the classification accuracy begins to stabilize.

6.6.3. Interpreting the learner models

Looking at the variables in the nodes at randomly selected time-slots, a list of observations is made:

- *TTAC/TTAW (Time to answer correctly/Time to answer wrongly)*: learners who are classified to the higher mastery class (class A) aggregate higher TTAC, while high TTAW better suits the class of lower mastery (class E); learners who are categorized to medium mastery level (class B) might have high TTAC, yet they aggregate non negligible amounts to TTAW, and conversely, although learners in class D are marked with higher TTAW, they gather appreciable TTAC. This finding is in agreement with previous results that response-times to answer correctly/wrongly have significant impact on the assessment result (e.g., Papamitsiou et al., 2014; Shih et al., 2008).
- *TTAE/TTAM/TTAH (Time to answer easy/Time to answer medium/Time to answer hard)*: regardless of the mastery class the learners belong to, all learners dedicate considerable TTAE, because they try to figure out the solution to the easy items. Similarly, learners in classes A and B devote high TTAM, due to internal motivation to prove themselves that they have gained remarkable mastery on medium items. On the other hand, learners in classes D and E

aggregate low TTAM. This is because these types of learners might find these items difficult enough for their skills and knowledge to get involved with. Learners in class A also gather higher TTAH. The reason is that learners who belong to this class try to increase their competences and thus, might find the hard items challenging enough for their abilities. On the contrary, learners in classes D and E aggregate low TTAH. These demanding items are more possible to be perceived as beyond their personal goals, and thus, their engagement is limited. This finding complies with previous results that the level of item difficulty influences the learners' dispositions regarding the self-assessment items, and respectively, the time they aggregate to answer the corresponding items (e.g., Conejo, Guzmán, Perez-de-la-Cruz, & Barros, 2014; Dodonova & Dodonov, 2013).

- *CERT/RTE (Level of certainty/Response Time Effort)*: learners in class A exhibit higher levels of certainty (CERT), but this feature is somewhat lower for learners in class B, who, nevertheless, demonstrate a trend to increase their certainty. On the contrary, learners in class D are less confident students, expressing lower CERT during self-assessment, and learners in class E exhibit the lower values in CERT. Similarly, the more committed the learners remain (high effort), they are classified in class A or in class B, while learners who are classified in classes D and E exhibit lower effort. This could be considered as an additional evidence of Humphreys & Revelle (1984, p.158) definition of effort as “the motivational state commonly understood to mean trying hard [...]. Effort is increased when the subject tries harder, when there are incentives to perform well, or when the task is important or difficult”.

As seen in Fig. 6-5, it is observed that at the start of the training process for each skill, insufficient examples exist, resulting in model under-fitting. The overall trend shows that, as more examples are trained, the classification accuracy steadily improves. Especially for skills 4 and 5 (analysing and evaluating, respectively), the estimation is even better, and the trendline gets its higher values and it reaches 88%. That is probably because, for assessing these skills, more difficult items are employed. These demanding items are more possible to be perceived as beyond the personal goals of less prepared learners, and thus, their engagement is limited. On the contrary, well prepared learners try to increase their competences and thus, might find the hard items challenging enough for their abilities. Accumulated time-spent according to the difficulty level of the items is a significant factor of the models that contributes to discriminating the learners' skill/knowledge mastery level. Overall, the trendline of accuracy is steadily above 76% for all skills being self-assessed, and HAT appears to provide accurate predictions throughout the self-assessment.

6.7. Conclusions

In the era of Big Data, the tremendous increase in online learning has signaled the momentum of real-time processing of the learner data and has enforced the demand for quality services that adapt to the individuals' learning states on-the-fly. Online learning has a strong

inherent temporal dimension: it is time-framed and it carries information about the current state of the learners, in real-time. The learners' states themselves uphold a temporal dimension, as well: they are sensitive-to-change and evolve over time. Towards efficient real-time adaptive learning services, building and updating accurate and easy-to-use learner models in real-time is required. For this purpose, the adaptive learning environments should be "*currently-aware*" of their learner's progress, as it evolves over time, whereas the parameters of the learner models should be frequently updated in run-time, and non-cognitive features should be considered as well, for gaining a more holistic perspective. *What* to model, *how* and *when* are core issues to address before making critical decisions on the learner models.

Previous prevailing methods either employ parameters that either are updated offline, or they require the human expertise to clarify task-related issues. Although they provide fine-grained models, yet, they process the learner-parameters in a batch fashion, and they lack temporal dynamics. Moreover, they missed out significant non-cognitive factors that have been acknowledged as fundamental moderators/facilitators of the learning processes (Zimmerman et al., 1992), or they endeavor to increase the learners' self-awareness, but lack in predictive capabilities regarding the future states of the learners.

This contribution envisages to open the discussion towards building learner models that maintain "current-awareness". It is acknowledged that the data-stream perspective takes a static view of a problem and adds a strong temporal dimension to it, making it feasible to have accurate temporal predictions in real-time and refitting of the parameters of the predictive models in run-time, as well. Building upon recent advances on adaptive data-stream classification, this work attempts to shape and update fine-grained learner models on-the-fly. The time-varying predictors employed in the models are inspired from the well-established BKT and logistic models. The proposed method identifies the learners' response-times patterns reflecting the gradually changing level of skill/knowledge mastery, and constantly incorporates this information in the learner models, and refreshes the learners' state. The constructed models incorporate time (either in the form of time-varying or in the form of temporal parameters) and are enhanced with a notion of "*temporal dynamics*": signaling changes in learners' current states and allowing for temporal predictions during the learning process. Moreover, the models are initialized with the learners' non-cognitive features, in an attempt to (a) associate non-cognitive traits to the learning attainments, and (b) to improve predictive accuracy of the learner models.

6.7.1. Findings and contributions

In order to assess the validity of the conclusions drawn by the learner models built with the proposed method, and evaluate its efficiency, an exploratory study with non-experimental setup was carried out with a web-based self-assessment environment. Five hundred and three (503) undergraduate students from a European University participated in a self-assessment procedure. The results from this study are encouraging:

1. the selected number of clusters ($k=5$) was a meaningful solution; the non-cognitive traits used in this phase provided sufficient information about the corresponding classes,
2. the aggregated response-times on answering the self-assessment items, in combination with the time-spent according to the items' level of difficulty, the students' time-varying level of certainty, their effort and their evolving mastery of skill/knowledge could efficiently be used for classifying students in real-time (average HAT accuracy > 82%),
3. multiple cognitive skills can be jointly represented in a single learner profile, and the progress gained for each of them can be accurately estimated,
4. changes in the distributions of the values of the time-varying parameters cause class drifting and indicate changes in learners' progress; as such, (a) class drift detection asserts an urgency for adaptation or other intervention to further support the learner in need, and (b) classes are re-defined in run-time and learners are re-classified accordingly,
5. the statistically significant 82% in classification accuracy provides strong indications regarding the overall efficiency of the method.

These findings contribute in five different ways:

- a. they open the discussion towards currently-aware learner models: the emerging need to enhance the models with temporal information and dynamics, as well as to process the collected interaction data in run-time, in a way that is both satisfactorily accurate and easy to make sense,
- b. the models successfully capture the temporal dimension of online learning, signal changes in learners' behavior and adapt to these changes, and predict the future states of the learners, as well,
- c. the models are easy to interpret and make sense of which are the exact learner traits at each point in time (for each one of the multiple skills measured), and how they have evolved to this point,
- d. the time-varying factors allow for gaining fruitful insight into learners' knowledge state ("current-awareness"), and should be incorporated for updating the learner models accordingly, and
- e. this combination of features and analysis methods make it feasible to detect differences between learners, discriminate and categorize them accordingly, and to detect differences in the behavior and knowledge of the same learner and forecast the learner's next.

The arising question is: *how we can exploit and utilize these findings towards developing efficient real-time adaptive applications or services?* Next we discuss about possible implications of the findings.

6.7.2. Reflections on learning modeling aspects and practical implications

Based on the abovementioned findings, we suggest that one can identify a set of time-driven, temporal or behavioral factors that could constitute the core foundations of a real-time

adaptive learning system. The grounded previous experience with models that sufficiently capture learners' progress (e.g., BKT, PFA) has resulted to a set of parameters that are appropriate for doing so. These parameters are mostly probabilities with high discrimination ability and reflect the level of learners' gaining, possession and attainment of knowledge. Inspired from these factors, in the suggested approach, a set of actual measurements that correspond to a similar group of variables has delivered a statistically significant explanation of learners' progress, as well. In addition, the data processing method (i.e., the adaptive data-stream classification) offers a potential for refitting these parameters in run-time and revising the models according to the detected changes in these parameters ("*current-awareness*"). Time-spent on correctly/wrongly responding to the self-assessment items, accumulated time-spent according to the level of the items' difficulty, required response time effort, and time-varying level of certainty are variables that could capture the complexity of real-time learning and model the learners' progress accordingly. The added-value of these variables is that they are simple to implement and easy to compute, their interpretation is straightforward, and they are enhanced with a notion of "*temporal dynamics*". Taking advantage of the fact that these predictors are direct measurements, the real-time adaptive learning system/service could efficiently identify the individuals' learning pace, misconceptions and needs, and respond with appropriate feedback (*what*) at the most appropriate timing (*when*). Examples of such feedback could be the recommendation of additional hints, learning components, alerts, etc. Furthermore, providing visualizations of related analytics to the learners could allow for further self-reflection, increase their self-awareness and guide self-regulation during the learning process, in general. Moreover, regarding the non-cognitive predictors included in the models for measuring self-awareness and motivation (e.g., GE, SE), they are only indicative variables that could be embedded into the system/service in order to model the learners. Other factors like emotions, mood, satisfaction and suitability of content, for example, could be investigated, as well.

6.7.3. Limitations and Future work

The added-value of this work stems from the successful detection of response-times patterns, the fusion of multiple skills in a joint representation, the accurate estimation of the learners' current cognitive state, and the straightforward interpretation of the models. To make the contribution of this method stronger and establish its significance, additional experimentation – with bigger samples, different learning disciplines and longer learning procedures – is required. Moreover, it is necessary to empirically compare the obtained results to the respective ones from the acknowledged BKT method.

Another future work plan is delivering this information to learners as real-time feedback, in the form of visualizations of assessment analytics, and measuring its effect on learners' self-awareness and self-regulation. Moreover, as already stated, no adaptations were implemented so

far. It is essential to investigate the bidirectional effect of learner models on the adaptation mechanisms and vice versa.

Furthermore, the nature of the data collected (time-based parameters) and the general-purpose methodology followed for the analysis of these data, render this approach replicable and/or transferable to other contexts. The mechanisms for tracking response-times data consume low computational resources, are cost-effective and can easily be implemented in any learning or assessment system. For example, time-related parameters could be tracked to measure the duration of solving/implementing sub-activities in the context of project-based learning, to measure the duration of studying/exercising with learning modules during inquiry-based learning, etc., along with the number of repeating the intermediate, facilitating steps (e.g. watch educational videos, use educational resources, participate in discussions). Applying the suggested method in other contexts is a future plan.

Finally, other features (e.g. frequencies of requesting a hint or consulting the statistics of the item, time on viewing-studying this information, frequencies of requesting similar assessment items and time-spent on solving these items, etc.) could also participate in the learner models as complementary behavioral attributes that reflect the students autonomy or creativity within the real-time learning environment. Furthermore, including time-varying affective factors in the student models (e.g., instant emotions/emotional stability, etc.) could complement the models with more sophisticated, fine-grained attributes from the affective dimension of learning which have been found to be strongly correlated to the self-assessment result. Considering these features, and capturing the complexity of learning as the whole picture in real-time is within our future work plans.

Chapter 7 : From motivational profiles to help-seeking strategies

“Not all those who wander are lost”

John Ronald Reuel Tolkien

Which help-seeking strategies do the learners use according to their motivational profiles? A pattern-based approach using learning analytics

7.1. Introduction

Contemporary learning theories highlight the significant role of feedback on the learners' personal development (Economides, 2009b; Hattie & Timperley, 2007). Feedback is a key tool for guiding and sustaining learners' involvement in the learning process, self-regulation and goal attainment (P R Pintrich, 2004; Zimmerman, 2002). However, feedback on its own might not impact learning as expected, unless the learners are willing to use it. A specific version of on-demand feedback, which directly involves the learners' enrolment and implies their intention to use it, is requesting for assistance, i.e., help-seeking. Help-seeking is “a behavior performed by individuals who perceive themselves as needing assistance with a problem, whereby the intended outcome of this behavior is addressing the problem faced” (Heerde & Hemphill, 2018, p. 2). In this process, it is the learner who initiates the communication loop: being consciously aware of the need for help, the learner defines the help-seeking goals, estimates the benefits and costs of (not) seeking help, selects the appropriate sources, and obtains and processes help (Karabenick & Berger, 2013; Nelson-Le Gall, 1985; Roll, Alevin, McLaren, & Koedinger, 2011).

The above identified steps, correspond to and describe the help-seeking process. These steps imply a complex mental mechanism, in which the learners' decision to (not) seek help is instigated by intrinsic motivational factors and externalized as different help-seeking strategies. However, what motivates the learners to (not) seek help, which is the distance from motivation to the decision to (not) seek help, and what guides the learners to determine and adjust their help-seeking strategies, accordingly? Is it a matter of achieving high performance, mastering concepts or a challenge to gain deeper and broader understandings? Or, on the contrary, is it the fear of feeling “dumb”, the internal insecurity of losing learning autonomy, or the doubt of self-worth by being exposed to external assistance and feeling incompetent to complete tasks? Answering these questions would shed light on our understanding of the relationship between motivation and help-seeking. Going a step beyond understanding, how addressing these questions could be beneficial in practice, in a meaningful and helpful manner?

7.2. What is already known on the topic

Exploring how motivational factors affect help-seeking has received increased attention in literature, in traditional classroom settings (Mäkitalo-Siegl, Kohnle, & Fischer, 2011; Ryan & Shin, 2011), in cognitive tutors and interactive learning environments (Huet, Moták, & Sakdavong, 2016; Vaessen, Prins, & Jeuring, 2014; Yang & Taylor, 2013) and in online learning

conditions (Finney, Barry, Jeanne Horst, & Johnston, 2018; Hao, Barnes, Wright, & Branch, 2016). Towards understanding what motivates the learners, as well as how and when they make the decision to (not) seek help, the following three parameters require attention and should be explicitly clarified: (a) how help-seeking is coded and measured, (b) how motivation is coded and measured, and (c) how their relationship is explored.

7.2.1. On the measurement of help-seeking in digital learning environments

During the past decade, the focus regarding explaining and understanding help-seeking behavior has shifted to digital learning environments due to the variety of help-seeking opportunities they offer. In these studies, help-seeking was measured via questionnaires, as the learners' self-reported intentions to ask for help (e.g., Hao, Wright, Barnes, & Branch, 2016; Karabenick, 2003; Pajares, Cheong, & Oberman, 2004; Pintrich, Smith, Garcia, & McKeachie, 1991). Measurements via questionnaires are fairly reliable (Pajares et al., 2004), yet, there is scepticism about their external validity due to the gap that often exists between intentions and real behaviors (Gollwitzer, Sheeran, Michalski, & Seifert, 2009).

More recently, help-seeking has been measured directly from tracking and analysing the learners' actual interactions with the help-seeking mechanisms that are available in the learning environments. Typical help implementations include examples of worked-out problems, glossaries or detailed solutions of the on-going problem (Huet, Escribe, Dupeyrat, & Sakdavong, 2011), hints on the steps required to solve a problem, asking the cognitive tutor to complete the exercise (Vaessen et al., 2014), explanations on errors, instructions for solving the problem, videos demonstrating the solutions (Huet et al., 2016), asking teachers/peers for online help, and online searching (Hao, Wright, et al., 2016). Help-seeking was coded in terms of frequencies of requests, as patterns of sequences of choices, or as binary options between taking/not-taking the help.

However, the cognitive help delivered to learners in these cases is quite unlikely to facilitate deeper thinking and processing mechanisms that are acknowledged to promote more permanent learning gains, or induce learners' self-reflection and self-awareness. Such metacognitive skills have been strongly associated with the learners' awareness of *when* they actually need help, the decision to *ask* for it, and the ability to *evaluate* the delivered feedback. Besides, lack of such metacognitive skills has been associated with the underuse of help facilities (Vaessen et al., 2014). To address this issue, some studies considered delivering metacognitive information to assist learners in engaging with the learning task and regulating help-seeking (Daley, Hillaire, & Sutherland, 2016; Roll et al., 2011), without, however, exploring the effect of motivation on help-seeking.

Moreover, researchers attempted to identify and discriminate help-seeking strategies by mining the learners' interaction data from practising help-seeking (Corrin, de Barba, & Bakharia, 2017; Cross, Waters, Kitto, & Zuccon, 2017), with configured advanced machine learning

techniques for pattern recognition. Nevertheless, in these cases, the help-seeking strategies either were not associated with motivation, or the findings did not provide clear evidence on the distinctiveness of the strategies. As such, additional research is required.

7.2.2. On the measurement of motivation

Establishing a link between motivational factors and help-seeking behavior has been intensively studied in different learning contexts. For the measurement of motivation, two factors were mostly considered: goal-orientation and self-efficacy. The majority of studies employed the 2x2 achievement goals framework (Elliot & McGregor, 2001) for measuring goal-orientation. The framework comprises four achievement goals constructs, i.e., mastery-approach, mastery-avoidance, performance-approach, and performance-avoidance goals. The other motivational factor extensively explored in relation to help-seeking is self-efficacy, measured with questionnaires (e.g., Cheng & Tsai, 2011; P. R Pintrich et al., 1991). Self-efficacy plays a major role in how learners approach goals, tasks and challenges (Bandura, 2006).

However, it has been argued that by focusing on the scores *across variables*, instead of using single variables, could contribute to describing the individual in a holistic manner (Magnusson, 1998) and deeper understanding and justifying her decisions and actions, accordingly. The findings from a recent study which adopted this perspective, lack insights from empirical help-seeking interactions (Finney et al., 2018). Considering joint motivational profiles needs to be further examined, combined with the analysis of help-seeking interactions.

7.2.3. On the impact of motivation on help-seeking behavior

Towards explaining how the motivational factors affect help-seeking, variance-based methods, correlations and multiple regression analyses were employed. Using these approaches, many studies consistently associated achievement goals with help-seeking (either as learners' intentions/perception, or as actual interactions). The key advantage of these methods is that they test hypotheses and obtain a concrete result (e.g., mean, p-values), whereas each statistical test offers a single solution that explains the outcome of interest. However, these methods fail to identify important *interrelationships* amongst the motivational variables. Using *only* these methods, it is unclear which *configurations* of the previously identified motivational variables (i.e., not only the combinations of variables, but also the participation/contribution of each variable in the combinations) better explain the distinctiveness in the outcome of interest, i.e., the help-seeking strategies. This is important because different configurations of the same variables may lead to the same or to different outcomes. This limitation is generally considered as a weakness of the variable-based methods, which do not provide information about the possible combinations of the motivational factors and do not allow for a holistic approach that will lead to multiple solutions. Thus, alternative research approaches should be employed.

Another reason why alternative analysis methods are required is that the nature of the relationship revealed from variance-based analyses is *vague*. Specifically, some studies found that

the more mastery-oriented the learners are, the less they seek and use help (Clarebout & Elen, 2009). Other studies identified mastery-orientation as a positive predictor of instrumental help-seeking, because these learners do not perceive help as a threat to their competence and self-worth (Roussel, Elliot, & Feltman, 2011). Yet, some studies could not correlate mastery goals and help-seeking (Hao, Barnes, et al., 2016; Huet et al., 2016), arguing that help-seeking is a threat to high mastery-oriented learners' autonomy. Regarding performance goal-orientation, some studies negatively correlated it with instrumental help-seeking (Huet et al., 2011), because these learners are very likely not to see intrinsic value in mastering course content. Other studies found no link between frequencies of help-seeking and performance goals (Geraldine Clarebout & Elen, 2009), or a positive effect was detected (Yang & Taylor, 2013).

Moreover, the results from studies that reported on the relationship between self-efficacy and help-seeking are *unclear*, as well. In particular, in some studies, high self-efficacy was related to low use of help, and low self-efficacy was related to high executive help-seeking, yet not in a linear fashion (Huet et al., 2016; Williams & Takaku, 2011). Other studies found that self-efficacy is positively related to instrumental help and to information seeking strategies (Cheng & Tsai, 2011; M. C. White & Bembenutty, 2013).

7.2.4. The present study and the research question

Considering the gaps and limitations identified in previous research, the goals of the present study are twofold: (a) it targets at investigating the differences in learners' actual usage of a metacognitive help-seeking mechanism, with respect to their motivational profiles (i.e., combinations of the motivational factors), and (b) it attempts to associate the different help-seeking strategies with multiple configurations of the learners' motivational factors. Thus, the arising research question that guided this study is the following:

RQ: *(a) Are there any differences in the usage of metacognitive help with respect to the learners' motivational profiles? If yes, how significant are these differences? (b) Which are the emerging help-seeking strategies? (c) How are these strategies associated with the motivational profiles?*

Determining the different configurations of motivational factors is expected to advance our knowledge on why learners adopt a specific help-seeking strategy over another, and how their help-seeking behavior reflects their decision to seek help.

7.3. Methods

This section demonstrates the methods employed in an exploratory study. Towards addressing the research question, the help mechanism implemented for the needs of the study delivers task-related metacognitive information to the learners. This information is extracted from the learners' interactions with the learning tasks and is used to increase the learners' awareness and help them regulate their time-spent and effort accordingly (see sub-section 3.3). Furthermore, a combination of variance-based and pattern-based analysis methods is employed.

The variance-based methods are expected to shed light to the statistical differences explored; the pattern-based methods will contribute to discovering the different configurations of goal-expectations and self-efficacy scores (i.e., interrelationships amongst the motivational factors), and the distinctive help-seeking strategies (according to the utilized learners' frequencies, timing and duration of interactions with the help mechanism), as well as to revealing non-linear relationships between motivation and help-seeking (see sub-section 3.4).

7.3.1. Research participants and study design

Overall, 88 undergraduate students (41 females [46.6%] and 47 males [53.4%], aged 19-26 years-old [M=20.334, SD=1.498, N=88]) at a European University were recruited in a self-assessment process for the Management Information Systems II course (related to databases, networks and e-commerce), at the University computers lab. All students had previously used the self-assessment environment.

The students had to answer to 15 multiple-choice questions (from now on referred to as "tasks") within 60 mins; each task had four possible answers, but only one was the correct. The tasks were delivered to the participants in predetermined order. The students were allowed to temporarily save their answers on the tasks, to review them, to alter their initial choices, and to save new answers; they could also skip a task and answer it (or not) later. Moreover, the students could request for task-related analytics visualizations for each task.

Prior to the self-assessment procedure, the level of difficulty of the tasks (easy, medium, hard) was determined by calibrating the tasks using prior assessment results, according to the number of correct answers. Each task's participation on the score was according to its difficulty level, varying from 0.5 points (easy) to 0.75 point (medium) to 1 points (hard), and only the correct answers were considered. If students chose not to submit an answer to task, they received zero points for this one.

The participation in the procedure was optional; it was offered to facilitate the students' self-preparation before the final exams. All participants signed an informed consent form prior to their participation. The informed consent explained to them the procedure and was giving the right to researchers to use the data collected for research purposes. Students were aware that their interactions had been anonymized prior to being analyzed, and that the collected data would be stored for 3 years.

7.3.2. Data collection and measurements

Data were collected with a web-based self-assessment environment (Appendix A). Measures commonly used in the field of learning analytics (e.g., response-times, frequencies) (Papamitsiou et al., 2014, 2017, 2018; Corrin et al., 2017; Kovanović, Gašević, Joksimović, Hatala, & Adesope, 2015) and indicative of the students' help-seeking interactions, were computed from the logged interactions trace data. Table 7-1 illustrates the measures captured and coded.

Table 7-1. Measurements used in the study.

Variable	Name	Description
TAVV	Time-spent on analytics visualizations viewing	<i>The average time the student spends on viewing the analytics visualizations</i>
FAVR	Frequency of analytics visualizations request	<i>How many times the student asks for analytics visualizations</i>

Specifically, Time-spent on Analytics Visualizations Viewing (TAVV) is the total time that the students spend on viewing the analytics visualizations and engage on sense-making. Similarly, Frequency of Analytics Visualizations Request (FAVR) is a counter that increases every time that the students make the respective request.

Moreover, for the measurement of the motivational factors, self-efficacy and goal-expectancy were configured. For self-efficacy (SE) three items were adopted from Bandura (2006), properly modified for the context of online self-assessment. For goal-expectancy (GE) three items were adopted from Terzis & Economides (2011), with a focus on the self-assessment. Goal-expectancy reflects the students' achievement expectations in the self-assessment, expressed in terms of how satisfied they are with their preparation, and their desirable level of success. Due to the fact that this construct has been designed and evaluated particularly for computer-based assessment procedures, it was preferred over the more general 2x2 goal-orientation framework (Elliot & McGregor, 2001). All items were answered on a 7-point Likert-like scale (1= strongly disagree to 7 = strongly agree) and are synopsisized in Table 7-2.

Table 7-2. Motivational constructs and items from the questionnaire

Construct	Items	Description
Self-Efficacy (SE) - Bandura (2006)	SE1	<i>I remember well information presented in class and textbooks.</i>
	SE2	<i>I get myself to study when there are other interesting things to do.</i>
	SE3	<i>I finish my homework assignments by deadlines.</i>
Goal Expectancy (GE) - Terzis & Economides (2011)	GE1	<i>Courses' preparation was sufficient for the self-assessment.</i>
	GE2	<i>My personal preparation for the self-assessment was sufficient.</i>
	GE3	<i>My performance expectations for the self-assessment are high.</i>

The system also calculates the learner performance (LP) for each learner in a zero to ten (0-10) scale, as $LP = \sum_{i=1}^N d_i z_i$ where $z_i \in (0,1)$ is the correctness of the learner's answer on task i , and d_i is the difficulty of the task.

7.3.3. The task-related learning analytics visualizations

During the design of task-related analytics visualizations as on-demand metacognitive feedback, two design models were considered: (a) the Contextualized Attention Metadata (CAM) schema (Wolpers, Najjar, Verbert, & Duval, 2007) for providing coordinated views over the data, and (b) the metacognitive computational model of help-seeking (Aleven et al., 2006) for guiding the desired help-seeking behavior (i.e., the learner should request for help only when she really

needs it, and receives meaningful assistance). Based on these principles, the content of the task-related visualizations includes three simple (easy-to-read) bar/column charts: (a) the average time to answer correctly vs. the average time to answer wrongly vs. the average time to answer the task, (b) the number of correct vs. the number of wrong answers submitted for this task, and (c) the effort expenditure vs. performance (i.e., correctness of answers) for this task. Figures 7-1 and 7-2 illustrate the task-related analytics visualizations delivered to students as metacognitive help. These visualizations are expected to provide fruitful and actionable insight to students about the actual difficulty of the tasks, about the actual effort needed to deal with the tasks, about the time required to allocate on the tasks (i.e., metacognitive attributes that promote deeper thinking, analytics translation, critical evaluation, and self-reflection).

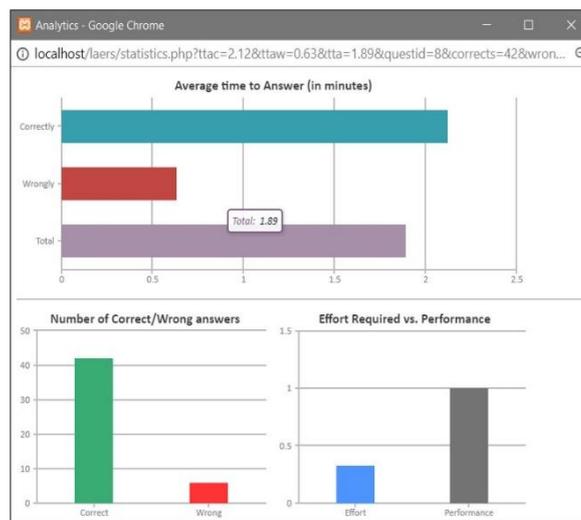


Figure 7-1. The task-related analytics visualizations - information about an easy task

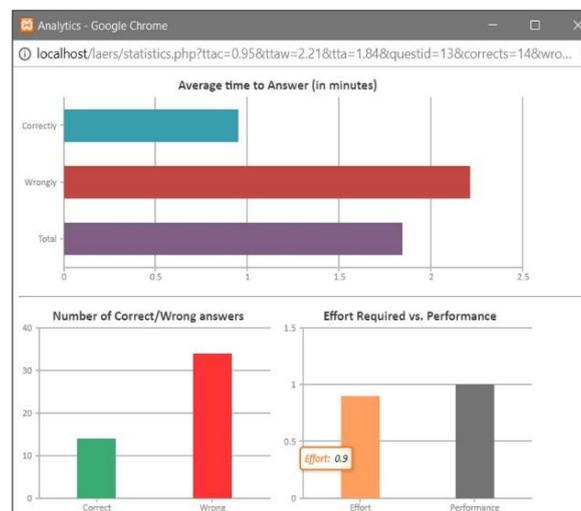


Figure 7-2. The task-related analytics visualizations - information about a hard task

The visualizations tool obtains the necessary temporal and performance indicators from the self-assessment environment, and generates the charts on-demand, by analysing the learners' logged interactions data. For resolving "cold-start" issues, (i.e., absence of data at the first time a task is being viewed by the students) the analytics from former self-assessment procedures were

employed. These analytics were extracted during the calibration of the task pool (see sub-section 3.1), and are updated with the arriving observations.

7.4. Data analysis

7.4.1. Discovering the motivational profiles and help-seeking strategies: clustering learners according to their motivational factors and their help-seeking interactions

For clustering similar learners according to their motivation using the self-reported goal-expectancy and self-efficacy scores, the popular *k-means* algorithm was used. Similarly, *k-means* clustering was applied to determine the different help-seeking strategies, according to the learner's actual interactions with the available metacognitive information (i.e., frequencies of requests and time-spent on viewing the visualizations). The purpose of *k-means* is to group the data by segregating a set of observations into *k* clusters according to a similarity measure. Each observation is assigned to the cluster with the minimum sum of squares of distances between data and the corresponding cluster centroid (Tan et al., 2005). The value of *k* is usually defined using the Silhouette measure of cohesion and separation for the interpretation and validation of consistency within clusters of data (de Amorim & Hennig, 2015). The minimum required sample size, given the desired statistical power level (≥ 0.8) and the probability level ($p < 0.05$), is 85. In this study, the sample size is 88, and thus, it is large enough.

7.4.2. Within group (between subjects) analysis

To investigate differences in the actual usage of help-seeking, i.e., in each one of the analytics parameters of the task-related visualizations usage (TAVV, FAVR), between the different learners' motivational profiles (clusters), ANOVA tests were performed. The impact of the analytics parameters was explored as well, and the η^2 effect size was computed for evaluating the strength of each one of these parameters. Ranges for η^2 effect size are small > 0.01 , medium > 0.06 and large > 0.14 . The decision to use ANOVA tests instead of multiple t-tests was based on the fact that ANOVA controls the Type I error so as it remains at 5%, when the number of groups is higher than two (in this study, three clusters of motivational profiles were detected during the previous step of analysis, as explained in next section).

For the estimation of the motivational factors, i.e., goal-expectancy and self-efficacy, which are latent variables, we used IBM "Statistical Package for the Social Sciences" (SPSS) SPSS-AMOS. The clustering and the within groups analysis tasks were performed using SPSS, version 20.0 for Windows.

7.5. Results

7.5.1. Exploratory data analysis: Convergent validity – Discriminant validity

All criteria for convergent validity are met: all factor loadings on their relative construct exceed 0.70, Cronbach's α and composite reliability of each construct exceeds 0.70 and Average

Variance Extracted values exceeds the variance due to measurement error for that construct (>0.50) (Table 7-3).

Table 7-3. Results for the Latent Constructs

Construct Items	Factor Loadings (>0.7)^a	Cronbach's Alpha (>0.7)^a	Composite Reliability (>0.7)^a	Average Variance Extracted (>0.5)^a
SE		0.878	0.925	0.804
SE1	0.884			
SE2	0.902			
SE3	0.905			
GE		0.866	0.918	0.789
GE1	0.894			
GE2	0.902			
GE3	0.869			

^a Indicates an acceptable level of reliability and validity

SE: Self-efficacy, **GE:** Goal-expectancy

Discriminant validity is also supported since the square root of the AVE of a construct is higher than any correlation with another construct (Fornell & Larcker, 1981) (Table 7-4).

Table 7-4. Discriminant validity

Construct	1	2	3
1. Goal-Expectancy	0.888		
2. Self-Efficacy	0.787	0.897	
3. Learning Performance	0.681	0.644	1.000

7.5.2. Clustering results: motivational profiles and help-seeking strategies

In this study, k-means clustering divided the learners into three groups (k=3), according to their goal-expectancy and self-efficacy scores (average Silhouette=0.7), with 21 (23.9%), 36 (40.9.5%) and 31 (35.2%) records each, respectively, illustrated in Figure 7-3. In addition, k-means identified three configurations of help-seeking strategies with respect to the learners' interactions with the task-related analytics visualizations (average Silhouette=0.7). Specifically, 31 (35.2%), 38 (43.2%) and 19 (21.6%) records were assigned to each cluster, as shown in Figure 7-4. Both help-seeking variables have been standardized in 0-1 range.

As apparent from these figures, the clusters are clearly distinguished from each other. This qualitative observation is further confirmed from the clustering analysis: as seen from k-means ANOVA output (Table 7-5), both goal-expectancy (GE) and self-efficacy (SE) as well as both frequencies of requests (FAVR) and time-spent on processing the metacognitive information (TAVV), contribute the most to the respective clusters, since the F-values are large (and the respective p-values are <.05) for all variables, and their means are significantly different between clusters (shown in Figures 7-5 and 7-6).

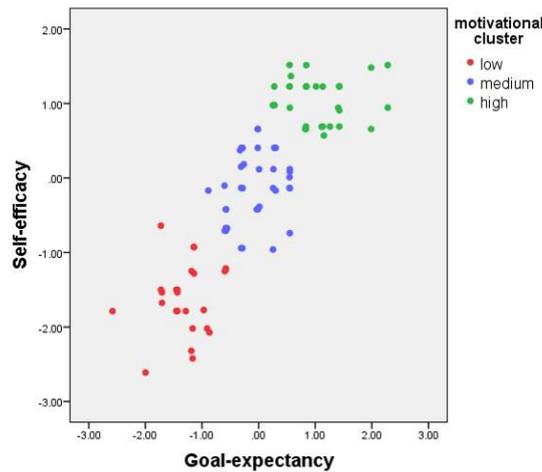


Figure 7-3. Clusters and configurations of goal-expectancy and self-efficacy scores

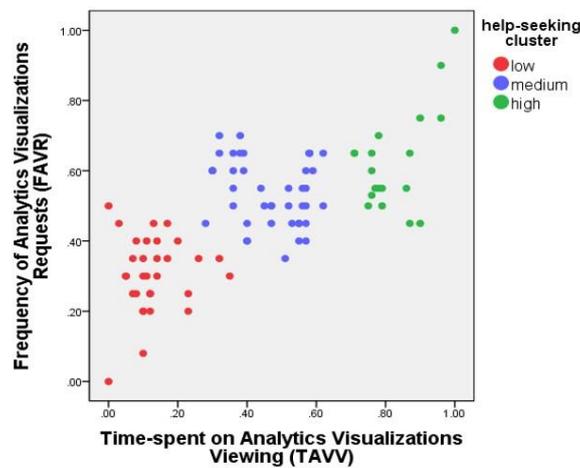


Figure 7-4. Clusters and configurations of help-seeking strategy

However, the clusters have been chosen to maximize the differences among cases in different clusters, and thus, the F-test should be used only for descriptive purposes. The observed significance levels cannot be interpreted as tests of the hypothesis that the cluster means are equal. Post-hoc Bonferroni test revealed that, statistically, the three motivation-based clusters and the three help-seeking strategies-based clusters respectively are significantly different from one another, i.e., the means of their within points are not equal ($p < 0.05$ in all pairwise comparisons – see Appendix C).

Table 7-5. k-means ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
GE	32.558	2	0.269	85	120.934	.000
SE	35.809	2	0.208	85	172.266	.000
FAVR	0.723	2	0.012	85	58.281	.000
TAVV	2.919	2	0.008	85	348.245	.000

GE: Goal-expectancy, SE: Self-efficacy, FAVR: Frequency of Analytics Visualizations Requests, TAVV: Time-spent on Analytics Visualizations Viewing

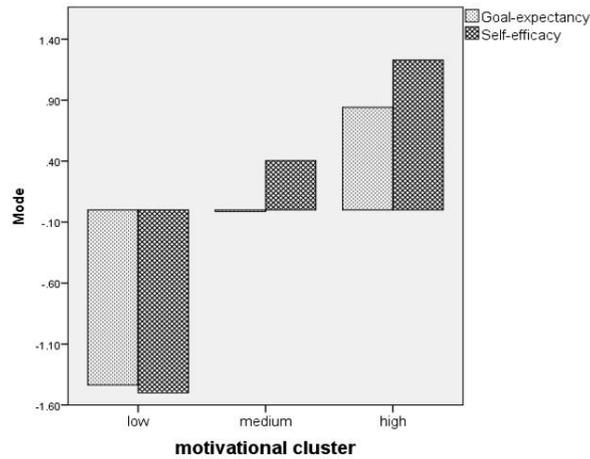


Figure 7-5. Final motivation-based cluster centers

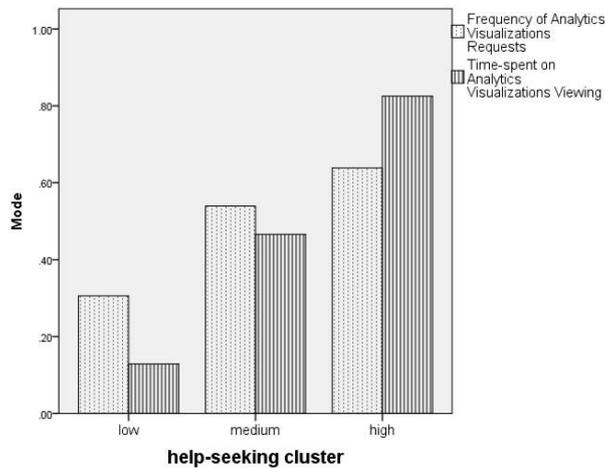


Figure 7-6. Final help-seeking strategies cluster centers

7.5.3. Variance-based analysis: ANOVA results of statistical differences

Table 7-6 presents the results for ANOVA tests for each one of the parameters of analytics visualizations usage (i.e., FAVR, TAVV), with respect to the learners' motivational profiles (clusters). The η^2 effect size was calculated, as well. In all cases, the Levene's test for homogeneity of variances could not reject the hypothesis of equal variances ($sig.>0.05$).

Table 7-6. ANOVA results for the learning analytics factors on the different performance-based student clusters

	F	p	η^2
FAVR	6.196	0.003	0.131*
TAVV	9.617	0.000	0.185*

* $p<0.05$

FAVR: Frequency of Analytics Visualizations Requests, **TAVV:** Time-spent on Analytics Visualizations Viewing

7.5.4. Pattern-based analysis: on the distinctiveness of help-seeking strategies using motivational profiles

Based on the statistical analysis, and since it revealed significant differences in the analytics parameters of the visualizations usage, we looked for patterns of metacognitive help-seeking behavior, with respect to the learners' motivational profiles configurations (exploratory analysis). Figure 7-7 demonstrates the results of the average requests for task-related analytics visualizations per task for each one of the motivation-based student clusters, whereas Figure 7-8 illustrates the respective results for average total-time spent on viewing the visualizations. The difficulty of the tasks increases from the easiest to the hardest (the first eight tasks are easy, the next four are medium, and the last three are hard).

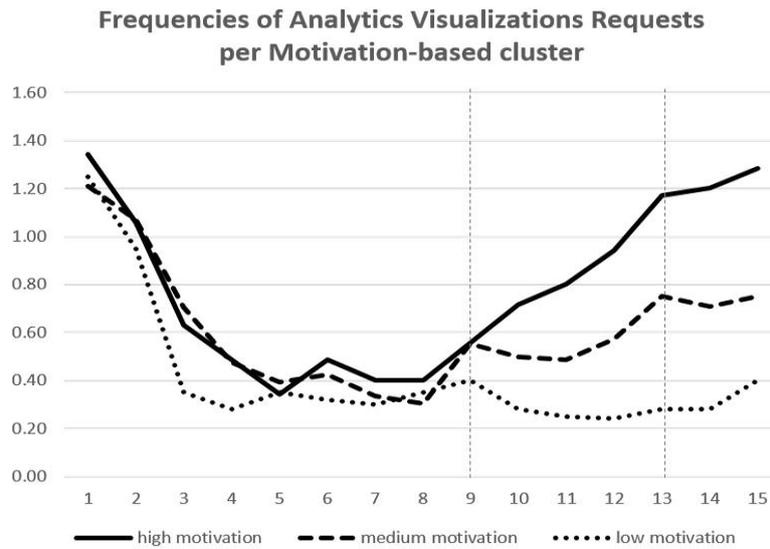


Figure 7-7. Average requests for task-related analytics visualizations per task, per motivational profile

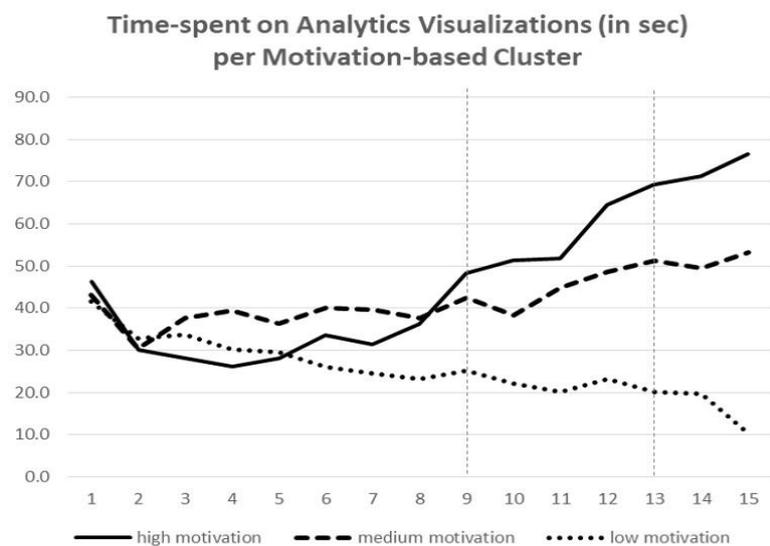


Figure 7-8. Average time-spent on task-related analytics visualizations per task, per motivational profile

As seen from these figures, there are significant differences in the patterns that learners follow regarding the treatment of task-related analytics visualizations, with respect to their motivation.

Next, we plotted the help-seeking strategies in order to identify which motivational clusters are associated with each of them, and explain the distinctive help-seeking profiles (Nagin, 2005): the learners in each of the different the motivational profiles, were identified in *more than one* of the help-seeking clusters, confirming that the same or different configurations of the same motivational factors may lead to the same or to different outcomes. Figure 7-9 demonstrates the percentages of learners from each motivational profile that adopt a help-seeking strategy.

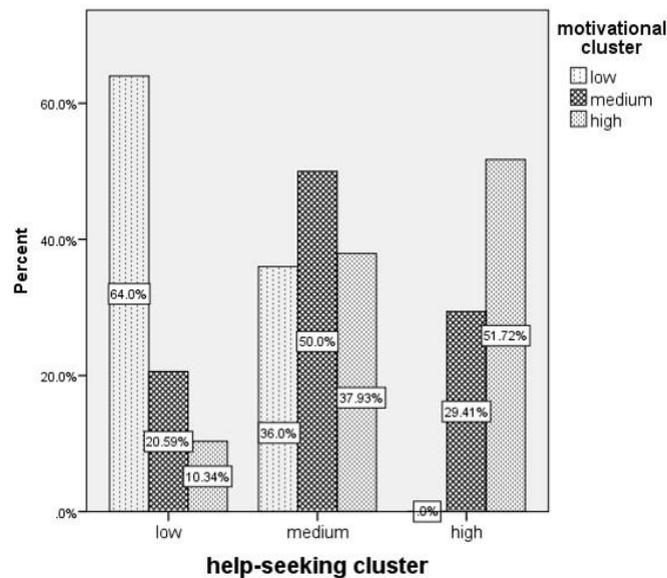


Figure 7-9. Percentages of learners assigned to help-seeking strategy clusters according to motivational profiles

7.6. Discussion

The overall results of this study revealed statistically significant differences in the learners' interactions with a metacognitive help mechanism, with respect to their motivation. Distinctive help-seeking strategies were discovered and associated with different configurations of the learners' motivational factors, as well. In this section, we explain in which ways the demonstrated results are interesting, and how they contribute to broadening current knowledge on what facilitates the learners' mental transition from motivation for seeking help to the decision to (not) do so, and how/when this transition is externalized.

The demonstrated approach was based on three core pillars concerning: (a) the help mechanism implementation, the help-seeking coding and measurement, and the help-seeking strategies configurations, (b) the motivation coding and measurement and the motivational profiles extraction, and (c) the investigation of the relationships and patterns between the motivational profiles and the help-seeking strategies.

Regarding the first pillar, previous studies suggested that delivering metacognitive help to the learners, instead of content-related cognitive help, could improve the learners' help-seeking regulation and on-task engagement (e.g., Daley et al., 2016). In addition, opposed to most studies that measured learners' help-seeking behavior using questionnaires (e.g., Huet et al., 2011), it was argued that employing other research tools which take under consideration the learners' interactions with a help mechanism, could provide valid empirical evidence (Hao, Barnes, et al., 2016). Moreover, discovering and examining help-seeking strategies was proposed towards better reflecting the learners' help-seeking behavior (e.g., Cross et al., 2017).

Considering these parameters, a help mechanism was implemented to support learners by delivering metacognitive information to them (i.e., task-related analytics visualizations). This information is extracted from the learners' interactions with the learning tasks and is used to increase their awareness and help them regulate their time and effort according to the real requirements of the tasks. However, considering learners as the main recipients of learning analytics data is not trivial; there is always a concern that the learners might not know how to translate and make-sense from this information (MacNeill, Campbell, & Hawksey, 2014). Despite this unease, former studies argued that the learners can interpret their own performance indices, yet they reserve a scepticism on how to practically convert this information into action (Corrin & de Barba, 2015). Next, help-seeking behavior was modeled in terms of frequencies of requests and time-spent on viewing this information and sense-making.

The following step was to detect and identify the learners' help-seeking strategies, by mining their interactions. Specifically, the k-means algorithm assigned the learners in three clusters according to their frequencies of help requests and the time-spent on processing this information. Both factors contributed the most to the respective clusters (Table 7-5), and the clusters were significantly different from one another (Bonferroni test, $p < 0.05$ for all pairwise comparisons). The selected $k=3$ is a meaningful solution since the clustering convergence to zero was achieved in less than ten iterations and the Silhouette measure of cohesion and separation regarding the quality of clustering was above 0.5 (average Silhouette=0.7). The cluster sizes were balanced; none of the clusters was underrepresented, indicating efficient (small number of clusters) and effective (comparable numbers of cases) clustering. The clustering results regarding learners' help-seeking profiles revealed the following patterns (strategies):

- learners in help-seeking-cluster 1 (C_{HS1} -low help-seeking) rarely seek help and they do not spent considerable time viewing the delivered information and trying to understand and make-sense from it;
- learners in help-seeking-cluster 2 (C_{HS2} -medium help-seeking) either ask for assistance relatively frequently, yet do not put time to understand the metacognitive information, or they request for help more rarely, but they spent a lot of time on trying to make meaning from the visualizations;

- learners in help-seeking-cluster 3 (C_{HS3} -high help-seeking) regularly request for additional support (medium and high frequency of requests), and they invest considerable time to processing the metacognitive information.

These results are slightly different, yet still comparable to previous ones (Corrin et al., 2017) in terms that in the previous study, the two of the clusters concerned the learners' assessment orientation solely (which does not apply in the present study, since the study itself is assessment-oriented), whereas the other three were similar to those detected in the present study. As such, the current findings extend previous results by identifying help-seeking strategies and examining their distinctiveness, according to the learners' motivational profiles.

Regarding the second pillar, the majority of previous studies adopted the 2x2 goal orientation framework (Clarebout & Elen, 2009; Hao, Barnes, et al., 2016) or explored the effect of self-efficacy as a distinct factor (Huet et al., 2016; Williams & Takaku, 2011). However, it has been argued that the individual is described in a holistic manner by focusing on the scores across variables (Magnusson, 1998). Thus, in the present study, the learners' motivational profiles were constructed and explored as different configurations of the scores of two motivational factors, i.e., self-efficacy and goal-expectations. The goal-expectancy measure was preferred over the goal orientation framework due to being designed for measuring learners' goal expectations in computer-based assessment procedures (Terzis & Economides, 2011). In this study, k-means divided the sample in three groups ($k=3$, average Silhouette=0.7) according to the learners' motivation, and the clusters were significantly different from one another (Bonferroni test, $p<0.05$ for all pairwise comparisons) and with balanced sizes, as well. In order to make sense of the results regarding the learners' motivational profiling, a preliminary look at the clusters outlines the configurations of learners' motivational traits in each of them:

- learners in motivation-cluster 1 (C_{MOT1} -low motivation) score low in both motivational factors, i.e., do not have increased self-expectations and persistence in achieving goals;
- learners in motivation-cluster 2 (C_{MOT2} -medium motivation) exhibit moderate motivation, either because they have not been sufficiently prepared or because they feel they lack abilities to accomplish tasks;
- learners in motivation-cluster 3 (C_{MOT3} -high motivation) score high in both self-efficacy and goal-expectations, indicating high motivation, high preparation and high belief in their abilities.

Regarding the third pillar, the majority of previous studies employed variance-based approaches (e.g., correlations, regression, factor analysis) to understand the nature of relationships between help-seeking and motivation (Hao, Barnes, et al., 2016; Hao, Wright, et al., 2016; Huet et al., 2011). Indeed, in this study, the one-way ANOVA results revealed statistically significant differences between the high, medium and low motivated students with respect to the frequency they request for analytics visualizations ($F(2, 85)= 6.196, p=0.003$), and to the time-

spent on viewing the metacognitive information ($F(2, 85) = 9.617, p = 0.000$). The effect sizes of both analytics parameters were large ($\eta^2 = 0.131$ for FAVR; $\eta^2 = 0.185$ for TAVV), as well. However, employing only variance-based methods fails to identify which *configurations* of the motivational factors (i.e., which different motivational profiles) can explain the distinctiveness in the outcome of interest, i.e., the help-seeking strategies.

The approach demonstrated in this study, followed a combination of variance-based analyses with a pattern-based methodology regarding the help-seeking strategies the learners use, with a focus on the subpopulations of learners identified by similar patterns of values on a *set of variables*. In line with the previously used mixture modeling (Finney et al., 2018), the aim of the pattern-based approach employed in this study was to plot the help-seeking strategies and to understand what external variables (motivational profiles) explain the *distinctiveness* of these strategies (Nagin, 2005); examining correlations between help-seeking and distinct motivational factors does not provide this information. The “mapping” between the two clustering results was conducted to understand the learners’ help-seeking strategies based on their motivational profiles. Plotting the help-seeking patterns according to the learners’ motivational configurations (Figures 7-7 and 7-8), revealed the following findings:

- low motivated learners (C_{MOT1}) are more likely to seek task-related metacognitive help in a more sparse and inefficient manner: they rarely request for analytics visualizations (regardless of the difficulty or the requirements of the tasks), and they do not allocate considerable time on trying to “read” the analytics and process the delivered help. This behavior best fits the strategy identified as C_{HS1} . Indeed, as seen in Figure 7-9, 64% of the low motivated students (C_{MOT1}) are mapped to the low help-seeking strategy (C_{HS1}). The other 36% of these learners are mapped to the medium help-seeking strategy (C_{HS2}). This finding extends previous results that found no link between frequencies of help-seeking and performance goals (Clarebout & Elen, 2009), and is in contrast with other studies which found that low self-efficacy was related to high help-seeking (Huet et al., 2016; Williams & Takaku, 2011).
- medium motivated learners (C_{MOT2}) appear to decide to seek help *when* they “sense” a change in the difficulty (requirements) of the tasks: there is a “peak” on both frequencies of requests and time-spent for task 9 (the first of the medium difficulty tasks) and for task 13 (the first one of the hard tasks). However, the help-seeking requests decrease for the rest tasks, implying a scepticism on whether these learners finally decide to use the help or not. Most of the medium motivated learners (C_{MOT2}) exhibit moderate help-seeking (C_{HS2}): 50% of these learners request for task-related analytics, yet the time they allocate on viewing them is considerably less than the time that highly motivated learners spent, probably indicating an unease with this help format. Almost 30% of the medium motivated learners engage more in help-seeking (they are mapped to the C_{HS3} cluster). This behaviour could imply that these learners are aware of their need for help, and they tend to externalize this need, instead of

avoiding help-seeking. Moreover, as seen in Figure 7-9, the medium motivated learners (C_{MOT2}) are less likely to persistently exhibit low help-seeking behavior, i.e., to belong to C_{HS1} . The lack of previous research – to the best of our knowledge – on explaining the medium motivated learners' help-seeking strategies, accounts for the added value of the present study.

- highly motivated learners (C_{MOT3}) consistently seek help, either in a regular basis or more rarely (Figure 7-7), but still persistently (Figure 7-8). As further confirmed from Figure 5, almost 52% of the highly motivated students (C_{MOT3}) exhibit high help-seeking strategy (C_{HS3}), and 38% of them are moderately seeking help (C_{HS2}). This finding provides additional evidence that mastery-orientation is a positive predictor of instrumental help-seeking (Roussel et al., 2011) (i.e., aided in understanding), and that self-efficacy is positively related to instrumental help and to information seeking strategies (Cheng & Tsai, 2011; M. C. White & Bembenuddy, 2013). Especially for tasks of increasing complexity, the highly motivated students persist on their efforts to handle these tasks, they frequently consult the analytics visualizations, and they allocate significant amounts of time on processing the task-related metacognitive help and making sense about the requirements of the tasks; these students seem determined to use the offered help in order to regulate their on-task engagement.

From the above one can notice that not all combinations of the two motivational factors were detected in the data: for example, no student scored high in goal-expectancy and low in self-efficacy. Similarly, no student scored low in goal-expectancy and high in self-efficacy. Analogous are the results for the help-seeking interactions, as well: the students who spent little time on viewing the visualizations did not frequently request for metacognitive support. More interestingly, the students who spent high amounts of time on making sense from the analytics visualizations did not always highly request for metacognitive assistance; in most of the cases, they dispatched medium help-seeking requests, yet none of them dispatched low requests. This means that not all configurations of the motivational/help-seeking factors are applicable or meaningful, yet those who are, were correctly identified by the analysis method employed.

Moreover, the contribution of the pattern-based analysis is the revealing of the *timing* (*when*) the learners *decide* to seek help, as well as of the *way* (*how*) they externalize this decision. From the preceding analysis it becomes apparent that the majority of learners, regardless of their motivational profiles, ask for analytics visualizations on the first task. From that point on, high-motivated students persistently seek for additional information as the complexity of the tasks increases, low-motivated students successively avoid requesting for metacognitive help, and medium-motivated students follow a more stable pattern and seek for help when the requirements of the tasks increase, but they do not allocate significant amounts of time on processing the information. This implies that these students are aware that they need additional support, but they are uncertain either regarding the translation of the metacognitive information or regarding the actions they should take afterwards, and as such they postpone their decision to

seek help. The findings of this study in conjunction with its research methodology have several implications for research and practice, discussed in next section.

7.7. Implications for research and practice, limitations and conclusions

Help-seeking is an inherently complex mechanism, associated with self-regulatory and motivational aspects of learning. It is a pro-active strategy that is instigated by learners' intrinsic motivational criteria and is externalized as an adaptive behaviour (i.e., not stable). The effects of motivational factors on help-seeking behavior have been intensively explored in literature in different learning contexts and through different research methodologies (from questionnaires to interaction analyses) (Clarebout & Elen, 2009; Hao, Barnes, et al., 2016; Hao, Wright, et al., 2016; Huet et al., 2011; Roussel et al., 2011; Vaessen et al., 2014).

However, *what* motivates the learners to seek/not-see help and *when* the learners decide to finally ask for assistance are difficult to determine using only variance-based methods. The innovation of the approach demonstrated in this study derives from the exploitation of a combination of variance-based methods with pattern-based techniques. The goal was to discover different configurations of the same motivational factors that are associated with and justify the same or different configurations of help-seeking strategies. In other words, the core contribution of this study stems from the adoption of analysis methods that go beyond the identification of single solutions (e.g., regression), to the discovery of behavioral patterns that holistically explain the learners' decision to seek help. Beyond confirming, contradicting or extending previous results, this study is the first one – to the best of our knowledge – that dives into the learners' interactions with the metacognitive support and associates the behavioral patterns (strategies) of this type of help-seeking with the learner's motivational profiles. Thus, methodologically, this study could guide researchers on how to utilize pattern-based methods to make sense of diverse analytics and take design decisions for various user groups.

Although learners with different motivational characteristics may form different sub-groups within a sample, these sub-groups have the potential to be analysed in-depth and receive targeted, learner-centric help-seeking support. Discriminating help-seeking strategies according to the learners' motivational profiles could potentially contribute to better adapting the delivery of help to better facilitate the learners' goals, abilities and expectations. This means, that researchers and practitioners could work together towards designing learning environments enhanced with adaptive, learner-centric help-seeking functionalities. The identified configurations of help-seeking strategies could contribute towards this objective as a roadmap for designing such innovative interfaces, targeting at effective help-seeking. These configurations could also comprise a reference model to train the learners on how to effectively organize their help-seeking strategies. For example, as shown from the findings, learners with medium self-efficacy, clustered in medium motivation, exhibit either medium or low help-seeking behavior (i.e., they request for help, but they do not allocate time on processing it). These learners could be

trained to read and understand the task-related analytics and to make sense from this information. This training is expected to help these learners to better regulate their effort (since they can persist longer to complete tasks, according to their self-efficacy index), and in longer term, to improve their engagement and performance. The adaptive system could provide to such learners analytics with more detailed explanations or it could deliver a shorter version of the dashboard, for example.

Of course, there are limitations as well. The sample size (N=88) of the present study is relatively small, yet sufficient for the analysis methods employed. Larger samples should be explored to further validate the demonstrated findings. Another limitation concerns the help format. From the analysis we could not reject the hypothesis that this metacognitive information did not confuse the learners. Additional experimentation is required with optional help formats (e.g., executive, content-related), to check whether the students (and which of them) would opted to the alternative support. It would be interesting to explore the learners' autonomous choice of help-format, i.e., if they would select executive cognitive help to complete the task, or if they would select the instrumental metacognitive help to regulate their efforts, and under which conditions (e.g., the complexity of the tasks, the time limitations of the self-assessment, etc.). This is within our future work plans.

Chapter 8 : Please Help! What I need to know is....

*“The man who asks a question is fool for a minute;
the man who does not ask is fool for life”*

Confucius

Fostering learners’ engagement and performance with on-demand metacognitive help: the case of task-related analytics visualizations

8.1. Introduction

Providing help to the learners is as old as the learning itself. Assisting learners during their learning is an important part of the process (Richardson, Abraham, & Bond, 2012). Contemporary learning theories highlight the significant role of feedback on the learners’ personal development. Feedback as the (physical/digital, teacher/peer) tutor’s response to the learners’ needs, actions, emotions, attitudes, intentions, etc., is assistive to the learners, either to motivate and reward them, or to help them deal with stressful/confusing learning conditions (Economides, 2009b; Hattie & Timperley, 2007): a key tool for guiding and catalysing learners’ involvement in the self-regulated learning process and goal attainment (Pintrich, 2004).

However, feedback might not impact learning as expected, unless the learners are willing to use it, and even more, to exhibit help-seeking behavior. Help-seeking is “a behavior performed by individuals who perceive themselves as needing assistance with a problem, whereby the intended outcome of this behavior is addressing the problem faced” (Heerde & Hemphill, 2018, p.2). However, help-seeking lurks two unwanted behaviors: (a) the learners might avoid asking for help, and prefer guessing instead, or (b) they might excessively request for hints that directly refer to the answer, without deeper thinking about the hints.

The challenge is to design learner-centered feedback, aiming at motivating learners to ask for assistance at the moment they actually need it, as well as at efficiently supporting their self-regulation (Daley et al., 2016; Roll et al., 2011). The rapid developments in the field of learning analytics have opened new opportunities and perspectives on the design of feedback (Durall & Gros, 2014). The goal is to deliver meaningful information to the receivers, and promote their metacognition: their awareness and sense-making, and finally, their decisions and actions (Schwendimann et al., 2017).

However, the learners are rarely considered as the main recipients of the learning analytics data or given the opportunity to gain access to that information and use it for self-reflection and self-regulation (Durall & Gros, 2014). Yet, when this happens, current learning analytics designs promote antagonism between learners rather than chasing knowledge mastery (Jivet, Scheffel, Drachsler, & Specht, 2017). Moreover, feeding the learners with this information encounters the danger that they may focus too much on their own self (ego), with unwanted

effects on the learning (e.g., might lose motivation if the performance indices are low, or stop trying if the indices are high, just to preserve their reputation and avoid failure).

This raises the question of how to provide meaningful metacognitive feedback to the learners, to encourage efficient help-seeking, to shift their focus on the learning task (rather than feeding the self), as well as to actually help them mastering the skill/knowledge. To address this issue, we investigated the potential of delivering task-related analytics visualizations to the learners (extracted from the learners' interaction trace data, i.e., learner-centered), and explored the effects of this type of on-demand metacognitive help on learners' engagement and performance from a learning analytics perspective.

8.2. Theoretical background and related work

Help-seeking is a pro-active strategy, and it is the learner who initiates the communication channel. It refers to a process that takes place in successive stages including: recognition of the need for help, definition of help-seeking goals, estimation of benefits and costs of seeking/not-seeking help, selection of the appropriate sources, and obtaining and processing help (Aleven, Stahl, Schworm, Fischer, & Wallace, 2003; Nelson-Le Gall, 1985).

These steps imply two facets about help-seeking: (a) its underlying relationship with learner's awareness of knowledge ("*Do I know enough to succeed on my own?*") and regulation of knowledge ("*How can I obtain additional information I may need?*") (Roll et al., 2011), and (b) the difficulty in motivating the learners towards efficient help-seeking ("*Why should I ask for help? Which are the benefits/costs for me?*") (Karabenick, 2011). These facets converge to the complexity of *when* to seek for help, *what* kind of help to seek for (Daley et al., 2016), and *how* to efficiently apply it. Different approaches have been studied in educational practice, using digital learning environments, due to the variety of help-seeking opportunities they offer.

8.2.1. Help-seeking in digital learning environments

Popular help implementations in current digital learning environments include executive help forms, aided in task completion (e.g., glossaries, hint buttons, online search, detailed solutions of the on-going problem) or instrumental help forms, aided in understanding (e.g., breaking problems down, providing to students their own data from the system usage).

Beyond the implementation format for delivering assistance to the learners, help-seeking in digital learning environments has been associated with a capacity for self-regulated learning (Kizilcec et al., 2017; Roll et al., 2011). The general conclusions regarding the impact of help-seeking on self-regulated learning gains are *inconclusive*. Higher learning gains are achieved when students pause to think and reason a hint, and elicit its implications (Shih et al., 2008), when time spent is properly allocated on help-seeking during problem solving (Arroyo & Woolf, 2005) or when the level of teacher's guidance is low, in highly structured classroom scripting conditions (Mäkitalo-Siegl et al., 2011). On the contrary, the achieved learning outcomes are low when learners intentionally misuse the help features and request for hints at a random time to obtain

answers (Aleven et al., 2003; R. Baker et al., 2004), or frequently use executive help (Mathews, Mitrović, & Thomson, 2008).

Moreover, although results have shown that effective help-seeking seems to depend largely on motivation, yet, the conclusions are *conflicting*, as well. Learners who mostly need help (e.g., due to lack of understanding on the topic, or to being afraid of being embarrassed if ask) are those who tend to avoid to find it in a timely manner. Yet, students in less need for support, often exhibit higher help-seeking behavior, by accessing hints and requesting for additional resources (Broadbent & Poon, 2015; Puustinen & Rouet, 2009; Rosé et al., 2015; Stahl & Bromme, 2009). Other studies shown that more mastery-oriented students use help less (Clarebout, Horz, & Elen, 2009; Kizilcec et al., 2017), and they achieve better the learning outcomes (Wood & Wood, 1999).

8.2.2. From self-regulation to metacognition: learning analytics approaches

The inherent complexity of help-seeking behavior as a process that incorporates the learners' self-regulatory and motivational aspects, is reflected on the inconclusiveness of the previously reviewed results. The conclusions drawn in these approaches were extracted mostly from the analysis of self-reported surveys. The need for more in-depth understanding of *why*, *when* and *how* learners seek help, as well as *which* are the effects of the respective help-seeking patterns on the learning gain, might be addressed with the exploitation of analytics methods.

Indeed, learning analytics have been employed by researchers from two perspectives. The first one concerns their capacity to identify and discriminate help-seeking patterns, by mining the learners' interaction data from practising help-seeking (Corrin et al., 2017; Cross et al., 2017). These approaches were explored in MOOCs and online communities, in which the traditional help formats are usually difficult to efficiently apply. In both cases, the authors configured machine learning techniques for pattern recognition, with encouraging, yet not directly applicable results: the authors acknowledged that there is still a long way to go until transforming these patterns into specific help facilities.

The second perspective involves the potential that learning analytics have to deliver metacognitive information to the learners about their own usage of the help-seeking facilities of the learning environment, and as such, to make explicit to the learners their own help-seeking behavior (Daley et al., 2016; Roll et al., 2011). The authors examined the degree to which the learners could translate the delivered metacognitive information, as well as the degree to which this information could trigger changes in learners' help-seeking. The results demonstrated a trend that metacognitive feedback can impact help-seeking behavior and learning performance, yet, these approaches either lack the discovery of specific behavioral patterns (Daley et al., 2016), or are difficult to be transferred to more open learning contexts (Roll et al., 2011).

8.2.3. Motivation of the research and research questions

The executive help formats are straightforwardly related to the self-regulatory (i.e., cognitive, motivational and developmental) needs of the help-seekers. However, it is very likely

that the learners shall only complete tasks, without activating self-reflection and metacognitive learning mechanisms. Moreover, other approaches focus on the metacognitive support of learners and mostly on improving their help-seeking skills, making explicit to the learners their own use of the system and identifying help-seeking errors. However, such approaches are very likely to direct attention away from the task and to hinder learners from mastering a skill/content. Moreover, most of these approaches do not deeper explore the actual learners' interactions with the provided help, to further understand the underlying effects of metacognitive feedback on actual engagement, in a learning analytics manner.

We aimed at exploiting and evaluating on-demand, task-related metacognitive feedback as instrumental help, from a learning analytics perspective. Specifically, this study built on the promising analytics visualizations capacity to promote critical thinking, deeper learning and data-driven sense making, and draws attention to task-related information (instead of the self). The core objective was to investigate whether providing help as task-related metacognitive feedback could lead to increased engagement and improved learning gain. The approach adopted a learning analytics perspective to explore the learners' direct engagement with the delivered feedback, based on the actual learners' interaction trace data. The first aim of the study was to investigate the extent to which the learners are capable of translating this kind of analytics visualizations into meaningful and actionable insight, as well as if/how this ability is reflected on their actual interactions/treatment of the visualizations. The second aim was to explore how this metacognitive help affect the learners' actual engagement with the learning tasks and their performance. Thus, the research questions were twofold:

RQ1(a) *Can learners make-sense from the task-related analytics visualizations? If yes, how the actual usage of visualizations is related to the learners' perceptions of visualizations' usefulness?*
(b) *Are there any differences in the usage of task-related analytics visualizations (i.e., in frequency of requests and in time-spent on viewing them) with respect to the learners' level of performance? If yes, how significant are these differences?*

RQ2(a) *Which is the effect of task-related analytics visualizations on learners' performance?*
(b) *Does the exploitation of task-related analytics visualizations contribute to enhancing the learners' performance? If yes, how significant is its effect on learning performance?* **(c)** *Do learners' interpretations of the task-related analytics visualizations actually help them to deeper engage with the task? If yes, how significant is this effect?*

8.3. Research model and hypotheses

The basic assumption of this study is that task-related analytics visualizations are expected to facilitate learners' data-driven sense-making (metacognition) to improve their performance. Thus:

H1: *Task-related analytics visualizations will have a positive effect on Learning Performance*

Moreover, the more the students use the analytics visualizations, it is more possible that they perceive them as helpful and meaningful. Thus, it is assumed:

***H2:** Task-related analytics visualizations will have a positive effect on Perceived Usefulness of Visualizations*

The rest of the research hypotheses on the effects of the analytics visualizations on engagement (i.e., on response-times and effort) are outlined as follows:

8.3.1. Response-time

Response-time is defined as the total time that the learners spend on interacting with a learning task. When the learners have to submit an answer to the specific task (i.e., a question or a problem), then the response-time can be discriminated according to the correctness of the submitted answer. In this case, it indicates the respective time-spent the learners constantly aggregate on answering the task correctly or wrongly (Papamitsiou et al., 2014). When learners view analytics visualizations about the tasks and receive information about how other students answer to these tasks (i.e., time-spent, correctness of answers, effort), it is more likely that they will critically assess the requirements of the tasks (e.g., difficulty). Thus, it is more possible that they will spend more time to answer on these tasks correctly, and consequently, they will accumulate less time on tasks that they will finally answer wrongly. Therefore:

***H3:** Task-related analytics visualizations will have a positive effect on Time to Answer Correctly*

***H4:** Task-related analytics visualizations will have a negative effect on Time to Answer Wrongly*

8.3.2. Effort

Effort is “the motivational state commonly understood to mean trying hard or being involved in a task. Effort is increased when the subject tries harder, when there are incentives to perform well, or when the task is important or difficult” (Humphreys & Revelle, 1984, p. 158). Thus, effort is about how much engaged the learners are in completing the tasks. We hypothesize that when the learners engage in translating the analytics and making-sense about them, it is more possible that they will remain engaged in these tasks, and they will demonstrate high effort exertion on dealing with them. Thus:

***H5:** Task-related analytics visualizations will have a positive effect on Effort*

Figure 8-1 illustrates the relationships among the factors considered in the model. The dashed arrows represent the relationships between factors that have been previously explored (Papamitsiou et al., 2016, 2014), whereas the rest arrows represent the research hypotheses explored in the present study.

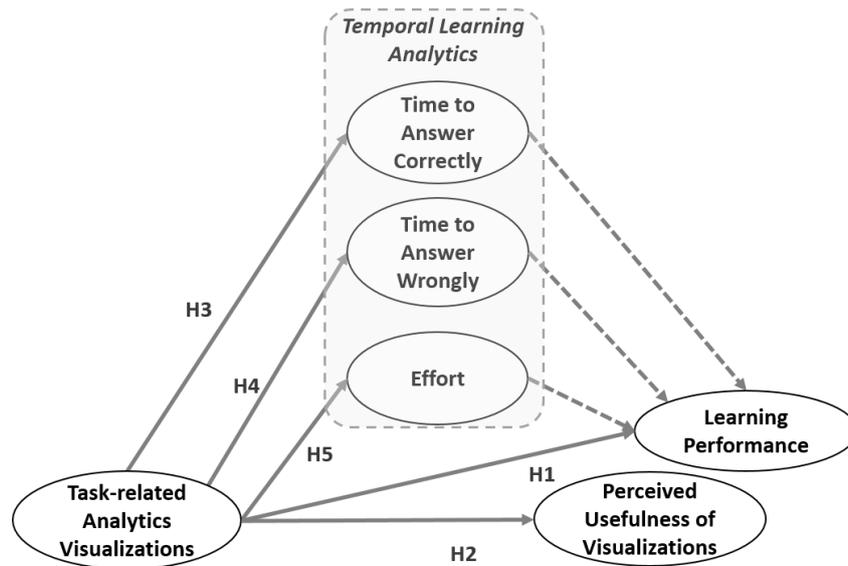


Figure 8-1. Overall research model and factor relationships with hypotheses

8.4. Methodology

8.4.1. Research participants and study design

Overall, 174 undergraduate students (93 females [53.4%] and 81 males [46.6%], aged 19-26 years-old [M=20.582, SD=1.519, N=174]) at a European University were enrolled in a self-assessment process for the Management Information Systems II course, at the University computers lab, for 60 mins. The study followed an experimental design (Cobb et al., 2003). All students had previously used the self-assessment environment, and they were randomly assigned into two groups: 88 students (50.6%) were assigned to the “*visualizations*” group (i.e., the treatment group), and 86 students (49.4%) were assigned to the “*no-visualizations*” group (i.e., the control group).

All students had to answer to 15 multiple-choice questions (from now on referred to as “tasks”); each task had four possible answers, but only one was the correct. The tasks were delivered to the participants in predetermined order. The students were allowed to temporarily save their answers on the tasks, to review them, to alter their initial choices, and to save new answers; they could also skip a task and answer it (or not) later. Moreover, the treatment group could request for task-related analytics visualizations for each task (see sub-sections 8.4.2).

Prior to the self-assessment, the difficulty of the tasks (easy, medium, hard) was determined by calibrating them using prior assessment results, according to the number of correct answers. Each task’s participation in the score was according to its difficulty, varying from 0.5 points (easy) to 0.75 points (medium) to 1 point (hard), and only the correct answers were considered. Students received zero points for task they chose not to submit an answer.

The participation in the procedure was optional; it was offered to facilitate the students’ self-preparation before the final exams. All participants signed an informed consent form prior to their participation. The informed consent explained to them the procedure and was giving the

right to researchers to use the data collected for research purposes. Students were aware that their interactions had been anonymized prior to being tracked and analyzed, and that the collected data would be stored for 3 years.

8.4.2. Data collection and measurements

Data were collected with an online self-assessment environment (Appendix A). For both groups, measures commonly used in the field of learning analytics and acknowledged to satisfactorily explain students' performance (e.g., response-times, frequencies) (Papamitsiou et al., 2018) were computed from the logged interactions trace data. In addition to these measures, for the treatment group, other similar measures, indicative of the students' help-seeking interactions, were computed, as well. Table 8-1 illustrates the measures captured and coded for each group.

Table 8-1. Measurements used in the study.

Variable	Name	Description	Treatment Group	Control Group
TTAC	Time to answer correctly	<i>The response-time a student aggregates on submitting correct answers</i>	✓	✓
TTAW	Time to answer wrongly	<i>The response-time a student aggregates on submitting the wrong answers</i>	✓	✓
RTE	Effort	<i>When a student exhibits solution behavior – a measure of engagement</i>	✓	✓
TAVV	Time-spent on analytics visualizations viewing	<i>The average time the student spends on viewing the analytics visualizations</i>	✓	
FAVR	Frequency of analytics visualizations request	<i>How many times the student asks for analytics visualizations</i>	✓	
PUV	Perceived usefulness of visualizations	<i>How useful and helpful does the student perceive the analytics visualizations</i>	✓	

As explained in sub-section 8.3.1, time to answer correctly (TTAC) and time to answer wrongly (TTAW) indicate the respective response-time the students constantly aggregate on answering the tasks. Similarly, Time-spent on Analytics Visualizations Viewing (TAVV) is the time that the students spend on viewing the analytics visualizations and engage on sense-making. Frequency of Analytics Visualizations Request (FAVR) is a counter that increases every time that the students make the respective request. For the effort calculation, the Response Time Effort (RTE) measures the proportion of tasks which the students try to solve (solution behavior) instead of guessing the answers (Wise & Kong, 2005) (Appendix B). For the perceived usefulness of visualizations (PUV), four items were adopted from Venkatesh, Morris, Davis, & Davis (2003) and particularized with a focus on the analytics visualizations (Appendix C). The system also calculates the learning performance (LP) for each learner according to the correctness of the learner's answer on each task and the difficulty of the task. The task-related learning analytics visualizations are demonstrated in Chapter 7.

8.4.3. Data analysis

8.4.3.1. Structural and measurement model

For addressing $RQ1(a)$, $RQ2(a)$ and $RQ2(c)$, the construction of a path diagram that contains the structural and measurement model was conducted with the Partial Least-Squares (PLS) analysis technique (Chin, 1998; Tenenhaus et al., 2005). Our sample size exceeds the recommended value of 40, i.e., 10 times larger than the number of items for the most complex construct. For the measurement and the structural model we used SmartPLS 3.2.

8.4.3.2. Between group analysis

Regarding $RQ2(b)$, independent samples t-test was used to investigate the impact of task-related analytics visualizations on learning performance. The minimum required total sample size and per-group sample size, given the probability level ($p < 0.05$), the anticipated effect size (Cohen's $d > 0.5$), and the desired statistical power level (≥ 0.8), is 128 and 64 respectively. In our study, the sample size is 174, and the sub-group sizes are 88 and 86 respectively, which are large enough to justify the above mentioned parameters. The statistical significance of the differences with respect to the learners' achievement scores was estimated between the treatment group and the control group of this study (see Supplementary material).

8.4.3.3. Within group (between subjects) analysis

Regarding $RQ1(b)$ the students in treatment and control groups were grouped into three classes according to their performance: High-performing: final grade > 7 , Medium-performing: final grade ≥ 5 , and Low-performing: final grade < 5 . Then, ANOVA tests were performed to investigate differences in each one of the analytics parameters of the visualizations usage (i.e., TAVV, FAVR) between the different performance-based student clusters. The impact of these parameters was explored as well, and the η^2 effect size was computed for evaluating the strength of each one of these parameters. The between groups and the within groups analyses were performed using the IBM SPSS, version 20.0 for Windows.

8.5. Results

8.5.1. Convergent validity – Discriminant validity

The results support the measurement model. Table 8-2 displays the construct items' reliabilities (Cronbach's α , Composite Reliability), Average Variance Extracted and factor loadings and confirms convergent validity for the latent constructs. Table 8-3 presents the variables' correlation matrix; the diagonal elements are the square root of the AVE of a construct, which is higher than the construct's highest squared correlation with any other construct, confirming discriminant validity.

Table 8-2. Results for the Latent Constructs of the Measurement Model

Construct Items	Factor Loadings (>0.7) ^a	Cronbach's a (>0.7) ^a	Composite Reliability (>0.7) ^a	Average Variance Extracted (>0.5) ^a
AV		0.719	0.877	0.781
FAVR	0.890			
TAVV	0.877			
PUV		0.849	0.896	0.683
PUV1	0.863			
PUV2	0.853			
PUV3	0.842			
PUV4	0.742			

^a Indicates an acceptable level of reliability and validity.

AV: Task-related Analytics Visualizations **FAVR:** Frequency of Analytics Visualizations Requests, **TAVV:** Time-spent on Analytics Visualizations Viewing, **PUV:** Perceived Usefulness of Visualizations

Table 8-3. Measurement Model (Discriminant validity) for the treatment group (n=88)

	1	2	3	4	5	6
1. Effort	1.000					
2. Time to Answer Correctly	0.246	1.000				
3. Time to Answer Wrongly	-0.219	-0.181	1.000			
4. Task-related Analytics Visualizations	0.397	0.482	-0.411	0.883		
5. Perceived Usefulness of Visualizations	-0.115	0.434	-0.219	0.326	0.826	
6. Learning Performance	0.404	0.656	-0.590	0.707	0.404	1.000

8.5.2. Testing hypotheses

A bootstrap procedure with 3000 resamples was used to test the statistical significance of the path coefficients (β value) in the model. The results for the hypotheses are summarized in Table 8-4 for the treatment group (n=88).

Table 8-4. Hypothesis testing results

Hypothesis	Path	β	<i>t</i>	<i>P</i>
H1	AV → Learning Performance	0.283	3.993*	0.001
H2	AV → Perceived Usefulness of Visualizations	0.326	3.431*	0.002
H3	AV → Time to Answer Correctly	0.482	4.789*	0.000
H4	AV → Time to Answer Wrongly	-0.411	4.251*	0.000
H5	AV → Effort	0.397	3.287*	0.003

* $p < 0.05$, **AV:** Task-related Analytics Visualizations

All hypotheses **H1**, **H3**, **H4** and **H5** regarding the direct effect of task-related analytics visualizations on learner engagement and performance were strongly supported.

8.5.3. Overall Model Fit

According to these results, the suggested model explains almost the 79% of the variance in performance for the treatment group, which is statistically significant. The cross-validated predictive relevance of the model was also confirmed ($Q^2 = 0.736$). Table 8-5 synthesizes the total effects of the selected factors, as well as the fit indices for the proposed model.

Table 8-5. Fit indices, Direct, Indirect and Total effects

<i>Endogenous</i>	<i>R</i> ²	<i>Q</i> ²	<i>Exogenous</i>	<i>Dir. effect</i>	<i>Indir. effect</i>	<i>t</i>	<i>Total effect</i>
LP	0.787	0.736	AV	0.283	0.423	10.243*	0.707
TTAC	0.231	0.215	AV	0.482		4.789*	
TTAW	0.170	0.162	AV	-0.411		4.251*	
RTE	0.158	0.144	AV	0.397		3.287*	
PUV	0.250	0.167	AV	0.326		3.431*	

* $p < 0.05$

AV: Task-related Analytics Visualizations, **LP:** Learning Performance, **TTAC:** Time to Answer Correctly, **TTAW:** Time to Answer Wrongly, **RTE:** Effort, **PUV:** Perceived Usefulness of Visualizations

The measurement results are summarized in Figure 8-2.

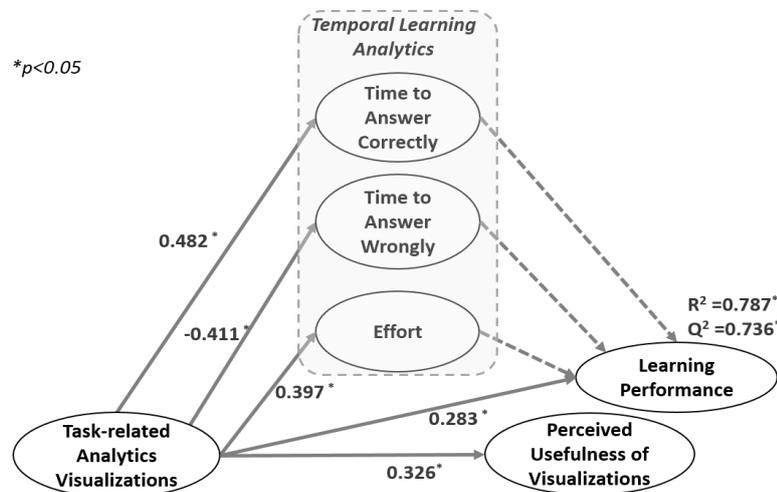


Figure 8-2. Path coefficients of the research model, overall variance explained (*R*²) for test score and cross-validated predictive relevance (*Q*²).

8.5.4. Independent samples t-test and ANOVA results

Table 8-6 demonstrates the descriptive statistics for the two groups with respect to the learning performance, whereas table 8-7 depicts the independent samples t-test results regarding students' learning outcomes between the treatment group and the control group. In this table, the last column illustrates the Hedge's *g* effect size.

Table 8-6. Descriptive statistics for performance for the treatment and control groups

Group	N	Mean	Std.Dev (SD)
Treatment	88	5.534	1.735
Control	86	4.372	1.681

Table 8-7. Independent samples t-test results for learning performance

Groups	F	df	t	95% CI		Hedges' g
				Lower	Upper	
Treatment vs. Control	0.009	172	4.486*	0.65069	1.67331	0.68

* $p < 0.05$

As seen from this table, there were significant differences in performance for the treatment and the control group, and the effect of task-related analytics visualizations on performance was large ($g=0.68$). Moreover, Table 8-8 presents the results for ANOVA tests for each one of the parameters of analytics visualizations usage (i.e., FAVR, TAVV). The η^2 effect size was calculated, as well. In all cases, the Levene's test for homogeneity of variances could not reject the hypothesis of equal variances ($sig.>0.05$).

Table 8-8. ANOVA results for the learning analytics factors on the performance-based clusters

	F	p	η^2
Frequency of Analytics Visualizations Requests	23.002	0.000	0.351
Time-spent on Analytics Visualizations Viewing	19.073	0.000	0.310

* $p<0.05$

Since the statistical analysis revealed differences in the parameters of analytics visualizations usage, we looked for patterns of metacognitive help-seeking behavior, with respect to the performance-based learner clusters (exploratory analysis). Figure 8-3 illustrates the results of the average requests for task-related analytics visualizations per task for each one of the performance-based learner clusters. Figure 8-4 demonstrates the respective results for average time-spent on viewing the visualizations. The difficulty of the tasks increases from the easiest to the hardest (tasks 1-8 are easy, tasks 9-12 are medium, and tasks 13-15 are hard).

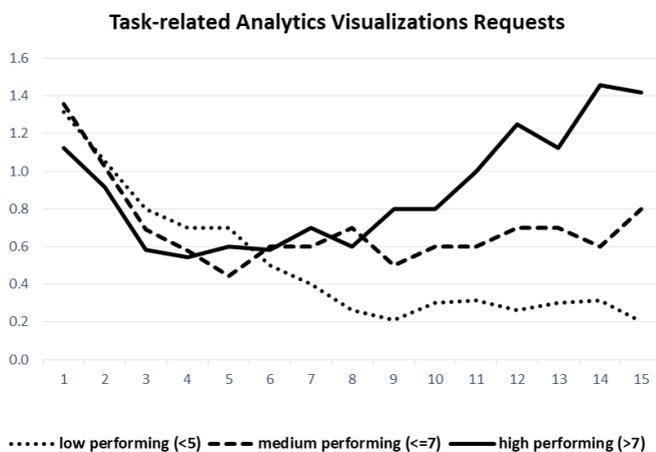


Figure 8-3. Average requests for task-related analytics visualizations per task

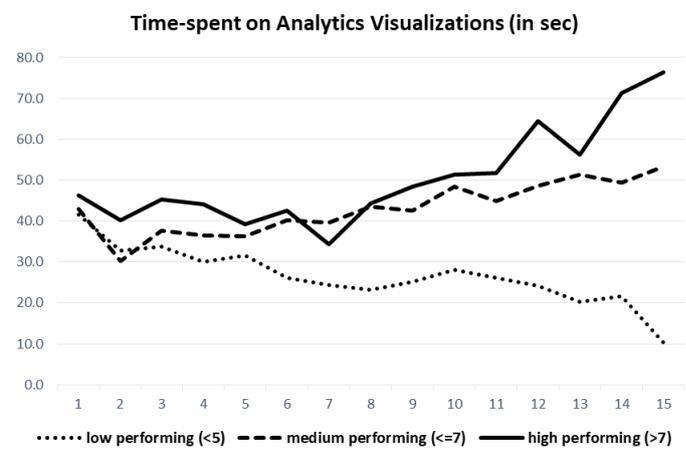


Figure 8-4. Average time-spent on task-related analytics visualizations per task

As seen from these figures, there are significant differences in the patterns that high, medium and low performing students follow regarding the treatment of task-related analytics visualizations. Each of these patterns can be further associated with a help-seeking strategy.

8.6. Discussion

The overall results of this study demonstrate a coherent relationship between the actual usage of the task-related analytics visualizations, the learners' engagement with the task, their

learning performance, and their perceptions about the comprehensibility and helpfulness of the metacognitive help. Additional consistent patterns of help-seeking behavior were identified, as well. In this section, we elaborate on these findings and discuss about how they address the research questions and how they contribute to broadening typical approaches to help-inquiry.

Considering learners as the main recipients of learning analytics data is not trivial; it might promote antagonism between the learners (Jivet et al., 2017), and there is always a concern that the learners might not know how to make-sense from this information (MacNeill et al., 2014). Despite this unease, previous studies argued that the learners can interpret their own performance indices, yet they reserve a scepticism on how to practically convert this information into action (Corrin & de Barba, 2015). The innovation of this work derives from exploiting easy-to-read *task-related* analytics visualizations in order to provide to the learners instrumental information about the tasks, and investigates *how* they actually use it, *how* they adjust their answering behavior, and *what* they believe about its usefulness.

In particular, the instrumental, metacognitive information about the tasks was extracted from the logged learners' interactions with the tasks (i.e., learner-centered), and was delivered to the learners on-demand, as simple bar/column charts. Next, help-seeking was modeled in terms of frequencies of requests for this additional information and time-spent on viewing it and sense-making. The consistency reliability of the respective factor was confirmed ($\alpha=0.72$).

Towards addressing **RQ1(a)**, we explored the relationship between the actual usage of the visualizations and the learners' perceptions about the usefulness of this type of information. The analysis shown a strong positive effect of using the visualizations on their perceived usefulness ($\beta=0.326$, $t=3.431$, $p=0.002$), confirming hypothesis **H2** (Table 8-4). This finding is in agreement with previous results that reported on the learners' ability to read and understand analytics visualizations as metacognitive feedback (Corrin & de Barba, 2015; Daley et al., 2016). Moreover, a consistent pattern of help-seeking and perceptions about the usefulness was revealed, as well. Specifically, students who extensively used the analytics visualizations, scored high in perceived usefulness of the visualizations, as well. In other words, students who regularly asked for additional task-related data and engaged into sense-making from the analytics visualizations, also found them meaningful. On the contrary, students who rarely used the help-seeking facility (i.e., low usage of the visualizations), considered them as confusing (i.e., low score in PUV). Moreover, students who critically assessed the visualized information and inferred the actual difficulty/requirements of the tasks, also perceived the delivered help as informative and actionable, and regulated their time-allocation and effort expenditure accordingly. The added value of this finding is that, in response to previous claims (Corrin & de Barba, 2015), it provides a preliminary insight on *how* learners could transfer the knowledge inferred from metacognitive information into practice and convert it into action: using the visualizations, they identify critical tasks and regulate their engagement accordingly.

In addition, towards addressing **RQ1(b)**, the one-way ANOVA revealed statistically significant differences between the high, medium and low performing students with respect to the frequency they requested for analytics visualizations ($F(2, 85)=23.002, p=0.000$), and to the time-spent on viewing the metacognitive information ($F(2, 85)=19.073, p=0.000$). The effect sizes of both analytics parameters about the actual usage of help-seeking were strong ($\eta^2=0.351$ for FAVR; $\eta^2=0.310$ for TAVV), as well (Table 8-8).

Combined with the results from the exploratory analysis (i.e., the graphical representation of help-seeking behavior with respect to the performance-based learner clusters), this finding can be interpreted as follows: high performing students use the analytics visualizations more often and they allocate considerable time to think and reflect about the received information, and elicit its implications. On the contrary, low-performing students rarely seek for assistance and request for metacognitive hints about the tasks (probably because they feel uncomfortable with this type of information and they don't know how to use it, as explained before). This finding provides additional empirical evidence to previously reported results that associated higher learning gains with time allocated on help-seeking and hint reasoning (Arroyo & Woolf, 2005; Shih et al., 2008). Furthermore, this finding is in line with other research works that claim that students in need usually don't seek for help, while students who can achieve higher – even without additional support – tend to ask for complementary hints and resources (Broadbent & Poon, 2015; Puustinen & Rouet, 2009; Rosé et al., 2015; Stahl & Bromme, 2009). It is also aligned with previous studies that associated lower learning gains with random help requests and misuse of help (Aleven et al., 2003; Baker et al., 2004). However, it should be noted that in the previous studies, the type of help was not learner-centered metacognitive information, and in no case, task-related.

Beyond confirming previous results, this study is the first one – to the best of our knowledge – that dives into the learners' interactions with the metacognitive support and associates the patterns of this type of help-seeking with the performance-based learner clusters. From the exploratory analysis (Figures 8-3, 8-4), it becomes apparent that the majority of students ask for analytics visualizations on the first task. From that point on, high-performing students seek for additional information mostly on the hard tasks, low-performing students successively avoid requesting for metacognitive help, and medium-performing students follow a more stable pattern and seek help on most of the tasks, regardless of their difficulty, but do not allocate significant amounts of time on processing the information. This implies that these students are aware that they need additional support, they seek for it, but they are uncertain regarding the translation of that information or the actions they should take afterwards.

Moreover, previous research yielded *inconclusive* results regarding the effect of help-seeking on performance, and precisely, on self-regulated learning gains (section 8.2.1). As such, additional research on the topic is required. This contribution attempted to shed light to this

relationship from a learning analytics perspective. Beyond exploring the direct impact of task-related metacognitive help on performance, this study also investigated possible indirect effects, and examined whether the visualizations foster learners' engagement with the tasks.

Specifically, towards addressing **RQ2(a)** and **RQ2(c)**, the effects of help-seeking (i.e., the actual usage of analytics visualizations) on engagement and performance were explored through a structural and measurement model. The overall variance in performance explained by the included factors reaches the statistically significant 78.7%. The analysis revealed a positive (direct and indirect) effect of using the visualizations on performance (Table 8-5), while additional analysis for the evaluation of the strength of this effect revealed a large effect size (Hedge's $g=0.68$) (Table 8-7) – addressing **RQ2(b)**. In more detail, the direct effect of help usage on performance was strong ($\beta=0.283$, $t=3.993$, $p=0.001$), confirming hypothesis **H1** and providing empirical evidence that extends current knowledge on the impact of metacognitive feedback on performance. Moreover, help-seeking was found to be a strong determinant of learners' engagement with the learning tasks, both in terms of effort ($\beta=0.397$, $t=3.287$, $p=0.003$) and response-times ($\beta=0.482$, $t=4.789$, $p=0.000$ for TTAC; $\beta= -0.411$, $t=4.254$, $p=0.000$ for TTAW). Thus, hypotheses **H3**, **H4** and **H5** were confirmed, as well.

Further confirming the exploratory analysis, and the findings from **RQ1(a)**, this finding contributes to hypothesizing that students who consciously seek for help are more possible to be more responsibly involved with the learning tasks, they are more likely to be more careful about their answers and to tend to avoid guessing. Elaborating more on that claim, these students seem to recognize *when* they are in need of additional support. They also seem to be aware that their interactions with the tasks, as well as the answers they finally submit, will affect what the other students receive as analytics. Since the metacognitive information about the tasks is extracted (cumulatively) from all students' interactions with the tasks, the task-related analytics reflect the knowledge/experience of the whole, in a manner. Requesting for that kind of information is like relying on, trusting and translating the “*collective intelligence*” – shifting the knowledge from the individual to the whole. Thus, the students who seek for that support, feel more engaged with the whole and more responsible for their own contribution. This finding opens new research directions towards what Karabenick (2011) highlighted as important: the consideration of help-seeking as a social-interactive strategy.

8.7. Implications for research and practice and Conclusions

Help-seeking is a pro-active learning strategy – initiated by the learner – acknowledged for its catalytic impact on the learning process. Different help formats have been explored in literature regarding their efficiency to foster learners' engagement and performance. However, in digital learning environments, the findings are inconclusive. Learning analytics have opened new perspectives on studying help-seeking behavior and supporting learners' help-inquiry.

We demonstrated a holistic study on the effects of task-related metacognitive help on learners' engagement and performance, by exploiting learning analytics. The objective was to explain *how* the actual usage and translation of task-related analytics can guide learners' sense-making and action (i.e., self-regulation) and to extend current knowledge on the topic. The findings of this study in conjunction with its research methodology have several contributions to the literature and numerous implications for research and practice.

Methodologically, the current study exploits learning analytics from two perspectives: (a) their direct feeding to the learners as metacognitive information about the tasks, and (b) their adoption as a consolidated research method to study the actual usage of metacognitive help, its effect on engagement, performance, and the learners' perceptions about its usefulness. As such, regarding its methodology, this study contributes by employing analytics both for the measurement of help-seeking, as well as for the analysis and reporting of its effect on learning. The analysis attempted to shed light to the previous inconclusive or conflicting results (Arroyo & Woolf, 2005; Baker et al., 2004; Broadbent & Poon, 2015; Kizilcec et al., 2017; Rosé et al., 2015), broadening the insights on *why*, *when* and *how* learners inquire for help, with respect to their performance-based classification. Consistent patterns of help-inquiry, engagement and performance were revealed (Table 8-4), as well as statistically significant differences between the high, medium and low performing students' help-seeking strategies (Table 8-7).

Furthermore, the study also evaluated a structural model for explaining engagement and performance by associating help-seeking with response-times and effort. The model explains almost 80% of the variance in performance. The large effect size of help-seeking on performance highlights the need to further investigate the role of task-related metacognitive help in the learning process (Table 8-5). Additional motivational factors, associated with help-seeking (e.g., goal-setting) should be incorporated into the model to further justify help-seeking.

Moreover, the study adds to the help-seeking literature by introducing and evaluating a task-related analytics approach as on-demand instrumental help. In other words, it suggests delivering to the learners metacognitive information about the tasks instead of their own behavior, in order to support them dealing with the tasks. Former studies examined the case of providing to the learners self-related metacognitive information to increase their awareness regarding their own usage of the help facilities and to support them on improving their help-seeking skills (i.e., reduce help-seeking errors) (Daley et al., 2016; Roll et al., 2011). However, neither of these studies investigated and assessed the variation in the level of learners' deeper engagement with the learning task/content that is attributed to the usage of the provided support, nor they did it in a learning analytics fashion. The added value of the shift from self-related to task-related analytics derives from the multiple benefits the latter have for the learners – as shown from the findings in this study – outlined as follows:

- The task-related analytics seem to boost the learners to practice higher order critical thinking for sense- and decision-making: the learners evaluate, assess and mostly contribute to generating peripheral information about the task (i.e., that is not directly related to its content), and they use this information to adjust their actions accordingly. The findings indicate that the more the learners consult this information and engage into translating it, the more efficient their on-task engagement and the better their learning performance (Table 8-5). As such, it appears to be mandatory to train the learners on how to read, interpret and use the information from the analytics visualizations in order to reduce antagonism and to improve data-driven decision-making. The easy-to-read data demonstrations used in this study seem to facilitate that goal.
- The fact that the information is about the task and not the learner herself adds in drawing attention on the task rather than the self. This type of information increases the learners' awareness about the actual requirements of the learning task because the analytics are measurements from actual interactions with the task. Consequently, the learners may develop time-regulation competences (i.e., regulation of time-allocation and on-task effort expenditure) according to the actual needs of the task. For example, if it's taking too long, probably something is wrong; if it's taking too short, probably something is wrong. In both cases, the learner is aware of the time/effort she needs to put on the task. And this is an example of how the students can convert the information into action.
- Instead of requesting for immediate assistance or getting informed about their own behavior –directly depending on the tutor – through the task-related analytics the learners elicit help from a self-reflecting, task-oriented process. Thus, they maintain their sense of independence from the tutor, potentially contributing to enhancing their autonomous learning capacity. Exploring this relation is within our future work plans.
- The most interesting finding of this study concerns the capacity that task-related metacognitive help has to promote responsible learning from a social-interactive perspective. Elaborating further on that claim, the task-related analytics are collectively generated and extracted from the actual interactions of all learners. The results of this study shown that the more the learners become aware that their engagement with the task affects the analytics the other learners receive, the more careful they might become, and thus, the more responsible for their choices and actions (as seen from the reduced guessing and increased effort). This opens new research directions towards integrating the “collective intelligence” factor in supporting the help-seeking process.

Overall, the findings showcased numerous benefits that the learners could acquire from this help format, and highlighted the need to examine the task-related analytics option more extensively, mostly by considering the social-interactive aspect of the provided support.

Chapter 9 : Practice makes perfect – Building capacity for efficient help-seeking

“The key question to keep asking is, Are you spending your time on the right things? Because time is all you have.”

Randy Pausch

The impact of metacognitive help-seeking on engagement and performance: A longitudinal study using learning analytics

9.1. Introduction

Assisting learners during their learning is an important part of the process (Richardson et al., 2012). It is even more critical, though, when the learners themselves are requesting for help, because it directly reflects their involvement and implies their intention to use the provided assistance. Indeed, help-seeking is a pro-active strategy, and it is the learner who initiates the communication loop: from recognizing the need for help, to defining the help-seeking goals, to estimating the benefits and costs of (not)seeking help and selecting the appropriate sources, the learner finally obtains and processes help (Aleven et al., 2003; Karabenick & Berger, 2013; Nelson-Le Gall, 1985). Help-seeking is “a behavior performed by individuals who perceive themselves as needing assistance with a problem, whereby the intended outcome of this behavior is addressing the problem faced” (Heerde & Hemphill, 2018, p.2).

However, help-seeking lurks two unwanted behaviors: (a) the learners might avoid requesting for help (underuse), or (b) they might excessively ask for hints that explicitly lead to the solution (overuse), without activating deeper thinking mechanisms. For example, the learners might feel that they lose their learning autonomy, since they rely on the instructor’s or the peers’ assistance, and for this reason they might avoid seeking help (Fletcher & Shaw, 2012; Huet et al., 2016). In another example, the learners might not be motivated in learning, and as such, it is very likely that they would not engage in help-seeking, as well (Hao, Barnes, et al., 2016; Huet et al., 2016). From another viewpoint, the learners might continuously ask for help because they actually do not know when they really need it, or might engage in “gaming the system” behavior (R. Baker et al., 2008) because they are highly performance oriented (Aleven et al., 2003).

It has been argued that both overuse and underuse of help-seeking seem to be detrimental to learning; although help-seeking overuse is associated with poor learning (Roll, Baker, Aleven, & Koedinger, 2014) because it bypasses the self-explanation and self-regulation processes, the lack of such metacognitive skills has been associated with underuse of help facilities, as well (Vaessen et al., 2014). The reason is that help-seeking is associated with and involves metacognitive characteristics that include the learner’s knowledge of knowledge (“*Do I know enough to succeed on my own?*”), regulation of knowledge (“*How can I obtain additional*

information I may need?") (Roll et al., 2011), and motivation to seek it ("*Why should I ask for help? Which are the benefits/costs for me?*") (Karabenick, 2011).

The above characteristics strongly indicate the existence of more complicated underlying processes: help-seeking is a self-directed strategy that involves motivational, cognitive, metacognitive and affective mechanisms (Mäkitalo-Siegl et al., 2011; A. M. Ryan & Shin, 2011; Vaessen et al., 2014). In a sense, the help-seeker needs to be persuaded to trust the help-provider in order to feel comfortable and safe to ask help, and that the delivered assistance shall actually help her to deal with the problem, at the moment it occurs. Yet, the help-seeker should be assured by the help-provider that her autonomy is not violated or prohibited, but instead, it is encouraged and promoted (as self-enforced decision/choice to resolve e.g., misunderstandings, when it is actually needed).

Therefore, beyond convincing the learner that help-seeking does not "expose" her to the help-provider (either in terms of knowledge mastery or in other psychological terms, e.g., insecurity), the goal is mostly to support her with appropriate forms of assistance that can actually help her mastering the skill/knowledge and potentially promote her deeper thinking mechanisms and metacognition: her cognitive self-awareness, her sense-making and finally, her decision-making and actions (Schwendimann et al., 2017). The present study targets at encouraging learners to seek help in a way that prevents them from feeling cognitively or psychologically "exposed", and at the same time assists them in their efforts to complete the learning tasks. For this reason, the study delivers on-demand metacognitive help to the learners in the form of *task-related analytics visualizations*. This help format is learner-oriented, but not self-centered, and is expected to provide meaningful information to the learners about how the other learners have treated the same learning tasks. The core idea is to extract analytics from the learners' interactions trace data, to convert them into meaningful indices and to deliver processed information to the learners, to improve their awareness and critical reflection on the actual requirements of the tasks. The study aims at investigating the effect of this help format on learners' on-task engagement and performance over a period of time. For this purpose, this study explores the changes in the same subjects' behavior prior to using the metacognitive help, during taking this treatment and after removing it again. In other words, this longitudinal study targets at examining how the task-related metacognition affects the learners' manipulations of learning tasks and learning performance, and how the learners' behavior changes over time due to the metacognitive help usage.

9.2. Related work

9.2.1. What is already know about the topic

Different approaches on exploring the usage and effects of help-seeking behavior on learning performance have been studied in educational practice, using digital learning environments, due to the variety of help-seeking opportunities they offer. For example, typical

help implementations include worked-out problems, glossaries or detailed solutions of the ongoing problem (Huet et al., 2011), hints on the steps required to solve a problem, asking the cognitive tutor to complete the exercise (Vaessen et al., 2014), explanations on errors, instructions for solving the problem, videos demonstrating the solutions (Huet et al., 2016), asking teachers/peers for online help, and online searching (Hao, Wright, et al., 2016). Help-seeking was coded in terms of frequencies of requests, as patterns of sequences of choices, or as binary options between taking/not-taking the help.

However, the general conclusions regarding the impact of help-seeking on learning gains are *inconclusive*. For example, higher learning gains are achieved when students pause to think and reason a hint, and elicit its implications (Shih et al., 2008), or when time-spent is properly allocated on help-seeking during problem solving (Arroyo & Woolf, 2005). On the contrary, the achieved learning outcomes are low when learners intentionally misuse the help features and request for hints at a random time to obtain answers (Aleven et al., 2003; Baker, Corbett, Koedinger, & Wagner, 2004), or frequently use executive help (Mathews et al., 2008).

Moreover, the cognitive help delivered to learners in the above cases is quite unlikely to facilitate deeper thinking and processing mechanisms that are acknowledged to promote more permanent learning gains, or induce learners' self-reflection and self-awareness. Such metacognitive skills have been strongly associated with the learners' awareness of *when* they actually need help, the decision to *ask* for it, and the ability to *evaluate* the delivered feedback.

To address this issue, some studies considered delivering metacognitive information to assist learners in engaging with the learning task and regulating help-seeking (Daley et al., 2016; Roll et al., 2011). For example, in intelligent tutoring system contexts, a geometry "tutored step-based problem-solving environment" (Roll et al., 2011, p. 268) modeled help-seeking by considering factors like the learners' knowledge level, previous help-seeking patterns and time spent on the problem. The system provided corrective feedback to the learners and encouraged them to change their behavior, whenever it detected a "help-seeking error" (p. 268) according to the help-seeking model. In another example, the metacognitive information delivered to students – in order to support their self-reflection and to encourage them to change their help-seeking behavior – was extracted from their own use of the online curriculum, and made explicit to them their own data from the online system. This information included content knowledge (feedback with the correctness of the response), strategic use of the curriculum (exploitation of hints and other available support facilities) and engagement (in terms of difficulty ratings) (Daley et al., 2016).

Nonetheless, delivering efficient metacognitive support is not a trivial task to accomplish. Recent developments in the learning analytics research domain have opened new opportunities and perspectives on the design and delivery of meaningful metacognitive information to the help-seekers, in the form of visualizations dashboards (Durall & Gros, 2014; Martinez-Maldonado et

al., 2016). So far, the dashboards typically visualize the learners' own interaction trace data, targeting at triggering the learners' self-awareness, self-reflection and self-regulation mechanisms (Rodríguez-Triana, Martínez-Monés, Asensio-Pérez, & Dimitriadis, 2014).

Considering learners as the main recipients of learning analytics data has received criticism: it has been argued that it might promote antagonism between the learners (Jivet et al., 2017), and there is always a concern that the learners might not know how to make-sense from this information (MacNeill et al., 2014). Despite this unease, previous studies shown that the learners can interpret their own performance indices, yet they reserve a scepticism on how to practically convert this information into action (Corrin & de Barba, 2015).

Indeed, learning analytics have been employed by researchers from two perspectives. The first one concerns their capacity to identify and discriminate help-seeking patterns, by mining the learners' interaction data from practising help-seeking (Corrin et al., 2017; Cross et al., 2017). These approaches were explored in MOOCs and online communities, in which the more "traditional" help formats (e.g., hints, explanation on errors, directly asking the instructor, etc.) are usually difficult to efficiently apply. In both cases, the authors configured machine learning techniques for pattern recognition, with encouraging, yet not directly applicable results: the authors acknowledged that there is still a long way to go until transforming these patterns into specific help facilities.

The second perspective involves the potential that learning analytics have to deliver metacognitive information to the learners about their own usage of the help-seeking facilities of the learning environment, and as such, to make explicit to the learners their own help-seeking behavior (Daley et al., 2016; Roll et al., 2011). The authors examined the degree to which the learners could translate the delivered metacognitive information, as well as the degree to which this information could trigger changes in learners' help-seeking. The results demonstrated a trend that metacognitive feedback can impact help-seeking behavior and learning performance.

9.2.2. Motivation of the research and research question

Although the demonstrated research results regarding the effect of metacognitive help on learning performance were encouraging, however, all previous studies followed cross-sectional research designs, i.e., describe a group of subjects at one particular point in time, without considering the time dimension, and rely on existing differences rather than change following intervention (Setia, 2016; Turner, 2013). However, individuals' behavior usually changes in essential ways over a period of time. In order to validate the effects of this type of help by considering changes in learners' behavior over time due to the usage of metacognitive support, continuous processes with repeated measurements for the same sample should be employed. In fact, it is difficult to imagine a theory (macro, meso, or micro) being purposely developed to explain a phenomenon at only a single point in time. As such, this study targets at addressing this objective. Specifically, the study considers a task-related metacognitive help format and aims at

investigating the changes in learners' engagement and performance that are caused by the mediating effect of help-seeking. The core research question that guided the research is:

RQ: *Are there any changes in learners' engagement and performance due to receiving metacognitive help, over time? If yes, how significant are these changes?*

9.3. Methodology

9.3.1. Study Design and Research Participants

The study followed a crossover longitudinal research design (Ployhart & Vandenberg, 2010). Crossover longitudinal studies follow the same sample at regular intervals and make repeated observations and measurements of the same variables for the same groups of people; every subject in the sample serves as their own "control", i.e., they belong both to the treatment group and the control group during the different points in time of the measurements. These observations shall enable researchers to track changes in independent variables (predictors - P) over time and to relate these changes to one or more treatment variables (mediators - M) that *might* explain why the changes in dependent variables (outcome - O) occur. Longitudinal designs permit the measurement of difference or change in a variable from one period to another, i.e., the description of patterns of change over time; it is hypothesized that *changes in the predictors and mediators* contribute to *change in the outcome*, and not static levels of some variables predicting static levels in another. Measurements are taken on each variable (P, M, O) over three or more distinct time periods (three is the minimum, but more is better).

During the design of the present study, we had to address the following three core methodological issues, and take decisions accordingly:

(a) Determine the optimal number of measurement occasions and their intervals to appropriately model the hypothesized form of change. Providing equally spaced repeated measurements is better to be avoided (Ployhart & Vandenberg, 2010). In this study, three measurements were carried out in total. The time-distance between the first and second was three weeks, whilst the interval between the second and third measurement was two weeks.

(b) Maintaining the integrity of the original sample can be difficult over an extended period of time. In this study, from the 122 participant students at the first phase, 67 (54.9% of the initial sample) where those who participated in all three phases: this percentage is acceptable (>50% of the initial sample) and maintains statistical power.

(c) It can be difficult to show more than one treatment variables at a time. In this study, the treatment variable (M) is the task-related analytics visualizations usage. The dependent variable (O) is learning performance, and the independent variables (P) include the students' on-task engagement, i.e., response-times and effort.

Specifically, in this study, 122 undergraduate students (55 females [45.1%] and 67 males [54.9%], aged 19-26 years-old [M=20.254, SD=1.411, N=122]) at a European University were

initially enrolled in a self-assessment procedure, consisting of four phases, for the Management Information Systems I course, at the University computers lab, for 60 mins, each phase. The participation in the procedure was optional; it was offered to facilitate the students' self-preparation before the final exams. Due to that "option", the students could take none, one, two, or all the intermediate self-assessment tests. Students who had not taken a test during the first phase were allowed to participate in any of the other phases, if they wanted to. As such, at the second phase, 95 students participated, whereas at the third phase, the attendee number was 86. Those who participated in all three phases were 67 (54.9% of the initial sample). The considered 67 undergraduate students (29 females [43.3%] and 38 males [56.7%], aged 19-26 years-old [M=20.182, SD=1.341, N=67]) had previously used the self-assessment environment (its default version, without the help mechanism) at least one time before the present study.

During the first phase of the study, all students took a self-assessment test: they had to answer to 15 multiple-choice questions (from now on referred to as "tasks"); each task had four possible answers, but only one was the correct. The tasks were delivered to the participants in predetermined order. The students were allowed to temporarily save their answers on the tasks, to review them, to alter their initial choices, and to save new answers; they could also skip a task and answer it (or not) later. At the second phase of the study, three weeks after the first one, the same self-assessment procedure was repeated (with different tasks). In this phase, the students could also request for task-related analytics visualizations for each task (treatment) (explained in section 3.3). The third phase, two weeks after the second one, was exactly like the first phase (the treatment was removed). In the between of the self-assessment tests, the students attended the regular course lectures. The overall process is illustrated in Figure 9-1.

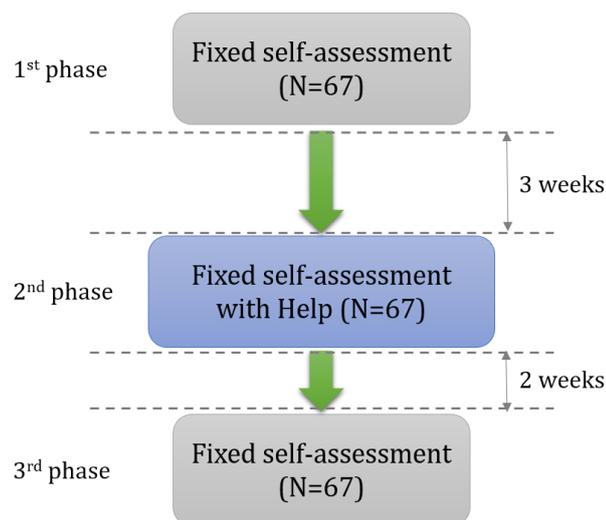


Figure 9-1. Overview of the longitudinal experimental study – Phases, duration and participants

Prior to all the self-assessment procedures, the difficulty of the tasks (easy, medium, hard) was determined by calibrating them using prior assessment results, according to the number of correct answers. Each task's participation in the self-assessment test score was according to its difficulty, varying from 0.5 points (easy) to 0.75 points (medium) to 1 point (hard), and only the

correct answers were considered. Students received zero points for the tasks that they chose not to submit an answer.

All participants signed an informed consent form prior to their participation. The informed consent explained the procedure to them, and was giving the right to researchers to use the data collected for research purposes. Students were aware that their interactions would be tracked, and anonymized prior to being analyzed, and that the collected data would be stored for 3 years.

9.3.2. Data collection and measurements

Data were collected with an online self-assessment environment (Appendix A). In all phases, measures commonly used in the field of learning analytics and acknowledged to satisfactorily explain students' performance (e.g., response-times, frequencies) (Papamitsiou et al., 2018; Corrin et al., 2017; Kovanović, Gašević, Joksimović, Hatala, & Adesope, 2015) were computed from the logged interactions trace data. In addition to these measures, during the treatment phase (i.e., when the task-related analytics visualizations were available), other measures, indicative of students' help-seeking interactions, were computed, as well. Table 9-1 illustrates the measures captured and coded for each phase.

Table 9-1. Measurements used in the study.

Variable	Name	Description	Treatment Phase	Control Phase
TTAC	Time to answer correctly	<i>The response-time a student aggregates on submitting correct answers</i>	✓	✓
TTAW	Time to answer wrongly	<i>The response-time a student aggregates on submitting the wrong answers</i>	✓	✓
RTE	Effort	<i>When a student exhibits solution behavior – a measure of engagement</i>	✓	✓
TAVV	Time-spent on analytics visualizations viewing	<i>The average time the student spends on viewing the analytics visualizations</i>	✓	
FAVR	Frequency of analytics visualizations request	<i>How many times the student asks for analytics visualizations</i>	✓	

Time to answer correctly (TTAC) and time to answer wrongly (TTAW) indicate the respective response-time the students constantly aggregate on answering the tasks (Papamitsiou et al., 2014). Similarly, Time-spent on Analytics Visualizations Viewing (TAVV) is the time that the students spend on viewing the analytics visualizations and engage on sense-making, and Frequency of Analytics Visualizations Request (FAVR) is a counter that increases every time that the students make the respective request. For the effort calculation, the Response Time Effort (RTE) measures the proportion of tasks which the students try to solve (solution behavior) instead of guessing the answers (Wise & Kong, 2005). The system also calculates the learning performance (LP) for each

learner according to the correctness of the learner's answer on each task and the difficulty of the task. The task-related learning analytics visualizations were demonstrated in Chapter 7.

9.3.3. Data analysis

Except for the methodological issues on the study design addressed in section 3.1, prior to running the measurements, one additional analytical issue should be resolved prior to the data analysis. Time is a metric for describing change and is frequently a predictor variable in longitudinal research. As such, defining the metric and coding it, is a core issue (Ployhart & Vandenberg, 2010). In this study, the variable that codes the time dimension is coded as ordinal, implying that there is an order in the measurements, but the interval length between the measurements is not taken into account. Elaborating on this decision, the treatment applied (i.e., the visualizations) was independent of the interval between the measurements because the students did not use it continuously between the self-assessment test, but only during the tests.

9.3.4. Panel data analysis

In preparation for the analysis, the three datasets with the measurements on the on-task engagement and learning performance (i.e., TTAC, TTAW, RTE and LP), collected throughout the study, were merged and transformed to long format, i.e., each subject had a row for each time point, and the repeated measurements were in a single column/variable. An additional identifier was generated as well, i.e., the number of the phase each measurement was collected at, for grouping the observations: we had three groups. This variable was also used to code the time dimension.

Next, a Hausman test was used to differentiate between fixed effects and random effects model. It tests whether the coefficients estimated by the efficient random effects estimator are the same as the ones estimated by the consistent fixed-effects estimator (Hausman, 1978). The result supported the assumption of correlation between observation's errors and predictors, thus, fixed effects model was used.

The initial model included 7 fixed-effects variables: self-assessment phase identifier (group variable), the three on-task engagement variables (i.e., the two response-times variables and the effort variable), and the interactions between the phase identifier (group) and each one of the on-task engagement variables. An interaction is simply two or more variables multiplied together. When including interaction terms in the model, the individual terms should be in the model as well, because otherwise it is impossible to say if the observed effect is caused by the interactions, or by the omitted individual term. The interactions terms reveal if there is a difference in changes of outcome of interest (in this study, the learning performance) between the groups (i.e., the self-assessment phases) due to the intervention (i.e., the usage of task-related analytics visualizations).

To remove nonlinearity from the time and effort variables, a simple log transformation was applied, and the transformed data were used instead of the original. Next, we fitted hierarchical linear mixed models using the restricted maximum likelihood (REML) estimation, having the learning

performance in the self-assessment as the outcome variable. To select the final model, we successively fitted the current model (the initial was the one with all variables), then computed the model's corrected Akaike Information Criterion (AICc) (Hurvich & Tsai, 1989) and removed the variables with the highest p-values. The finally selected model was the one with the smallest AICc.

Our analysis was performed in Stata 12.

9.4. Results

The final model included the self-assessment phase identifier (group variable), the two response-times variables, the effort variable, as well as the interaction between the phase identifier (group) and the two response-times. The hierarchical linear mixed model further revealed that both response times and effort were statistically significant determinants of learning performance in self-assessment tests (Table 9-2), confirming once again the previous findings with other analysis methods.

Table 9-2. The final hierarchical linear mixed model for explaining the change in learning performance

Variable	β	SE	95% CI		g
			Lower	Upper	
Phase (1)	2.401*	0.016	1.478	3.324	
Phase (2)	2.445*	0.010	1.792	3.089	
Phase (3)	2.413*	0.012	1.496	3.330	
Time to answer correctly (TTAC)	0.023*	0.001	0.004	0.041	0.27
Time to answer wrongly (TTAW)	-0.025*	0.002	-0.038	-0.012	0.26
Effort (RTE)	0.190	0.013	-0.411	0.791	0.17
TTAC*Phase (1)	0.022*	0.002	0.018	0.036	0.25
TTAC*Phase (2)	0.024*	0.002	0.014	0.044	0.26
TTAC*Phase (3)	0.028*	0.001	0.017	0.039	0.27
TTAW*Phase (1)	-0.018*	0.002	-0.022	-0.014	0.24
TTAW*Phase (2)	-0.013*	0.001	-0.018	-0.008	0.25
TTAW*Phase (3)	-0.017*	0.002	-0.021	-0.012	0.24

* $p < 0.05$

β : mean for the factor variable, **SE**: standard error, **CI**: confidence intervals, **g**: Hedges' g effect size. Ranges for Hedge's g effect size are small > 0.2, medium > 0.5 and large > 0.8.

Given the statistically significant effects of the interactions between phase identifier and time to answer correctly, as well as between phase identifier and time to answer wrongly, additional analyses were performed to shed light to the nature of these interactions. Table 9-3 demonstrates these results.

Table 9-3. Statistical differences between the phases of self-assessment with respect to fixed effects of the response-times variables on learning performance

Fixed Effect	Between Phases	df	F	g
Time to answer correctly (TTAC)	Phase (1) vs. Phase (2)	66	5.023	0.18
	Phase (2) vs. Phase (3)	66	1.024	0.07
	Phase (1) vs. Phase (3)	66	7.333	0.22
Time to answer wrongly (TTAW)	Phase (1) vs. Phase (2)	66	4.551	0.16
	Phase (2) vs. Phase (3)	66	1.127	0.08
	Phase (1) vs. Phase (3)	66	5.386	0.19

Bold represents statistically significant values.

As seen in this table, statistically significant differences in the effects of time to answer correctly and time to answer wrongly on performance are detected between the first and second phase, as well as between the first and third phase, while, the difference in the effects of these variables on learning performance between the second and third phase are only moderate.

9.5. Discussion

9.5.1. Interpretation of the results

The overall results of the study, as demonstrated in Table 9-2, revealed a statistically significant effect of the distinct response-times to answer correctly/wrongly on learning performance, in a setting of a longitudinal study, where the subjects serve as their own control. Both response-time factors contribute the most to explaining the variance in learning performance in all stages of the study. These findings provide additional proof to previous claims that the distinctive response-times are strong determinants of learning performance in self-assessment tests (Papamitsiou et al., 2014, 2016, 2018; Shih et al., 2008; Wang & Hanson, 2005). However, the previously reported strong positive effect of effort on learning performance was not further confirmed (Papamitsiou & Economides, 2015; Papamitsiou et al., 2016; Setzer, Wise, van den Heuvel, & Ling, 2013; Silm, Must, & Täht, 2013).

In addition, Table 9-3 shows direct comparisons between the three phases of the study with respect to the response-times and their effect magnitudes. As shown in this Table, between the first measurements of analytics parameters, prior to exposing learners to the treatment (i.e., the task-related analytics visualizations), and the second measurement, when the metacognitive help was available, the difference in the response-times was statistically significant and the effect of the difference on explaining the difference on learning performance was statistically medium ($F(1, 66)=5.023, g=0.18$ for TTAC; $F(1, 66)=4.551, g=0.16$ for TTAW). Similarly, the difference in response-times explains satisfactorily the difference in performance between the first and the third phase of the repeated measurements ($F(1, 66)=7.333, g=0.22$ for TTAC; $F(1, 66)=5.386, g=0.19$ for TTAW), whilst the effects of these differences are statistically small between the second and third phases ($F(1, 66)=1.024, g=0.07$ for TTAC; $F(1, 66)=1.127, g=0.08$ for TTAW). What

mediated and caused these differences was the usage of the analytics visualizations. Thus, what these findings imply is that the intervention employed, i.e., the usage of the available metacognitive help, strongly contributes in increasing learners' on-task engagement, which in turn, results in improved performance. Furthermore, this finding contributes to hypothesizing that students who consciously seek for help are more possible to be more responsibly involved with the learning tasks and they are more likely to be more careful about their answers.

These findings are in agreement with previous results that reported on the learners' ability to read and understand analytics visualizations as metacognitive feedback (Corrin & de Barba, 2015) and to make use of the metacognitive information (Daley et al., 2016) to improve their performance. Considering learners as the main recipients of learning analytics data might put in question how efficiently the learners could make-sense from this information (MacNeill, Campbell, & Hawksey, 2014). Despite this unease, previous studies argued that the learners can interpret their own performance indices, yet they reserve a scepticism on how to practically convert this information into action (Corrin & de Barba, 2015). Taking us a step further from visualizing the learners' own interaction trace data that the dashboards typically do (Rodríguez-Triana et al., 2014), the core innovation of this work derives from exploiting easy-to-read task-related analytics visualizations in order to provide to the learners instrumental information about the tasks, and investigates *how* their on-task engagement changes due to this intervention.

9.5.2. Implications for research and practice

Individuals' behavior usually changes in essential ways over a period of time. Prior studies followed cross-sectional research designs and detected a significant effect of metacognitive support on the learning gain. The core contributions of this study were threefold:

- Methodologically, it was one of the very limited in number studies in the field of learning analytics that implemented a longitudinal research design. This study showcased how the time metric for describing change can be coded to facilitate the research design. Time of measurements is frequently a predictor variable in longitudinal research. As shown from the findings, this factor was included in the final fixed effect model, and it indeed was one of the strong determinants of the change in learning performance. As such, this study provided the description of a coherent longitudinal study in the area of learning analytics and showcased how former hypotheses can be further explored and validated with respect to the changes in learners' behavior over time (*methodological implication*).
- This study provides further insight and evidence into existing body of research on the role of metacognitive help to increase learning gains. The theoretical model suggested in this study considers learners' engagement and investigates changes in learners' behavior and performance due to the metacognitive help. From the findings became apparent that the metacognitive help seeking caused significant changes in learners' behavior in terms of response-times, which in turn resulted in changes in performance. Investigating the effects of

changes in the usage of the on-demand task-related visualizations on the changes in performance is necessary to be clarified, as well (*implication for research*). Designing and implementing longer longitudinal studies, with more phases of exposing the sample to the treatments (i.e., more points in time) would facilitate that objective. In addition, providing alternative forms of assistance (e.g., executive help formats, explicit hints, etc.), measuring the effects of the changes in response-times and effort, and comparing these changes to the ones estimated in this study is expected to shed light to the effect size of the employed intervention.

- Combining the findings of this study with previous results that indicated an alignment of using the metacognitive help with perceiving them as useful and helpful (Chapter 8), further justified the role and significance of the intervention. As such, it provided a strong indication that training learners to use, read, and make-sense from the learning analytics fosters their metacognition and assists them to ask for assistance at the moment they actually need it (*practical implication*). In a sense, the findings provided empirical evidence on the added-value of enhancing the learning environments with interaction features that facilitate the learners' self-directed decisions. Accordingly, further training the learners on how to efficiently use such features is expected to build their capacity for autonomous learning (*practical implication*).

9.5.3. Limitations

First and foremost, a basic limitation of this study is that it assumed that the self-assessment procedures were of similar difficulty for the learners, despite the fact that the content of the self-assessment tests differed from measurement to measurement. In a sense, we treated the self-assessment procedures as “black boxes”. Considering the same number of items of similar levels of difficulty do not establish that the procedures are identical. The effects of the content itself on explaining the variance in on-task engagement and performance should be considered, as well.

One other limitation of the present study is the size of the sample which is marginal and the number of points in time selected is minimum as well. Experimentation with bigger sample sizes should be conducted and further longitudinal research following the same students over different self-assessment procedures, enhanced with metacognitive help, is needed in order to understand how responsible and effective learners become when using the task-related analytics visualizations.

Another limitation is that in this study only response-times and effort were considered in the hierarchical model. Other factors that have been strongly associated to the learners' performance (e.g., motivational constructs or affective states) should be explored as well.

9.6. Conclusions

The benefits of help-seeking for the overall learning gains are beyond question, and the role of help-seeking in knowledge acquisition is catalytic. However, help-seeking is an inherently complex mechanism, instigated by learners' intrinsic motivational criteria and externalized as an adaptive behavior (i.e., that evolves over time and is not stable), and it is an often phenomenon that the learners either underuse or overuse the available help-seeking facilities within the digital learning environments.

Former studies followed cross-sectional research designs to investigate the effects of help-seeking on learning performance. However, in order a theory to be consolidated, multiple measurements are required. The present longitudinal study explored the changes in learners' on-task engagement and performance that were caused by the changes in the usage of metacognitive help. The results provided strong evidence that this help format contributes to increasing the learners' attention when completing a learning task and to improving their learning outcomes. Additional research is required on the role and effect of the effort factor, as well as on exploring other significant learning/learner factors that have been previously found to affect performance. The most interesting finding of this study, though, is that the learners were not "afraid" to use the analytics visualizations and make-sense out of it, resulting in increased response-times and better self-assessment outcome, over time.

Chapter 10 : Taking control of the self-assessment

*"All stable processes we shall predict.
All unstable processes we shall control"*

John von Neumann

Exploring autonomous learning capacity from a self-regulated learning perspective using learning analytics

10.1. Introduction

Contemporary learning theories highlight the significant role of self-regulation on the learners' personal development. Self-regulated learning (SRL) refers to a "self-directive process by which learners transform their mental abilities into academic skills" (Zimmerman, 2002, p. 65). It is also conceptualized as "an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment" (Pintrich, 2000, p. 435). This process involves the systematic planning, regulation and evaluation of learning goals and their attainment (Narciss, Proske, & Koerndle, 2007). Self-regulated learners are aware of their learning processes and strategies, and adjust their behaviors to the specifications of the learning environments (McCardle & Hadwin, 2015).

And, the more the learning turns online, the higher the need for learners to develop and sustain self-regulation skills. The reason why SRL is critical in online learning environments, is that these environments are more autonomous than traditional classrooms or blended learning environments (Broadbent & Poon, 2015; Xu & Jaggars, 2014). Through the "lenses" of autonomy, the learners understand their needs, are aware of their self-directed learning goals, take control of and become responsible for their learning choices, monitor their progress, and critically reflect on their learning (Benson, 2001; Dickinson, 1995; Littlewood, 1996). In Benson's (2001) conceptualization of learner autonomy, the autonomous and self-directed learners take control over the cognitive, emotional, motivational, and behavioral processes, and the independent use of learning material/technology, i.e., the autonomous learners practice SRL strategies. A broader conceptualization of autonomy focuses on learners' capacity that allow them to accept responsibility and take control of their own learning processes (Vanijdee, 2003).

However, learners do not intuitively know how to achieve autonomy; they need to be trained in building the competences and capacity for efficient autonomous choices (McDevitt, 1997; White, 1995). For Oxford (2008), the use of learning strategies can promote learner autonomy. Andrade (2014) prompted that towards achieving autonomy, there is a need to develop technology enhanced learning environments within which the learners shall be given the opportunity for practising their SRL processes; they need to be consciously involved in their own learning, as they move along the SRL continuum towards autonomy.

Accordingly, we investigated how specific self-regulated learning strategies can predict and explain the learners' autonomous interactions, from a learning analytics perspective, in an online learning setting that allows for exercising autonomous control.

10.2. Autonomous and self-regulated learning

The concepts of autonomous and self-regulated learning share some similarities, and the relationship between them seems to be bidirectional. Loyens et al. (2008) elaborated on the conceptual clarity between autonomous and self-regulated learning. Both concepts involve the learners' active engagement and goal-directed behavior, and whilst effective SRL focuses on helping learners become autonomous thinkers, the autonomous learners appear to exercise control over their SRL strategies. This is a reason why both terms are usually used in literature interchangeably (Andrade & Bunker, 2009; Lewis & Vialleton, 2011).

However, the degree of control the learners have over the learning process significantly differs (Loyens et al., 2008). Learners exercise autonomy when they make choices and act on them: freedom of choice is central to the idea of autonomy – it is always the learners who choose what, where, and how to learn. Autonomy targets at fostering learners' responsible self-initiative and allows them to determine the selection of what shall be learned, as well as the critical evaluation (reflection) of the learning tasks that were selected (Candy, 1991). On the other hand, the concept of SRL places less emphasis on choices and more on guiding learners towards efficiently employing strategies. SRL seems more concerned with the subsequent steps in the learning process, such as setting goals, monitoring their progress, reflecting on the steps that were taken and changing their plans accordingly, and is usually described as a favorable learner characteristic, i.e., as the processes that the learner substantiates (Zimmerman, 2000). For example, selecting which learning materials to study (according to the learners' self-defined goals and priorities) corresponds to autonomous learning behavior, whereas deciding on the time to allocate on studying the learning materials (which are very possible to have been determined by the instructor or the curriculum) corresponds to an SRL strategy employed.

In line with Oxford (2008) and Andrade (2014), some studies have endorsed the importance of SRL strategies to students' autonomous learning in online contexts (Barnard et al., 2009), yet, little research has explicitly investigated the effects of SRL on learners' autonomy in online settings. Due to the high similarity of the two concepts, empirical research on how the learning strategies affect the development of capacity for autonomous learning is rather sparse. For example, Kormos and Csizér (2014) examined the mediation effect of self-regulation on autonomy in computer-assisted learning conditions and found that time-management predicted learners' perception of autonomous use of learning resources, but a strong motivation was a prerequisite for the adoption of the strategy. Sierens et al. (2009) and Schuitema et al. (2016) found that the regulation of cognition (a metacognitive strategy) was positively related to students' perceptions of autonomy support. In another study, Vansteenkiste et al. (2012) found that perceived autonomy support was positively related to learners' deep-level learning strategies and effort-regulation during the learning process.

10.2.1. Motivation of the research and research questions

Although the studies on the topic are limited, they converge to the conclusion that SRL strategies are expected to positively influence the learners' autonomous behaviour. This is not surprising due to the high similarity of the concepts. However, previous research is based on learners' perceptions of autonomy, resulting to scepticism about the external validity of the findings due to the gap that often exists between intentions and real behaviors (Gollwitzer et al., 2009). Moreover, these studies are not oriented to online learning environments that are inherently autonomous, and they do not report findings about SRL strategies that are commonly used in online learning setting (e.g., help-seeking, time-management, effort-regulation) (Barnard et al., 2009). Besides, there is a lack of knowledge on how SRL strategies are related to the actual learners' exercise of autonomous control. It is very likely that self-regulation might affect the actual self-enforced choices that learners exhibit in online conditions, in terms of interactions measured with learning analytics parameters: the question is how. And, it is very possible that SRL strategies might explain and justify the extent to which students take advantage of autonomous control in contexts or conditions that it is available, towards becoming autonomous learners: the question is which are dominant strategies. Additional empirical evidence is required towards better understanding the role of the SRL strategies in the development of autonomous learning capacity. Therefore, the research question that guided this study is defined as follows:

***RQ:** Which is the effect of SRL strategies on the learners' control of autonomous learning?*

10.3. Research model and hypotheses

In a contextualization of SRL to mostly apply in online learning environments, Broadbent (2017) and Barnard et al. (2009) suggested six prevalent strategies, i.e., goal-setting, time-management/study-environment, help-seeking, task-strategies (e.g., effort regulation, rehearsal), peer-learning, and self-evaluation. In the present study, peer-learning and self-evaluation do not apply because the online learning environment is self-assessment oriented.

The research hypotheses on the causal relationships between the considered factors are outlined as follows.

10.3.1. Effort-regulation

Effort regulation is the control of "one's effort expenditure" (Halisch & Heckhausen, 1977, p. 724). It incorporates the learners' ability to exert effort and to persist in their engagement with the learning items. We believe that students who regulate their efforts are expected to exhibit higher autonomous control, as well. That is because the sense of acting freely and making independent choices might be seen as an opportunity for students who properly regulate their efforts according to the requirements of the tasks. Thus:

***H1:** Effort-regulation will have a positive effect on Autonomous Control.*

10.3.2. Goal-setting

Goal-setting is a skill associated with the learners' self-set learning expectations, whereas self-set goals produce higher goal commitment (Zimmerman, 2000). We believe that in autonomous conditions, the learners who are aware of their goals and have high learning and achievement expectations, will take advantage of the autonomous control and will try to select learning items that will facilitate their goals. In other words, learners with higher goal-settings will go for harder learning items, whereas less motivated learners are less likely to exploit the available autonomous learning opportunities. Therefore:

H2: Goal-setting will have a positive effect on Autonomous Control.

10.3.3. Help-seeking

Help-seeking is "a behavior performed by individuals who perceive themselves as needing assistance with a problem, whereby the intended outcome of this behavior is addressing the problem faced" (Heerde & Hemphill, 2018, p.2). It is the learner who initiates the communication loop: being consciously aware of the need for help, the learner defines the help-seeking goals, estimates the benefits/costs of (not) seeking help, selects the appropriate sources, and obtains and processes help (Karabenick & Berger, 2013; Nelson-Le Gall, 1985). Based on Nelson-Le Gall's (1985) arguments that seeking instrumental help (i.e., aided in understanding) can result in greater autonomy, we believe that in autonomous learning conditions, the learners who have strong help-seeking skills will feel free to ask for multiple levels of hints during dealing with learning items. Therefore:

H3: Help-seeking will have a positive effect on Autonomous Control.

10.3.4. Time-management

Time-management is a meta-skill aiming at efficiently allocating time on each of the tasks, within a limited amount of time (Michinov, Brunot, Bohec, Juhel, & Delaval, 2011). It is associated with students' exercising conscious control over the amount of time-spent on learning items during the learning process. Students who believe that have good time-management skills are expected to take advantage of the autonomous control in order to adjust and fine-tune their time-management practices. In other words, these students are more likely to select items that they believe that best fit their knowledge-level and goal-orientation, they will allocate time on the items they have chosen, and they will make their decision strategies quickly. On the contrary, students who don't have high time-management skills are more likely to feel confused within the autonomous activity. These students are more possible to lose time on decision making, as well as to straggle to answer random questions, while they are expected to not exhibit autonomous control. Thus:

H4: Time-management will have a positive effect on Autonomous Control.

Figure 10-1 illustrates the causal relationships among the considered factors.

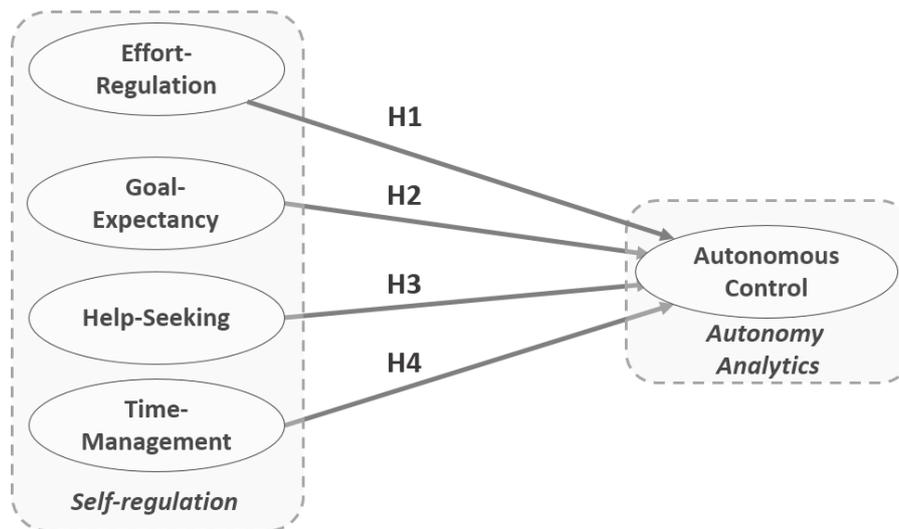


Figure 10-1. Overall research model and factor relationships with hypotheses

10.4. Methodology

10.4.1. Study design and research participants

The study followed an exploratory design. Overall, one hundred and thirteen (113) undergraduate students (51 females [45.1%] and 62 males [54.9%], aged 19-26 years-old (M=20.74, SD=1.755, N=113)) at a European University were enrolled in an online self-assessment procedure for the Management Information Systems I course (related to Information Systems, databases, and Business Intelligence), at the University lab, for 60 mins.

For the needs of the self-assessment, 150 multiple-choice questions (items) in total were calibrated before being available and exposed to the students. Each item had four possible answers, but only one was the correct. For the calibration of the item-bank, prior self-assessment tests were used. The calibration of the item-bank included the determination of the discrimination ability of the items, and the clarification of their difficulty. The discrimination ability of an item corresponds to the probability of students in each mastery class responding correctly to each item, i.e., how likely is a student of the given class to answer correctly to this item. In previous tests, three mastery classes of students were identified: High-performing: final grade \geq 7, Medium performing: final grade \geq 4, and Low performing: final grade $<$ 4. Moreover, two experts agreed on the items' difficulty (easy, medium, hard).

The students had to answer to up-to-12 items, and they had full-autonomy to select the next self-assessment item according to the desired level of difficulty of that item: the students could ask for an item of the same difficulty with the current one, they could ask for an easier or harder item, or they could ask for a random item (delivered to them according to the discrimination ability of the item and the students' currently diagnosed mastery level). For the score computation, only the correct answers were considered, without penalizing the incorrect answers. Each item's participation on the score was according to its difficulty, varying from 0.5 points (easy) to 1 point (medium) to 1.5 points (hard).

Before taking the self-assessment, each participant had to answer to a pre-test questionnaire that measures the SRL strategies, i.e., their perceptions of effort-regulation, goal-setting, help-seeking, and time-management. The participation to the procedure was optional; it was offered to facilitate the students' self-preparation before the final exams. All participants signed an informed consent form prior to their participation. The informed consent explained to them the procedure and was giving the right to researchers to use the data collected for research purposes. Students were aware that their interactions had been anonymized prior to being tracked and analyzed, and that the collected data would be stored for 3 years.

10.4.2. Data collection and measurements

Data were collected with an online self-assessment environment (Appendix A). Measures commonly used in the field of learning analytics (e.g., response-times, frequencies), indicative of the learners' autonomous control (AC) interactions, were computed from the logged trace data. Table 10-1 illustrates the measures captured and coded.

Time-spent on Decision Making (TTDM) is the total time that the students spend from the moment that they answer to an item until they make a decision on what item they want to answer next, and they ask for it. Moreover, Frequency of choosing easier (FEAS), Frequency of choosing harder (FHAR), Frequency of choosing same (FSAM), and Frequency of choosing random (FRAN) are simple counters for the respective choices, and they increase every time that the students make the respective selection of next item.

Table 10-1. Measurements used in the study.

Variable	Name	Description
FEAS	Frequency of choosing easier	<i>How many times the student asks for easier questions (compared to the current one)</i>
FHAR	Frequency of choosing harder	<i>How many times the student asks for harder questions (compared to the current one)</i>
FSAM	Frequency of choosing same	<i>How many times the student asks for question of the same difficulty with the current one</i>
FRAN	Frequency of choosing random	<i>How many times the student asks for a random question</i>
TTDM	Time-spent on decision making	<i>The time interval between answering a question and choosing the next one</i>

In order to develop the instrument for measuring the four SRL strategies, we adapted some items of the constructs from previously validated instruments. For students' effort-regulation (ER) we adopted three items from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich et al., 1993). For goal-setting we used the goal-expectancy construct, and configured three items from Terzis & Economides (2011). Goal-expectancy (GE) is a measure of

goal-setting particularized on assessment procedures. It reflects the students' achievement expectations in the self-assessment, expressed in terms of how satisfied they are with their preparation, and their desirable level of success. For help-seeking (HS), three items were adopted from the Online Self-Regulated Learning Questionnaire (OSLQ; Barnard et al., 2009). Finally, for time-management (TM) we configured three items from the OSLQ, as well. The questionnaire was first developed in English and then translated into the native language of the students. The translation was made by certified translators to ensure linguistic equivalence. All items were measured in a 7 point Likert-like scale (1 = strongly disagree to 7 = strongly agree, Appendix C).

10.4.3. Data analysis

10.4.3.1. Structural and measurement model

For addressing the research question, the construction of a path diagram that contains the structural and measurement model was conducted with the Partial least-squares (PLS) analysis technique (Chin, 1998; Tenenhaus et al., 2005). Our sample of 113 participants exceeds the recommended value of 50 i.e., a) 10 times larger than the number of items for the most complex construct (AC with five items), and b) 10 times the largest number of independent variables impact a dependent variable (ER, GE, HS, TM to AC).

10.4.3.2. Measures and Evaluation Criteria

The structural model evaluates the relationship between exogenous and endogenous latent variables by examining the variance measured (R^2). R^2 values of 0.67, 0.33, and 0.19 are substantial, moderate, and weak, respectively (Chin, 1998). The quality of path model can be evaluated by the Stone-Geisser's Q^2 value (Geisser, 1974; Stone, 1974), an evaluation criterion for the cross-validated predictive relevance of the PLS path model. The Q^2 statistic measures the predictive relevance of the model by reproducing the observed values by the model itself. A Q^2 greater than 0 means the model has predictive relevance; Q^2 statistic less than 0 mean that the model lacks predictive relevance. Finally, a bootstrap procedure evaluates the significance of the path coefficients (β value) and total effects, by calculating t-values. For the measurement and the structural model we used SmartPLS 3.2.

10.5. Results

10.5.1. Convergent validity - Discriminant validity

The results support the measurement model (Table 10-2). All criteria for convergent validity are met: the items' factor loadings on the corresponded constructs, Cronbach's α and composite reliability of all constructs are higher than 0.7 (Chin, 1998), and confirm reliability of the measurement model, and Average Variance Extracted is higher than 0.5, exceeding the variance due to measurement error for that construct.

Discriminant validity is also confirmed since the AVE of each construct is higher than the construct's highest squared correlation with any other construct. Table 10-3 presents the variables' correlation matrix; the diagonal elements are the square root of the AVE of a construct.

Table 10-2. Results for the Latent Constructs of the Measurement Model

Construct	Factor Loadings	Cronbach's a	Composite	Average Variance
Items	(>0.7)^a	(>0.7)^a	Reliability (>0.7)^a	Extracted (>0.5)^a
ER		0.777	0.857	0.602
ER1	0.818			
ER2	0.853			
ER3	0.823			
GE		0.871	0.920	0.794
GE1	0.885			
GE2	0.893			
GE3	0.895			
HS		0.797	0.871	0.693
HS1	0.906			
HS2	0.857			
HS3	0.725			
TM		0.853	0.911	0.773
TM1	0.908			
TM2	0.851			
TM3	0.877			
AC		0.791	0.864	0.625
FEAS	0.803			
FHAR	0.871			
FSAM	0.702			
FRAN	0.709			
TTDM	0.716			

^a Indicates an acceptable level of reliability and validity.

ER: Effort-regulation, **GE:** Goal-expectancy, **HS:** Help-seeking, **TM:** Time-management, **AC:** Autonomous Control, **FEAS:** Frequency of choosing easier, **FHAR:** Frequency of choosing harder, **FSAM:** Frequency of choosing same, **FRAN:** Frequency of choosing random, **TTDM:** Time-spent on decision making

Table 10-3. Measurement Model (Discriminant validity)

	1	2	3	4	5
1. Effort-Regulation	0.831				
2. Goal-Expectancy	0.604	0.891			
3. Help-Seeking	0.334	0.595	0.833		
4. Time-Management	0.616	0.708	0.480	0.879	
5. Autonomous Control	0.422	0.501	0.187	0.488	0.776

10.5.2. Testing hypotheses

A bootstrap procedure with 3000 resamples was used to test the statistical significance (t-value) of the path coefficients (β value) in the model. Table 10-4 summarizes the results for the hypotheses testing.

Table 10-4. Hypothesis testing results

Hypothesis	Path	β	t	P	Result
H1	Effort-regulation \rightarrow Autonomous Control	0.105	1.076	0.283	Not Support
H2	Goal-expectancy \rightarrow Autonomous Control	0.374	3.388*	0.001	Support
H3	Help-seeking \rightarrow Autonomous Control	-0.190	2.154*	0.032	Support Opposite
H4	Time-management \rightarrow Autonomous Control	0.250	2.244*	0.025	Support

* $p < 0.05$

As seen from this table, two of the initial hypotheses are supported, one is not supported and for one hypothesis, its negation is supported. These results are further discussed in next section.

10.5.3. Overall Model Fit

The suggested model explains almost the 33% of the variance in autonomous control, which is statistically moderate. The cross-validated predictive relevance of the model was confirmed ($Q^2=0.261$). Table 10-5 synthesizes the total effects of the selected factors, as well as the variance (R^2) and cross-validated predictive relevance (Q^2) explained by the proposed model.

Table 10-5. R^2 , Q^2 and Total effects

<i>Endogenous</i>	R^2	Q^2	<i>Exogenous</i>	<i>Total effect</i>	<i>t</i>
Autonomous Control	0.332	0.261	Effort-regulation	0.105	1.076
			Goal-expectancy	0.374	3.388*
			Help-seeking	-0.190	2.154*
			Time-management	0.250	2.244*

* $p<0.05$

The measurement results are summarized in Figure 10-2. This figure illustrates the path coefficients for the initial hypotheses of the research model.

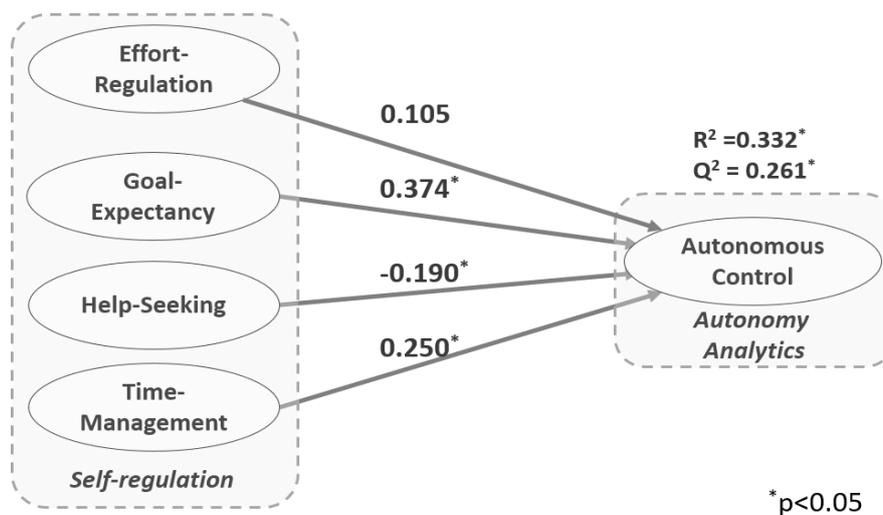


Figure 10-2. Path coefficients of the research model, overall variance explained (R^2) for test score and cross-validated predictive relevance (Q^2).

10.6. Discussion

10.6.1. Findings and interpretations

Online learning environments are more autonomous than traditional classrooms or blended learning environments (Broadbent & Poon, 2015; Xu & Jaggars, 2014). The more the learning turns online, the higher the need for learners to develop and sustain autonomous learning competences. However, learners do not intuitively know how to achieve autonomy. Towards achieving autonomy, exercising SRL strategies has been acknowledged as critical

(Oxford, 2008). The question is *which* SLR strategies are more effective to employ towards building capacity for autonomous learning, and *how* these strategies influence the learners' actual autonomous interactions, within inherently autonomous online learning environments.

The research on the topic is rather limited and previous studies relied on learners' perceptions of autonomy, and were not conducted in online learning contexts. The present study targets at bridging this gap and at providing empirical evidence on which SLR strategies better predict and justify the learners' self-enforced choices. The current study takes us a step ahead by considering the learners' autonomous choices, using a learning analytics perspective.

In particular, in the demonstrated approach, autonomy was modeled in terms of frequencies of interaction types and time-spent on decision making, both commonly used in the learning analytics research practise (Kovanović et al., 2015; Papamitsiou & Economides, 2017). The consistency reliability of the factor that measures autonomy was confirmed ($\alpha=0.791$). Next, a set of specific SRL strategies that have been identified in previous research as critical in online learning contexts (Barnard et al., 2009; Broadbent & Poon, 2015), and that could be associated to or infer autonomous behaviour, were explored through a structural and measurement model. These factors were measured with configured versions of previously validated instruments, and their consistency reliability was also confirmed (Table 10-2).

The overall prediction accuracy of the model proposed in this study was 33.2%, which is moderate, and the cross-validated predictive relevance was confirmed ($Q^2=26.1\%$) (Table 10-5). Although this seems an *intuitive* finding, however, one would expect a higher prediction accuracy, due to the high similarity between the concepts of autonomy and self-regulation, acknowledged in literature. Further elaborating on that, on the one hand, this finding contributes to understanding the diversity of the two concepts and provides empirical evidence that justifies their discrimination. On the other hand, it becomes apparent that additional factors should be considered as well. For example, Kormos and Csizér (2014) argued that this relationship is catalysed by the learners motivation to adopt an SRL strategy.

Moreover, the analysis revealed a strong direct positive effect of goal-setting and time-management on autonomous control ($\beta=0.374, t=3.388, p=0.001; \beta=0.250, t=2.244, p=0.025$ for GE and TM respectively), confirming hypotheses **H2** and **H4** (Table 10-4). Aligned with and further confirming previous results (Kormos & Csizér, 2014), this finding indicates the strong role of goal-orientation in both autonomous and self-regulated learning (Vanijdee, 2003). Practically, it extends previous findings by adding empirical evidence, and it implies two posits: (a) learners who are aware of their goals and have high learning and achievement expectations, will try to select items that will facilitate their goals, and (b) learners who believe that they have good time-management skills, will take the opportunity for autonomous control in order to adjust and fine-tune these skills. In a sense, learners want to be consistent with their choices and to take responsibility for their learning (Oxford, 2015).

In addition, a moderate positive effect of effort-regulation on autonomy ($\beta=0.105$, $t=1.076$, $p=0.283$) was detected, neither supporting nor rejecting hypothesis **H1** (Table 10-4). In previous research (Vansteenkiste et al., 2012) effort-regulation was found to be a strong determinant of perceived autonomy. The slight divergence of our finding might be due to *how autonomy was measured*, highlighting the difference between intentions and real behaviors (Gollwitzer et al., 2009). Based on our results, the role of this factor is not clear: although it seems that learners' perceptions of persistence in their engagement with the learning tasks might influence their self-enforced choice of tasks, however, in practise, effort expenditure on the selected tasks would possibly clarify more this relationship.

The most intriguing finding of this study concerns the strong direct negative effect of the help-seeking factor on autonomous interactions ($\beta=-0.190$, $t=2.154$, $p=0.032$), resulting in supporting the opposite of hypothesis **H3** (Table 10-4). Due to the lack of previous research on this relationship, due to the highly self-initiating nature of help-seeking behavior, and in line with Nelson-Le Gall's (1985) claims that help-seeking can promote autonomy, we assumed that help-seeking would positively influence autonomous control. We hypothesized that, in online learning environments, the learners who have strong help-seeking skills would feel free to ask for multiple levels of hints during dealing with learning items. Online learning environments provide increased opportunities for free help-seeking and they can preserve learners' anonymity. Surprisingly, not only this hypothesis was not confirmed, yet, the opposite results was discovered. Based on our results, help-seeking contradicts autonomous control. A possible explanation of this result would be the following: learners might perceive help-seeking as a threat to their autonomy (Huet et al., 2011). They might be negatively positioned to seek help because of the fear of feeling "dumb", the internal insecurity of losing learning autonomy, or the doubt of self-worth by being exposed to external assistance and feeling incompetent to complete tasks. This finding requires attention and should be further explored from different perspectives, including actual measurements of help-seeking behavior in environments that facilitate autonomous learning conditions.

10.6.2. Implications for research and practice

The findings offer implications for research and practice. Understanding how learners' self-regulation contributes to and influences their autonomy can provide insight on how to plan the learners' SRL support in online learning environments (research). This means that, in the next step, practitioners shall be able to integrate specific features into online learning environment, in order to train learners on effectively using the SRL strategies, accordingly (practice). For example, supporting learners' effort-regulation (e.g., with learning analytics visualizations) might help the learners to persist on their on-task engagement and guide them to choose tasks that better correspond to their learning goals.

Moreover, attention is required regarding the impact of help-seeking on self-directed learning. The divergence of our finding from previous theoretical claims indicates that the relationship between both self-initiated behaviors (i.e., autonomous choices and help-seeking) should be further explored (research). For example, providing a help-seeking functionality to the learners, measuring their interactions with this facility, and exploring the next autonomous choices of learning tasks, would provide additional empirical evidence on the effect of help-seeking on autonomy. Clarifying this relationship shall next open new directions towards making decisions on how to efficiently guide the learners to feel free to seek help (with all the learning gains that help-seeking brings) (practice).

10.6.3. Limitations and Future work

Of course, there are limitations as well. First, the sample size (N=113) of the present study is relatively small, yet sufficient for the analysis methods employed. Larger samples should be explored to further validate the demonstrated findings. Second, two SRL strategies, i.e., peer-learning and self-evaluation did not apply on the context of this study, and need to be considered for analysis, as well. Last, individual differences among participants might have contributed to the difference in adoption of SLR strategies, and in autonomous choices.

10.7. Conclusions

Self-regulation and autonomy share common characteristics and they are usually treated as synonyms in literature. However, autonomy is a broader concept, strongly oriented to the learners' freedom of choice and degree of control over the learning tasks, whereas SRL seems more concerned with the subsequent steps and comprises sets of strategies that the learner can employ towards achieving autonomy. And, whilst autonomous learning can encompass SRL, the opposite does not hold. We suggest that understanding *how* and *which* SRL strategies mediate and facilitate the development of capacity for autonomous learning is critical nowadays that learning increasingly turns online.

Chapter 11 : Towards autonomous decision making

*“It’s ok to have your eggs in one basket
As long as you control what happens to that basket”*

Elon Musk

A learning analytics approach on the impact of learners’ autonomy on performance, response-time and effort

11.1. Introduction

Adaptivity in learning environments is in the epicentre of the technology enhanced learning research community. The focus of these systems is on providing the best possible support to the learners (Brusilovsky et al., 2016). The core idea is to increase their “awareness” regarding the learners’ cognitive and emotional states, as well as regarding their degree of non-cognitive skills and competences acquisition, and to accurately predict what kind of personalized assistance the learners would need, accordingly (Economides, 2009a). The support is usually delivered either in the form of personalized course content or visualizations (Essalmi, Ayed, Jemni, Kinshuk, & Graf, 2010; Vesin, Ivanović, Klašnja-Milićević, & Budimac, 2012), as recommendations (individualized /group) (Anaya, Luque, & Peinado, 2016; Chen, Li, Liu, & Ying, 2018; Papamitsiou & Economides, 2018), or by adjusting the learning design to match the learners’ requirements (Mavroudi, Giannakos, & Krogstie, 2018; Towle & Halm, 2005). In each case, it is a pre-requisite for these systems to deeper “learn” and “understand” the learners, in order to make informed decisions and to best scaffold them throughout the learning process (Brusilovsky et al., 2016).

However, although the adaptive learning systems tend to be enhanced with artificial intelligence and to “know” (to infer) what the learners need, *would the learners themselves make the same choices, if they had the opportunity to freely and responsibly choose?* What if a learning environment could provide its users with the opportunity to take charge of their own learning and to independently select what to study, with respect to their own learning objectives and learning gain? Would this sense of *learning autonomy* lead the learners to increased engagement and improved learning outcomes? And what would be the “*lessons learnt*” from this intervention for the intelligent, adaptive learning environments?

11.2. Theoretical background and related work

These questions are neither hypothetical, nor rhetorical. Extended theoretical research on the topic has determined the traits of autonomous learner: through the “lenses” of autonomy, the learners understand their needs, are aware of their self-directed learning goals, take control of and become responsible for their learning choices, monitor their progress, and critically reflect on their learning (Benson, 2001; Cotterall, 1995; Dickinson, 1995; Holec, 1981; Little, 1991; Littlewood, 1996; White, 1995). In Benson’s (2001) conceptualization of learner autonomy, the

autonomous and self-directed learners take control over the cognitive, emotional, motivational, and behavioral processes of learning, as well as the independent use of learning material and technology. This approach implies two facets: (a) the autonomous learners exhibit self-regulation strategies, and (b) they make decisions independently.

Aligned with the first facet, within Self-Determination Theory (SDT), autonomy is driven by the learners' personal interests and inner values, and concerns volition and self-regulation (Ryan & Deci, 2009). SDT outlines a continuum on which the learners experience autonomy differently, and regulate their learning processes accordingly. However, autonomy is beyond self-regulation only; autonomous learners are also capable of taking responsibility for their learning choices (Oxford, 2015), as well.

Regarding the second facet, autonomy in SDT has been mistakenly equated with independence: whether individuals are independent, dependent, or interdependent is not a defining characteristic for autonomy (Ryan & Deci, 2009). Little (1995) argues that autonomy is beyond independence; the learners should not only feel that they are independent, but to be guided to develop their capacity for autonomy, as well.

However, learners do not intuitively know how to achieve autonomy; they need to be trained in building the competences and capacity for efficient autonomous choices (McDevitt, 1997; White, 1995). Andrade (2014) prompted that towards achieving autonomy, there is a need to develop technology enhanced learning environments within which the learners shall be given the opportunity for exercising control over their self-regulated learning processes and to be consciously involved in their own learning.

11.2.1. Perceived measures of learner autonomy and previous results

For measuring the learners' perceived autonomy or to assess their autonomous motivation, different Instruments have been constructed in literature. For example, the questionnaire designed by Cotterall (1995) targets at investigating learners' perceptions of themselves as autonomous learners and at measuring their readiness for autonomy. Other popular instrument for autonomous/controlled motivation was adapted from the Learning Self-Regulation Questionnaire (SRQ-A - Ryan & Connell, 1989). The initial questionnaire was developed to measure students' regulatory style in 4 subscales: intrinsic motivation, identified regulation, introjected regulation and external regulation. Combined with the Relative Autonomy Index (RAI - Ryan & Connell, 1989), SRQ-A is used to compute the students' level of autonomy as the sum score for each subscale. Moreover, the Basic Psychological Need Satisfaction (BPNS) Questionnaire (Baard, Deci, & Ryan, 2004) and the Intrinsic Motivation Inventory (IMI) Questionnaire (McAuley, Duncan, & Tammen, 1989) were also adopted to measure learners' perceived autonomy (e.g., Nikou & Economides, 2017a; Sergis, Sampson, & Pelliccione, 2018). Other studies (e.g., Yen & Liu, 2009) used the short version of the Learner Autonomy Profile (LAP-

Confessore G., 2004) to measure learner autonomy, whereas the Academic Motivational Scale (AMS-Vallerand et al., 1992) was used to assess students' autonomous motivation.

Using these instruments, it has been found that autonomous learners achieve their learning outcomes with positive feelings towards the learning activity (Deci, Eghrari, Patrick, & Leone, 1994), reduced anxiety (Ghorbandordinejad & Ahmadabad, 2016), and increased intrinsic motivation (Guay, Boggiano, & Vallerand, 2001). Classroom climates that are supportive of students' need for autonomy, cultivate to students the feeling of autonomous motivation to learn and to achieve. When learners perceive themselves as more autonomous in classroom activities and they find their schoolwork meaningful, purposeful and useful in future learning, they become more interested and goal-orientated, and they tend to exhibit higher self-efficacy, engagement and responsibility in learning (Luke, 2006; Newby & Winterbottom, 2011). Moreover, feeling autonomously motivated in the classroom, the students are likely to engage in effort regulation and deep-processing (León, Núñez, & Liew, 2015) in turn, they progressively need less structure, increasing their capacity for autonomy (Andrade, 2014).

Studies on the effects of learners' autonomy from a SDT perspective (e.g., Hu & Zhang, 2017; Sergis et al., 2018), agree that internalizing extrinsic motivation, as well as fulfilling the learners' needs for autonomy, competence, and relatedness, facilitated them to move from dependence to autonomy, and improved their proficiency. It was also argued that students could invest more time on hands-on activities and collaboration instead of being exposed to restricting teacher-led lecturing.

Moreover, results suggest that students benefit more from self-directed assessment than from teacher-directed assessment in terms of learning outcomes (Fletcher & Shaw, 2012). Higher level of student autonomy in the assessment processes was positively related to the learning gain and outcome.

11.2.2. Objective measures of learner autonomy – A learning analytics approach

As seen from the literature review, although autonomy is considered as an important factor impacting positively the learning outcome, however, previous studies rely mostly on – the potentially biased – subjective learners' perceptions. These perceptions are difficult to be transformed into actionable interventions, because they don't reveal anything about the learners' autonomous, self-enforced choices (*what* they choose and *why*). At the same time, contemporary learning environments allow for gathering learner-generated interaction data; these data, coded as learning analytics, have been extensively used and acknowledged as significant factors that explain learners' performance, in different contexts.

Some researchers have investigated the role of learner autonomy in open learning environments such as MOOCs (Kop, Fournier, & Mak, 2011; Tschofen & Mackness, 2012), online distance courses (Giesbers et al., 2013; Hartnett, George, & Dron, 2011) or online collaborative learning settings (Rienties, Tempelaar, Giesbers, Segers, & Gijsselaers, 2014). These studies focus

on examining the level of learner autonomy in this type of courses. However, although they adopt an analytics viewpoint to measure other factors (e.g., participation) associated with autonomy, they rely on learners' perceptions to measure autonomy itself. Moreover, in an attempt to measure the actual learners' autonomy in a MOOC environment, Dawson et al. (2015) adopted typical variables from the learning analytics research, including the number of active users and the number of assignments completed. However, no clear conclusions were extracted due to the high dropout rate.

In addition, although Sergis et al. (2018) measured autonomy with data describing the time invested on the engagement of students in learning activities related to hands-on practice in the classroom, as captured by the practitioners' observations, beyond the questionnaire items, however, the authors didn't proceed to separately processing these data in a learning analytics fashion.

11.2.3. Motivation of the research and research questions

Although metrics extracted from the logged learners' interactions within a learning environment have been acknowledged for providing valuable insight into the patterns of student engagement (Papamitsiou & Economides, 2014a), however, these measures lack a strong theoretical background, regarding their ability to explain the learners' autonomy. This contribution aims at analysing logged interaction data for exploring the effect of autonomous control on learning performance, as well as to develop a conceptual model of this relationship.

Specifically, this study investigates the extent to which the utilized learning analytics (i.e., frequencies of decisions and choices, and time-spent on decision making) influence the learning outcome, measured as the score achieved in a self-assessment procedure. Moreover, in order to shed light to the learners' self-enforced choices, and consequently, to better understand the role of autonomous control in the learning experience, it is required to investigate the relationship of autonomous control with other objective factors that are strongly associated with learning performance. For example, when a learner makes a self-enforced decision and chooses a learning item, it is important to know whether the chosen item was randomly selected or if it was concisely requested. Yet, it is more likely that the learner will remain more engaged with this one, resulting in more time-spent and effort on understanding it and interacting with it. Learners' response-times and effort have been acknowledged for their capacity to reflect learners' engagement in the learning activity and to explain the learning outcome, accordingly. Therefore, the research questions are twofold:

***RQ1(a):** Which is the impact of autonomous control on learners' performance, response-times and effort? **(b):** Does the exploitation of autonomous control (measured with utilized analytics) contribute to enhancing the learners' performance? If yes, how significant is its effect on learning performance?*

RQ2(a): Are there any differences in the analytics parameters of autonomous control, with respect to the learners' level of performance? If yes, how significant is the effect of each one of these parameters? **(b):** Are there any differences in the other factors (i.e., the response-times and effort), with respect to the learners' level of performance? If yes, how significant is the effect of each one of these parameters?

11.3. Research model and hypotheses

By default and according to the previous studies, it is expected that:

H1: *Autonomous Control will have a positive effect on Learning Performance*

The rest of the research hypotheses on the causal relationships between the autonomous control, and response-times and effort are outlined as follows:

11.3.1. Response-times

Response-time is defined as the total time that the learners spend on interacting with a learning object. When the learners have to submit an answer to the specific learning object (i.e., the object is a question or a problem), then the response-times can be discriminated according to the correctness of the submitted answer. In this case, they indicate the respective response-time the learners constantly aggregate on answering the objects correctly or wrongly. When learners make self-enforced choices of learning objects, it is more likely that they will spend more time to answer on these objects correctly, and consequently, they will accumulate less time on objects that they will finally answer wrongly. Thus:

H2: *Autonomous control will have a positive effect on Total Time to Answer Correctly*

H3: *Autonomous control will have a negative effect on Total Time to Answer Wrongly*

11.3.2. Effort

Effort is "the motivational state commonly understood to mean trying hard or being involved in a task. Effort is increased when the subject tries harder, when there are incentives to perform well, or when the task is important or difficult" (Humphreys & Revelle, 1984). Thus, effort is about how much engaged the learners are in answering the items. We hypothesize that when the learners concisely request for specific items, it is more possible that they will remain engaged in these items, and they will demonstrate high effort exertion on dealing with them. Thus:

H4: *Autonomous control will have a positive effect on Response Time Effort*

Figure 11-1 illustrates the causal relationships among the factors considered. The dashed arrows represent the relationships between factors that have been previously explored and have been found to satisfactorily explain performance (Papamitsiou et al., 2016, 2014), whereas the rest arrows represent the research hypotheses explored in this study.

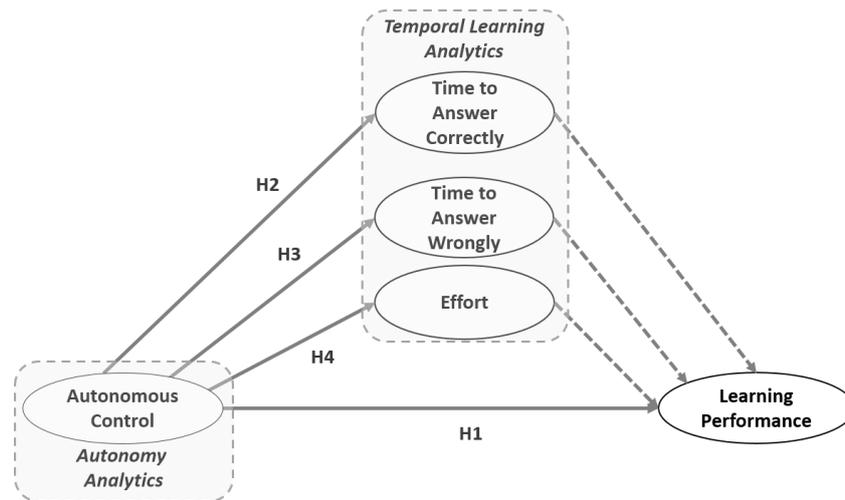


Figure 11-1. Overall research model and factor relationships with hypotheses

11.4. Methodology

11.4.1. Study design and research participants

This study followed an experimental design, i.e., experimentation in controlled conditions which involves applying some treatment to one of the groups, while holding other factors constant (Cobb et al., 2003). Overall, 168 undergraduate students (79 females [47.1%] and 89 males [52.9%], aged 19-26 years-old [M=20.67, SD=1.825, N=168]) at a European University were enrolled in an online self-assessment procedure for the Management Information Systems I course (related to Information Systems, databases, and Business Intelligence), at the University computers laboratory, for 60 minutes.

According to the experimental protocol, based on a pre-test questionnaire, all students had similar experience with computer-based assessment and similar computer efficacy. Next, they were randomly assigned into three groups, i.e., the two treatment groups and one control group. More precisely, 57 students (33.9% of the participants) were assigned to the “*fully-autonomous*” group, 56 students (33.3% of the participants) were assigned to the “*semi-autonomous*” (SA) group, and 55 students (32.8% of the participants) were assigned to the “*no-autonomous*” (NA) group (i.e., the control group). The FA group had *full-autonomy* to select the next self-assessment question, according to the desired level of difficulty of that item: the students could ask for a question of the same difficulty with the current one, they could ask for an easier or harder question, or they could ask for a random question. The SA group had *guided autonomy*: the system delivered a question in a Computerized Adaptive Testing (CAT) manner, according to the correctness of the previous answer and the discrimination ability of the question (Algorithm 1, Appendix B), but the students could either answer that question or ask for another one to replace it (which again was selected according to the correctness of the previous answer, based on the similarity ranking of the items). The NA group had *no autonomy*: all students had to answer on a fixed test consisting of a set of pre-determined questions of varying difficulty.

In all cases (FA, SA, NA), the students had to answer to up-to-12 questions. For the score computation, only the correct answers were considered, without penalizing the incorrect answers (i.e., without negative scores). Further, each question's participation on the score was according to its difficulty level, varying from 0.5 points (easy) to 1 point (medium) to 1.5 points (hard). If students in the NA group chose not to submit an answer to an item, they received zero points for this one.

For the needs of the self-assessment, 150 multiple-choice questions in total were calibrated before being available and exposed to the students. Each question had four possible answers, but only one was the correct. For the calibration of the question-bank, prior testing (involving students who had already been classified) was used. Three mastery classes of students were identified: Class A (advanced): final grade \geq 7, Class B (basic): final grade \geq 5, and Class C (below basic): final grade $<$ 5. The calibration of the question-bank included the determination of the discrimination ability of the questions, as well as the clarification of their level of difficulty. The discrimination ability of a question corresponds to the probability of students in each mastery class responding correctly to each question, i.e., how likely is a student of the given class to answer correctly to this question. Moreover, two experts agreed on the questions' level of difficulty (easy, medium, hard).

Table 11-1 synthesizes the experimental study (description of the sample and the self-assessment process).

Table 11-1. Synopsis of the experimental study.

Sample		Procedure
Group ID	Group size	(Self-assessment)
Full-Autonomous	57 (33.9%)	<i>Select the next self-assessment question according to the desired level of difficulty.</i>
Semi-Autonomous	56 (33.3%)	<i>Answer the question that the system delivers or ask for another one to replace it.</i>
No-Autonomous	55 (32.8%)	<i>Answer on a fixed test consisting of a set of pre-determined questions</i>

The participation in the procedure was optional; it was offered to facilitate the students' self-preparation before the final exams. All participants signed an informed consent form prior to their participation. The informed consent explained to them the procedure and was giving the right to researchers to use the data collected for research purposes. Students were aware that their interactions had been anonymized prior to being tracked and analyzed, and that the collected data would be stored for 3 years.

11.4.2. Data collection and measurements

Data were collected with an online self-assessment environment (Appendix A). For all three groups of students, measures commonly used in the field of learning analytics and acknowledged to satisfactorily explain students' performance (e.g., response-times, frequencies)

(Joksimović et al., 2015; Papamitsiou et al., 2018) were computed from the logged interactions trace data. In addition to these measures, for the FA and the SA groups, other similar measures, indicative of the students' autonomous control interactions, were computed, as well. Table 11-2 illustrates the measures captured and coded for each group.

Table 11-2. Measurements used in the study.

Variable	Name	Description	FA	SA	NA
TTAC	Total time to answer correctly	<i>The response-time a student aggregates on submitting correct answers</i>	✓	✓	✓
TTAW	Total time to answer wrongly	<i>The response-time a student aggregates on submitting the wrong answers</i>	✓	✓	✓
RTE	Effort	<i>When a student exhibits solution behavior – a measure of engagement</i>	✓	✓	✓
FEAS	Frequency of choosing easier	<i>How many times the student asks for easier questions (compared to the current one)</i>	✓		
FHAR	Frequency of choosing harder	<i>How many times the student asks for harder questions (compared to the current one)</i>	✓		
FSAM	Frequency of choosing same	<i>How many times the student asks for question of the same difficulty with the current one</i>	✓		
FRAN	Frequency of choosing random	<i>How many times the student asks for a random question</i>	✓		
TTDM	Time-spent on decision making	<i>The time interval between answering a question and choosing the next one</i>	✓	✓	
FACC	Frequency of acceptance	<i>How many times the student takes the question that the system delivers</i>		✓	
FREP	Frequency of replacement	<i>How many times the student asks for another question</i>		✓	

FA: Full-autonomous, SA: Semi-autonomous, NA: No-autonomous

As explained in sub-section 11.3.1, total time to answer correctly (TTAC) and total time to answer wrongly (TTAW) indicate the respective response-time the students constantly aggregate on answering the questions (Papamitsiou et al., 2014). Similarly, Total Time on Decision Making (TTDM) is the total time that the students spend from the moment that they answer a question until they make a decision on what item they want to answer next, and they ask for it. Moreover, Frequency of choosing easier (FEAS), Frequency of choosing harder (FHAR), Frequency of choosing same (FSAM), Frequency of choosing random (FRAN), Frequency of acceptance (FACC), Frequency of replacement (FREP) are simple counters for the respective choices, and they increase every time that the students make the respective selection of next item. For the effort calculation we used the Response Time Effort (RTE) measure. RTE measures the proportion of items which the students try to solve (solution behaviour – SB) instead of guessing the answers (Wise & Kong, 2005). The RTE for a student

j is: $RTE_j = \frac{\sum SB_{ij}}{k}$, where k is the number of items, and $SB_{ij} = \begin{cases} 1, & \text{if } RT_{ij} \geq T_i \\ 0, & \text{otherwise} \end{cases}$, where RT_{ij} is student's

j response time to item i , and T_i is a threshold value that discriminates solution behaviour from guessing.

The system also calculates the learning performance (LP) for each learner according to the correctness of the learner's answer on each question i , and the difficulty d_i of the question.

11.4.3. Data analysis

11.4.3.1. Structural and measurement model

For addressing RQ1(a), the construction of a path diagram that contains the structural and measurement model was conducted with the Partial least-squares (PLS) analysis technique (Chin, 1998; Tenenhaus et al., 2005). Moreover, PLS is suitable for studies that have small samples. In PLS the sample size has to be a) 10 times larger than the number of items for the most complex construct, and b) 10 times the largest number of independent variables impact a dependent variable (Chin, 1998). In this study, the most complex predictor is AC with five items, and the largest number of independent variables impacting a dependent variable is four (TTAC, TTAW, RTE, AC to LP). Thus, our sample is large enough (>50).

11.4.3.2. Measure and Evaluation Criteria

The structural model evaluates the relationship between exogenous and endogenous latent variables by examining the variance measured (R^2). R^2 values of 0.67, 0.33, and 0.19 are substantial, moderate, and weak, respectively (Chin, 1998). The quality of path model can be evaluated by the Stone-Geisser's Q^2 value (Geisser, 1974; Stone, 1974), an evaluation criterion for the cross-validated predictive relevance of the PLS path model. The Q^2 statistic measures the predictive relevance of the model by reproducing the observed values by the model itself. A Q^2 greater than 0 means the model has predictive relevance; Q^2 statistic less than 0 mean that the model lacks predictive relevance. Finally, a bootstrap procedure evaluates the significance of the path coefficients (β value) and total effects, by calculating t-values.

Since PLS path modeling is a distribution-free method (Chin & Dibbern, 2010), t-test is not a suitable method regarding the multigroup analysis (MGA), in order to investigate if the difference in path coefficients of different groups is statistically significant. Sarstedt, Henseler, & Ringle (2011) proposed a method (that does not follow distributional assumptions) for PLS-based MGA, deriving from bootstrapping in combination with a rank sum test, which makes it conceptually sound. In particular, after having exposed the sub-samples to separate bootstrap analyses and having made assumptions about the distributions of the parameter standard errors, one can calculate the statistic for the difference in paths between groups.

For the measurement and the structural model we used SmartPLS 3.2. The PLS-MGA method, as implemented in SmartPLS, is an extension of the original non-parametric Henseler's MGA method

(Sarstedt et al., 2011). We used this method in order to evaluate the differences between the path coefficients of the three groups of the study.

11.4.3.3. *Between groups analysis*

Regarding RQ1(b), independent samples t-tests were used to investigate the impact of the different scales of autonomous control (i.e., fully-autonomous, semi-autonomous, no-autonomous) on learning performance. Thus, the statistical significance of the differences with respect to the learners' achievement scores was estimated between the treatment groups and the control group of this study. However, not every significant result refers to an effect of high impact. As such, calculating the effect size is necessary for evaluating the strength of a treatment. Therefore, Hedge's g effect size was considered, because the sample size of each sub-group is considered small. Ranges for Hedge's g effect size are small > 0.2 , medium > 0.5 and large > 0.8 .

11.4.3.4. *Within groups (between subjects) analysis*

Regarding RQ2(a) and RQ2(b), the students in treatment and control groups were initially clustered into three mastery classes according to their performance (see section 12.4.1). Then, ANOVA tests were performed to investigate differences in each one of the analytics parameters of autonomous control (i.e., FEAS, FHAR, FSAM, FRAN, TTDM, and FACC, FRAN, TTDM), as well as in the other temporal learning analytics (i.e., TTAC, TTAW, RTE) between the different performance-based student clusters. The impact of these parameters was explored as well, and the η^2 effect size was computed for evaluating the strength of each one of these parameters. Ranges for η^2 effect size are small > 0.01 , medium > 0.06 and large > 0.14 . The decision to use ANOVA tests instead of multiple t-tests was based on the fact that ANOVA controls the Type I error so as it remains at 5%, when the number of groups is higher than two.

The between groups and the within groups analysis tasks were performed using the IBM "Statistical Package for the Social Sciences" (SPSS), version 20.0 for Windows.

11.5. Results

11.5.1. Convergent validity - Discriminant validity

The results support the measurement model. All criteria for convergent validity are met: the items' factor loadings on the corresponded constructs are higher than 0.7 (Chin, 1998), Cronbach's α and composite reliability of all constructs are higher than 0.7 and confirm reliability of the measurement model, and Average Variance Extracted (AVE) is higher than 0.5, exceeding the variance due to measurement error for that construct. Table 11-3 displays the construct items' reliabilities (Cronbach's α , Composite Reliability), Average Variance Extracted (AVE) and factor loadings and confirms convergent validity for the latent constructs. Discriminant validity is also confirmed, i.e., the AVE of each construct is higher than the construct's highest squared correlation with any other construct (Fornell & Larcker, 1981). Tables 11-4 and 11-5 presents the variables' correlation matrix. In this table, the diagonal elements are the square root of the AVE of a construct.

Table 11-3. Results for the Latent Constructs of the Measurement Model

Construct Items	Factor Loadings (>0.7) ^a	Cronbach Alpha (>0.7) ^a	Composite Reliability (>0.7) ^a	Average Variance Extracted (>0.5) ^a
AC¹_{FA}		0.791	0.864	0.625
FHAR	0.884			
FEAS	0.822			
FSAM	0.706			
FRAN	0.709			
TTDM	0.702			
AC²_{SA}		0.832	0.889	0.668
FACC	0.833			
FREP	0.856			
TTDM	0.706			

^a Indicates an acceptable level of reliability and validity.

¹ AC_{FA}: Autonomous Control factor, computed only for the full-autonomous group (n=57)

² AC_{SA}: Autonomous Control factor, computed only for the semi-autonomous group (n=56)

TTDM: Time on decision making, FHAR: Frequency of choosing harder, FEAS: Frequency of choosing easier, FSAM: Frequency of choosing same, FRAN: Frequency of choosing random, FACC: Frequency of acceptance, FREP: Frequency of replacement

Table 11-4. Measurement Model (Discriminant validity) for the full-autonomous group (n=57)

	1	2	3	4	5
1. Response Time Effort	1.000				
2. Time to answer correctly	0.412	1.000			
3. Time to answer wrongly	-0.380	-0.129	1.000		
4. Autonomous Control	0.680	0.534	-0.424	0.791	
5. Learning performance	0.688	0.645	-0.441	0.782	1.000

Table 11-5. Measurement Model (Discriminant validity) for the semi-autonomous group (n=56)

	1	2	3	4	5
1. Response Time Effort	1.000				
2. Time to answer correctly	0.377	1.000			
3. Time to answer wrongly	-0.350	-0.073	1.000		
4. Autonomous Control	0.764	0.380	-0.366	0.817	
5. Learning performance	0.825	0.565	-0.444	0.783	1.000

11.5.2. Testing hypotheses

A bootstrap procedure with 3000 resamples was used to test the statistical significance of the path coefficients (β value) in the model, as well as the differences in path coefficients between the two treatment groups. The results are summarized in Table 11-6.

Table 11-6. Hypothesis testing results

Hypothesis	Path	FA (N=57)			SA (N=56)			Diff. between FA/SA	Sig. of diff.	Result
		β_{FA}	t	P	β_{SA}	t	P			
H1	AC → LP	0.429	4.525*	0.000	0.460	4.184*	0.000	0.032	0.827	Equal
H2	AC → TTAC	0.534	8.416*	0.000	0.380	3.461*	0.001	0.154	0.226	Equal
H3	AC → TTAW	-0.424	5.119*	0.000	-0.366	2.530*	0.012	0.058	0.728	Equal
H4	AC → RTE	0.680	15.537*	0.000	0.764	14.737*	0.000	0.084	0.216	Equal

* $p < 0.05$, AC: Autonomous Control, LP: Learning performance, TTAC: Time to answer correctly, TTAW: Time to answer wrongly RTE: Response Time Effort

As seen from table 11-6, all hypotheses **H1**, **H2**, **H3** and **H4** regarding the direct effect of autonomous control on learner performance strongly supported for both treatment groups, whereas no statistically significant difference were identified between the path coefficients of FA and SA learners in the structural model, based on the method of Sarstedt et al. (2011).

11.5.3. Overall Model Fit

The suggested model explains almost the 75% of the variance in performance for the FA group, and 81% of the variance in performance for the SA group, which both are statistically significant. The cross-validated predictive relevance of the model was confirmed in both cases ($Q_{FA}^2=0.707$; $Q_{SA}^2=0.749$). Table 11-7 synthesizes the total effects of the selected factors, as well as the variance (R^2) and cross-validated predictive relevance (Q^2) explained by the proposed model.

Table 11-7. R^2 , Q^2 and Direct, Indirect and Total effects

<i>Endogenous</i>	R^2	Q^2	<i>Exogenous</i>	<i>Dir. effect</i>	<i>Indir. effect</i>	<i>t</i>	<i>Total effect</i>
LP	0.752	0.707	AC_{FA} ¹	0.429	0.371	4.666*	0.800
			TTAC	0.309		5.845*	
			TTAW	-0.137		2.630*	
			RTE	0.218		2.018*	
			AC_{FA}	0.534		8.202*	
			AC_{FA}	-0.424		5.072*	
			AC_{FA}	0.680		16.526*	
TTAC	0.817	0.749	AC_{SA} ²	0.460	0.365	3.911*	0.825
			TTAC	0.306		3.545*	
			TTAW	-0.199		2.627*	
			RTE	0.230		2.347*	
			AC_{SA}	0.380		3.225*	
			AC_{SA}	-0.366		2.647*	
			AC_{SA}	0.764		15.509*	

* $p < 0.05$

¹ **AC_{FA}**: Autonomous Control factor, computed only for the full-autonomous group ($n=57$)

² **AC_{SA}**: Autonomous Control factor, computed only for the semi-autonomous group ($n=56$)

LP: Learning performance, **TTAC**: Time to answer correctly, **TTAW**: Time to answer wrongly **RTE**: Response Time Effort

The measurement results are summarized in Figure 11-2.

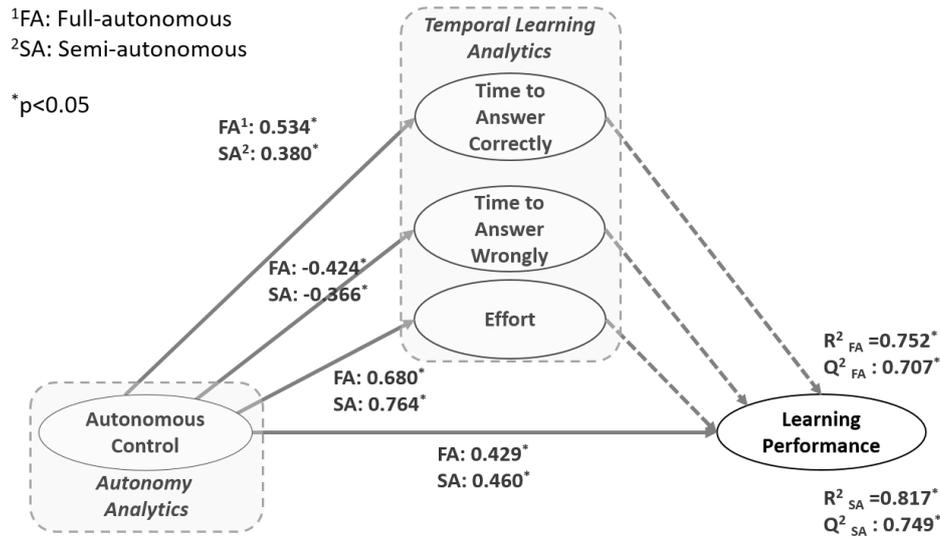


Figure 11-2. Path coefficients of the research model, overall variance explained (R^2) for test score and cross-validated predictive relevance (Q^2).

11.5.4. Independent samples t-tests and ANOVA results

Table 11-8 demonstrates the descriptive statistics for the three groups with respect to the learning performance, whereas table 11-9 depicts the independent samples t-test results regarding students' learning outcomes between the two treatment groups, as well as between each of the treatment groups and the control group. In this table, the last column illustrates the Hedge's g effect size in each case.

Table 11-8. Descriptive statistics for learning performance for the treatment and control groups

GroupID	N	Mean	Std.Dev (SD)
Full-autonomous	57	5.667	2.173
Semi-autonomous	56	5.679	1.936
No-autonomous (control)	55	4.509	1.631

Table 11-9. Independent samples t-test results for learning performance

Groups	F	df	t	95% CI		Hedges' g
				Lower	Upper	
FA vs. SA	1.000	111	-0.031	-0.7798	0.7560	0.005
FA vs. NA	7.055	110	3.178*	0.4358	1.8793	0.602
SA vs. NA	2.999	109	-3.438*	-1.8436	-0.4953	0.652

* $p < 0.05$, **FA:** Full-autonomous, **SA:** Semi-autonomous, **NA:** No-autonomous

As seen from this table, there were significant differences in performance for both treatment groups and the control group. Specifically, for the FA ($M=5.667$, $SD=2.176$) and NA ($M=4.509$, $SD=1.631$) groups, as well as for the SA ($M=5.679$, $SD=1.936$) and NA ($M=4.509$, $SD=1.631$) groups, the effects of autonomous control on performance were large ($g_{FA}=0.602$; $g_{SA}=0.652$).

Table 11-10 presents the results for ANOVA tests for each one of the analytics parameters of autonomous control (i.e., FEAS, FHAR, FSAM, FRAN, TTDM, and FACC, FRAN, TTDM), as well as of the other learning analytics (i.e., TTAC, TTAW, RTE) with respect to the different performance-based student clusters. The η^2 effect size was calculated, as well. In all cases, the Levene's test for homogeneity of variances could not reject the hypothesis of equal variances (*sig.*>0.05).

Table 11-10. ANOVA results for the learning analytics factors on the different performance-based clusters

	Full-Autonomous			Semi-Autonomous			No-Autonomous		
	F	p	η^2	F	p	η^2	F	p	η^2
TTAC	15.846*	0.000	0.370	8.919*	0.000	0.252	21.334*	0.000	0.451
TTAW	5.005*	0.010	0.156	7.844*	0.001	0.228	7.850*	0.001	0.232
RTE	17.485*	0.000	0.393	49.936*	0.000	0.653	11.200*	0.000	0.301
FHAR	48.385*	0.000	0.642						
FEAS	6.547*	0.003	0.195						
FSAM	14.533*	0.000	0.350						
FRAN	1.670	0.198	0.058						
TTDM	3.771	0.029	0.103	4.461*	0.000	0.120			
FACC				50.769*	0.000	0.657			
FREP				27.141*	0.000	0.506			

**p*<0.05

TTAC: Time to answer correctly, *TTAW*: Time to answer wrongly, *RTE*: Response Time Effort, *FHAR*: Frequency of choosing harder, *FEAS*: Frequency of choosing easier, *FSAM*: Frequency of choosing same, *FRAN*: Frequency of choosing random, *TTDM*: Time on decision making, *FACC*: Frequency of acceptance, *FREP*: Frequency of replacement

11.6. Discussion

The results of this study revealed consistent patterns of learners' behaviour in (controlled) autonomous self-assessment conditions. Other interesting (i.e., informative and actionable) findings regarding the role of autonomous control were encountered as well. In this section we further elaborate on these findings and their contribution to deeper understand the impact of autonomy on learning performance, effort and response-times from a learning analytics point of view, with respect to the research questions.

Exploring the effect of learner autonomy on performance is not a new field of inquiry. Autonomy has been extensively studied through the lenses of learners' perceptions and has been acknowledged for motivating learners (i.e., increasing their intrinsic motivation), and for its positive impact not only on the learning outcome, but on the general learning gain, as well (Fletcher & Shaw, 2012; Ghorbandordinejad & Ahmadabad, 2016; P. Hu & Zhang, 2017; León et al., 2015; Luke, 2006). The innovation and contribution of the present study stems from the adoption of a learning analytics perspective and from the exploitation of learner interaction data to investigate the relationship between autonomy and other performance factors.

In particular, the demonstrated approach was based on two core pillars: (a) previous research studies used the TLA model to explain performance, with statistically significant results (Papamitsiou et al., 2016, 2014), and (b) previous research studies associated learners' perceptions

of autonomy with commonly used learning analytics measures (e.g., frequencies of events and response-times), although these studies either did not model autonomy in terms of analytics or they did not further processed these data (Dawson et al., 2015; Giesbers et al., 2013; Rienties et al., 2014; Sergis et al., 2018). In the current study, autonomy is modelled in terms of frequencies of interaction types and time-spent on decision making. The consistency reliability of both factors that measure autonomy (in the FA and SA conditions) was confirmed (Table 11-3).

Next, patterns that can be associated to or infer specific autonomous behaviors were explored through a structural and measurement model. The overall prediction accuracy in this study was 75.2% for the FA group and 81.7% for the SA group, which both are statistically significant (Table 11-7). Moreover, the analysis revealed a consistent and positive (direct and indirect) effect of autonomous control on performance for both groups (Table 11-7), whereas the evaluation of the strength of this effect with additional analyses shown a large effect size (Hedge's $g=0.60$, $g=0.65$ for the FA and SA group respectively) (Table 11-9). Specifically, the direct effect of autonomous control on performance was strong ($\beta=0.429$, $t=4.525$, $p=0.000$; $\beta=0.460$, $t=4.184$, $p=0.000$ for the FA and SA group respectively), confirming hypothesis **H1** (Table 11-7). This finding is in line with and extends prior claims (Fletcher & Shaw, 2012; Ghorbandordinejad & Ahmadabad, 2016; Luke, 2006), which based their conclusions on learners' perceptions, by adding empirical evidence.

Moreover, autonomous control was found to be a consistent and strong direct determinant of both types of response-times (i.e., TTAC, TTAW), confirming hypotheses **H2** and **H3** (Table 11-7). Indicating a relationship between autonomously choosing a leaning item and time-spent on processing it, this finding constitutes empirical evidence that when learners make their own self-enforced choices of learning items, they are likely to treat these items more responsibly, by devoting more time on trying to deal with them. As such, it verifies the respective assumptions in former studies (León et al., 2015; Sergis et al., 2018). It also explains *why* learners who feel autonomous are likely to engage in deeper-processing mode and to invest more time on hands-on activities: they want to be consistent with their choices and to take responsibility (Oxford, 2015).

The finding that explains *how* this happens (i.e., being responsible learners) concerns the consistently strong positive effect of autonomous control on learners' effort. Specifically, both FA and SA groups appear to be strongly engaged in terms of effort exertion, exhibiting high solution behaviour, and avoiding guessing ($\beta=0.680$, $t=15.537$, $p=0.000$; $\beta=0.764$, $t=14.737$, $p=0.000$ for the FA and SA group respectively). Beyond confirming hypothesis **H4** (Table 11-7), this finding is very interesting in the following different ways:

- It provides additional evidence to previous research studies that claimed that learners who perceive themselves as autonomous tend to exhibit higher engagement (León et al., 2015; Luke, 2006).
- Due to the very high positive impact of autonomy on effort, the latest could be perceived as an indicator of *autonomous capacity development*. Elaborating further on that claim, one can

support that, since students had the opportunity to (freely or guided) choose the next item, they had to make strategic decisions on their own, by considering which choice would foster learning achievement. Accordingly, they had to take advantage of the available options, yet responsibly select an item, being consciously involved with it, and consequently, modifying their autonomous capacity. Furthermore, the students in the SA group appear to exhibit higher effort – although not statistically significantly different. This might be due to the effect of the guided autonomy: learners are not by default experienced on how to be autonomous (Andrade, 2014; Little, 1995). In the adaptive testing conditions, the assessment itself guides the learners to achieve the best possible score. Enhancing the adaptive testing with autonomous options, the assessment environment offers to the learner the opportunity to practice autonomy in guided conditions. As a result, the learner progressively engages more in the self-enforced choices, makes better decisions, needs less guidance on how to become autonomous (i.e., increases autonomous capacity), and finally, performs better.

The lack of previous research – to the best of our knowledge – on exploiting analytics methods for exploring the differences on behavioural patterns of autonomy, with respect to performance-based student clusters, accounts for the added value of the present study.

The one-way ANOVA revealed a consistent pattern: statistically significant differences were discovered between the performance-based student clusters, regarding the temporal learning analytics (i.e., TTAC, TTAW and RTE) for both the two treatment groups and the control group. Moreover, the effect sizes of these parameters were strong in all cases ($\eta^2 > 0.14$). One can notice that for the NA group, the factor with the strongest effect size was TTAC ($\eta^2 = 0.451$) – in consistence with previous results (Papamitsiou et al., 2016, 2014) – whereas for the two experimental groups, the respective factor was RTE ($\eta^2 = 0.393$, and $\eta^2 = 0.653$ for FA and SA respectively) (Table 11-10). This finding further supports our claim that effort could be considered as an indicator of autonomous capacity development: since this factor is the one that accounts for the higher performance in the FA and SA groups, better discriminating advanced performers from basic and below basic ones, one can infer that this factor is responsible for balancing autonomous capacity development within these groups, as well.

In addition, and more precisely for the group of students who had full autonomy of choices (i.e., FA), there were statistically significant differences between the advanced, basic and below basic students with respect to the frequencies they selected items with higher ($F(2,54) = 48.385$, $p = 0.000$), lower ($F(2, 54) = 6.547$, $p = 0.003$) or same ($F(=2, 54) = 14.533$, $p = 0.000$) level of difficulty. However, these students exhibited similar behaviour with respect to randomly selecting the next item ($F(2, 54) = 1.670$, $p = 0.198$). Within the FA group, the highest difference between the differently performing students was on the frequency they selected harder items ($F(2,54) = 48.385$, $p = 0.000$) (Table 11-10). This finding can be interpreted as follows: when students have the option to make their own choices of assessment items, it is more possible that they will strategically

select items according to their known level of difficulty, rather than randomly. In addition, selecting items of same difficulty as the current one, is a “safe” choice for medium performing students, whereas the high performing students will go for the harder items in order to achieve a higher score, while the below basic students will prefer easier items, hoping for the best possible score. Furthermore, the effect size of the time-spent on decision making was moderate ($\eta^2=0.103$). The time that students spent on decision making is considerable, but it is expected that regardless of their expertise level, most students will take time to think about their next choice.

Furthermore, for the SA group of students, who received guided autonomy (in terms of proposing to them an appropriate question to answer, as well as providing them the option to reject the suggestion and replace the item), there were statistically significant differences between the high, medium and low performing students with respect to both frequencies of choices. The advanced students, although initially chose to replace the items that the system delivered to them, they gradually accepted the suggested items ($F(2,53)=50.769, p=0.000$), whereas the below basic students mostly replaced these items ($F(2,53)=27.141, p=0.000$), trying to get items that were easier to them. The effect sizes of both frequencies were strong ($\eta^2=0.657$ for FACC; $\eta^2=0.506$ for FREP) (Table 11-10). This finding indicates that, in the guided environment, the students shaped different achievement trajectories. Even though the assessment followed an adaptive testing approach (i.e., personalized), the differently performing students adopted different answering strategies, motivated by an autonomously enforced criterion. Moreover, as in the FA group, the effect size of the time-spent on decision making was moderate ($\eta^2=0.120$), indicating that most students will take considerable time to think about their next choice.

11.7. Implications for research and practice

This study presented a consolidated analysis of two implementations of controlled autonomy conditions in the context of self-assessment, in order to extend current knowledge on the impact of autonomous control on learning performance. The findings contribute to the literature in a number of ways and have several implications for research and practice.

First, this study adds to the autonomy literature by presenting a learning analytics-driven research method that revealed strong association between learners’ interactions within controlled autonomous and semi-autonomous learning environments and the learners’ achievement. Previous studies explained the effect of autonomy from a self-reported point of view (Fletcher & Shaw, 2012; Ghorbandordinejad & Ahmadabad, 2016; P. Hu & Zhang, 2017; León et al., 2015; Luke, 2006). However these studies focused on perceptions and failed to shed light to *how* actually the learners interact with the learning objects within (scalable levels of) autonomous learning conditions. Moreover, previous analytics-driven studies either did not model autonomy in terms of analytics or they did not further processed these data (Dawson et al., 2015; Giesbers et al., 2013; Rienties et al., 2014; Sergis et al., 2018). As such, regarding its methodology, this study is one of the first to employ learning analytics in the measurement of

autonomous control. The thorough analysis showcased that learning analytics can produce results that are comparable to – and (conditionally) can extend – those measured with self-reported instruments. Specifically, this study contributes to better understanding the specific patterns of interactions that the different performance-based student clusters adopt as achievement trajectories.

Another contribution of this study is that it proposed and evaluated a theoretical model for explaining students' performance by considering their autonomy, effort and response-times. The overall insights from the analysis are promising and reveal a consistent pattern: it was shown that most of the variance in learning outcome was explained by autonomous control (Table 11-7). It is indicative that the extracted effect sizes (Table 11-9) are large, but should be further explored through the lenses of a self-regulation learning perspective. Exploring the complex relationships between motivational and self-regulatory factors (Ryan & Deci, 2009) and the analytics-driven autonomous control factors and effort exertion, is expected to better justify the learners' performance achievements. This is within our future work plans.

The most intriguing finding of this study concerns the useful insights it offers for instructors and researchers, to help them identify mechanisms for autonomous capacity development. Learners are not intuitively aware on how to be autonomous (Andrade, 2014; Little, 1995), yet it was claimed that they should not only feel that they are independent, but to be guided to develop their capacity for autonomy, as well (Little, 1995). This study provides empirical evidence on the assumption that towards achieving autonomy, it is necessary to enhance learning environments with features that offer to the learners the opportunity for exercising control over their learning processes and being consciously involved in their own learning (Andrade, 2014). In particular, this study shown that enhancing the adaptive testing process with a notion of semi-autonomous control could contribute towards increasing autonomous capacity. Adaptive testing algorithms have been acknowledged for their efficiency to accurately discriminate the students' level of knowledge mastery, and classify them accordingly (Ala-Mutka, 2005). Enhancing an adaptive testing mechanism with semi-autonomous control acts as follows: regardless of accepting or replacing the suggested assessment item, the difficulty and the appropriateness of the item remain the same, and best fit the students' knowledge mastery level. As the self-assessment progresses, for the advanced students, the need for autonomously choosing the next item gradually fades-out, as they consciously engage more in answering the items (increased autonomous capacity). On the contrary, the medium and low performing students need more to maintain their control over the assessment, trying to get easier items. In this case, the effort (indicating their involvement) increases as well. The sense of autonomy triggered by the option of individually choosing the next item, is in fact a "pseudo-autonomous" control that increases the students' degree on engagement in the test (expressed as higher effort and response-times, as explained above). This can be interpreted as a "placebo" effect. Within guided conditions, the "pseudo-autonomous" control increased students'

autonomous capacity: they perceived the process as autonomous, they believed that they were making their self-enforced choices and they were consciously involved in the self-assessment, yet they were seamlessly guided by the system to increase their effort and improve their performance. However, these findings warrant further research and need to be validated by additional experimentation and bigger participant samples, and in different learning contexts (e.g., open-ended), as well.

11.8. Conclusions

Autonomy has been acknowledged for its positive impact on the learning outcome. Three core concepts can describe learner autonomy: (a) freedom of choice, (b) responsibility of choice, and (c) guidance to develop the capacity of the previous two, within the learning environments. This study adopted this conceptualization of autonomous learning, and shifted beyond self-reported perceptions. Towards better understanding and explaining learner autonomy, this study invested on the importance of using interaction data collected by learning environments. We exploited learning analytics to explore the learners' freedom of choices and to associate the level of autonomy with the learners' effort exertion to achieve high performance – as an indicator of engagement and thus, as a measure of responsible choice and autonomous capacity. The findings discovered consistent patterns of strategic decisions related to autonomous behaviour, that otherwise would not be possible to reveal. We suggest that autonomy should be perceived as a form of learning that should be considered in and facilitated by adaptive learning systems (i.e., taking advantage of their intelligence in order to guide the learners on how to become autonomous learners).

Chapter 12 : Modeling autonomous learning capacity development

*“All models are approximations.
Essentially, all models are wrong, but some are useful”*

George Edward Pelham Box

Towards an analytics-driven model for assessing autonomous learning capacity development in online self-assessment conditions.

12.1. Introduction

Adaptivity and adaptive learning environments are in the epicentre of the technology enhanced learning research community. The recent 2018 NMC Horizon Report Preview highlights the emergence of these systems. In the same report, it is acknowledged that the focus of these systems is on modifying the instruction anytime and providing the best possible support to the learners “to accurately and logically move [students] through a learning path, empowering active learning” (New Media Consortium, 2018, p. 9). The adaptive systems adjust their features to meet the learners’ characteristics and offer them a personalized learning experience (Brusilovsky et al., 2016). This chain of adaptive interactions results in continuously engaging the learners in controlling their own learning, as they move along the self-regulated learning continuum towards autonomy.

And, the more the learning turns online, the higher the need for learners to take control of their learning, and develop and sustain self-regulated and autonomous learning competences. Contemporary learning theories highlight the role of self-regulation in the learners’ personal development. Self-regulated learning (SRL) refers to a “self-directive process by which learners transform their mental abilities into academic skills” (Zimmerman, 2002, p. 65). It is also conceptualized as “an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment” (Pintrich, 2000, p. 435). This process involves the systematic planning, regulation and evaluation of learning goals and their attainment (Narciss et al., 2007). Self-regulated learners are aware of their learning processes and strategies, and adjust their behaviors to the specifications of the learning environments (McCardle & Hadwin, 2015). For Oxford (2008), the use of self-regulated learning strategies can develop learner’s capacity for autonomous learning.

Autonomy targets at fostering learners’ responsible self-initiative and allows them to determine the *selection* of what shall be learned, as well as the *critical evaluation* (reflection) of the learning tasks that were selected (Candy, 1991). Through the “lenses” of autonomy, the learners understand their needs, are aware of their self-directed learning goals, take control of and become responsible for their learning choices, monitor their progress, and critically reflect on their learning (Benson, 2001; Holec, 1981; Little, 1991). According to Holec’s (Holec, 1981, p.

3) definition, autonomous learning is “the *capacity* to take charge of one’s own learning”. A broader conceptualization of autonomy focuses on learners’ *capacity* that allows them to accept responsibility and take control of their own learning processes (Vanijdee, 2003). In Benson’s (2001) conceptualization of learner autonomy, the autonomous and self-directed learners take control over the cognitive, emotional, motivational, and behavioral processes of learning, as well as the independent use of learning material and technology. This approach implies that the autonomous learners exhibit self-regulation strategies, and that they make self-enforced decisions independently.

12.2. From self-regulated learning to autonomy

The concepts of autonomous and self-regulated learning share some similarities, and the relationship between them seems to be bidirectional. Loyens et al. (2008) elaborated on the conceptual clarity between autonomous and self-regulated learning. Both concepts involve the learners’ active engagement and goal-directed behavior, and whilst effective SRL focuses on helping learners become autonomous thinkers, the autonomous learners appear to exercise control over their SRL strategies. This is a reason why both terms are usually used in literature interchangeably (Andrade & Bunker, 2009; Lewis & Vialleton, 2011).

However, the *degree of control* the learners have over the learning process significantly differs (Loyens et al., 2008). Learners exercise autonomy when they make choices and act on them: freedom of choice is central to the idea of autonomy – it is always the learners who choose *what, where, and how* to learn. On the other hand, the concept of SRL places less emphasis on choices and more on guiding learners towards efficiently employing strategies. SRL seems more concerned with the *subsequent steps* in the learning process, such as setting goals, monitoring their progress, reflecting on the steps that were taken and changing their plans accordingly, and is usually described as a *favorable* learner characteristic, i.e., as the processes that the learner substantiates (Zimmerman, 2000).

For example, selecting *which* learning materials to study, according to the learners’ self-defined goals and priorities, corresponds to autonomous learning behavior, because the learner has the absolute control and responsibility of choice. In another example, deciding on the *time to allocate* on studying the learning materials corresponds to an SRL strategy employed. In the later example, although it might seem that the learner has control and responsibility for learning, however, the learning materials are very possible to have been determined by the instructor or the curriculum. In this case, the learner only adjusts her treatments and behavior regarding the materials: she is *not free* to choose *what* to study, but only *how much* she wants to engage with it.

12.2.1. Motivation of the research and research question

As apparent, autonomy is broader than self-regulation; autonomous learners are also capable of taking responsibility for their learning choices (Oxford, 2015), as well. Furthermore,

Little (1995) argues that autonomy is beyond independence; the learners should not only feel that they are independent, but to be guided to develop their capacity for autonomy.

However, learners do not intuitively know how to develop their capacity and achieve autonomy; they need to be guided in building the competences for efficient autonomous choices (McDevitt, 1997; C. White, 1995). For Oxford (2008), the use of self-regulated learning strategies can develop learner's capacity for autonomous learning. Figure 12-1 illustrates this approach.

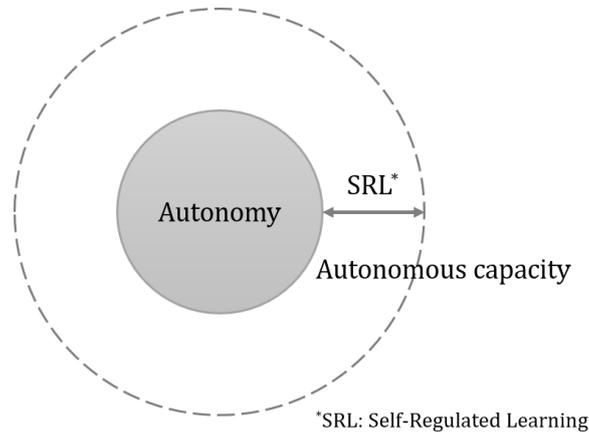


Figure 12-1. Developing capacity for autonomous learning with self-regulated learning strategies

However, although self-regulated forms of learning may *require* the exercise of strategies, yet, it is a “black box” how their usage *develops* autonomous capacity. As apparent, it is a challenge to address the issue of *how* to appropriately support learners in building their capacity towards achieving autonomy. This study targets at taking us beyond current knowledge on the topic, and explores how this could be achieved, using insights from learning analytics: it aims at identifying which self-regulated learning strategies and which facilitating factors/conditions affect the learners' autonomous capacity. A model for assessing the development of autonomy is absent from current literature: this is the first study – to the best of our knowledge – that aims at introducing a holistic, analytics-driven model for assessing autonomous learning capacity development. Thus, the research question is defined as follows:

RQ: *To what extent can we exploit learning analytics to assess learners' autonomous capacity development?*

12.3. The model for assessing autonomous learning capacity development

“Capacity development (or capacity building) is the process by which individuals, organizations, institutions and societies develop abilities to perform functions, solve problems and set and achieve objectives” (United Nations Economic and Social Council, 2006, p. 7). Accordingly, autonomous learning capacity development (ALCD) corresponds to the process by which learners (individuals) develop abilities to achieve learning autonomy (objective). Since SRL is a means towards achieving autonomy (Oxford, 2008), the abilities that the learners should develop can be mapped to the respective SRL strategies and should be appended to include the

independent, autonomous learners' choices, as well. These strategies, along with the autonomous interactions can next be measured with appropriately configured learning analytics parameters. Finally, the analytics shall be fed to the learners to trigger their awareness regarding their personal progress and development.

As such, ALCD is a four-step process:

- the learners' self-reported perceptions of self-regulation are measured (input),
- the learners practice SRL strategies with guided autonomous choices (according to Andrade (2014), McDevitt (1997) and White (1995)), and their interactions are tracked and measured (process),
- the respective analytics (i.e., autonomous learning capacity analytics - ALCA) are extracted (output), and
- the autonomous learning capacity analytics (ALCA) are forwarded to back to the learners (feedback).

This process is illustrated in Figure 12-2.

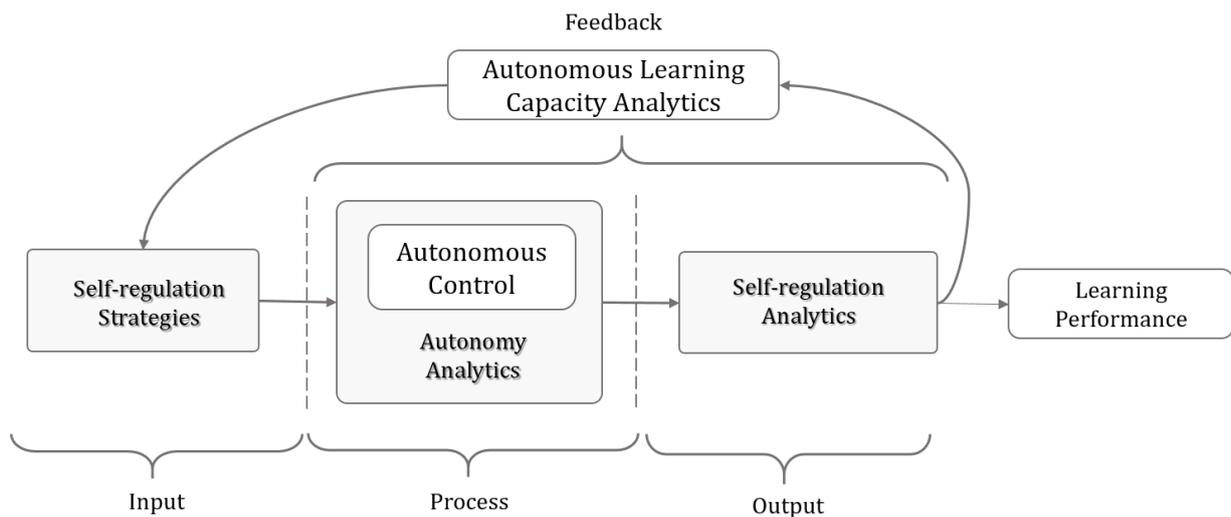


Figure 12-2. Overview of the model of autonomous learning capacity development

The difference in ALCA corresponds to the development in the learners' autonomous learning capacity, and can be used for assessing the progress made.

It should be noted that the development of autonomous learning capacity is a long-term process: it requires a relatively long period of time, during which the learners progressively gain their autonomy, by being trained in controlled conditions, until they reach the desired level of freedom to make responsible choices. Sustaining autonomy is a more complex process: it requires continuous practice of the gained skills, within less controlled and more open learning settings.

Andrade (2014) prompted that towards achieving autonomy, there is a need to develop technology enhanced learning environments within which the learners shall be given the

opportunity for exercising control over their self-regulated learning processes and to be consciously involved in their own learning. In order to facilitate this objective, the learning environment should provide functionalities to the learners, that: (a) support SRL strategies (e.g., implement a help-seeking mechanism, a time-awareness mechanism, etc.), (b) allow for self-directed autonomous choices, and (c) guide the learners to make decisions (e.g., with an adaptation mechanism).

12.4. Particularizing the ALCD model in online self-assessment

In view of the above, this study targets at defining the specific SRL strategies and autonomous learning capacity analytics to be integrated into the suggested ALCD model, and is contextualized in online self-assessment conditions.

It has been argued that self-assessment leads students to a greater awareness, by training them to self-regulate their motivation and behavior, as well as by fostering reflection on their own progress in knowledge/skills, and finally, to understanding themselves as learners (McMillan & Hearn, 2008; Nicol & Macfarlane-Dick, 2006). As a learning process, self-assessment encompasses learners' evaluation of the quality of their own cognition and their own behavior and choices: *"How much have I progressed towards achieving my learning goals?", "Have my choices been correct and did they lead me to achieve my learning objectives?", "What could I have done differently?"*

By emphasizing on the learners' personal choices and control over their learning outcome, self-assessment is expected to promote the development of learners' capacity to self-regulate their performance and to preserve their autonomy. For this reason, self-assessment procedures are ideal for measuring learners' actual self-regulation and autonomous interactions.

12.4.1. Selection of SRL strategies and learning analytics parameters

12.4.1.1. Determination of the autonomy facilitating conditions and autonomous learning analytics

It was found that in online self-assessment conditions, facilitated by autonomy-enhanced Computerized Adaptive Testing (CAT) procedures, the self-directed choices increased learners' on-task engagement (in terms of response-times) and effort, and resulted in higher achievement levels. Based on these findings, it was suggested that autonomy-enhanced CAT could be used as a means to guide learners to developing their autonomous capacity (Chapter 11).

Accordingly, in order to enhance CAT with autonomous choices, a set of adjustments is required. Typically, a CAT self-assessment system delivers the next self-assessment task to the learners, according to the correctness of the previous answer and the discrimination ability of the task (Algorithm 1, Appendix B). In an autonomy-enhanced CAT, the students can either accept that task, or ask for another one to replace it (which again is selected according to the correctness of the previous answer, based on the similarity ranking of the tasks). The autonomy-enhanced "semi-adaptive" self-assessment test captures the learners' interactions with the tasks, and the extracted autonomy analytics are measures commonly used in the field of learning analytics (e.g.,

response-times, frequencies) (Joksimović et al., 2015; Kizilcec et al., 2017; Kovanović et al., 2015), and include the frequency of accepting the task (FACC), the frequency of replacing the task (FREP), and the Time on Decision Making (TTDM). Both analytics measures

12.4.1.2. Determination of SRL strategies

In a contextualization of SRL to mostly apply in online learning environments, Broadbent (2017) and Barnard et al. (2009) suggested six prevalent strategies, i.e., goal-setting, time-management/study-environment, help-seeking, task-strategies (e.g., effort regulation, rehearsal), peer-learning, and self-evaluation. However, peer-learning and self-evaluation do not apply when the online learning environment is self-assessment oriented. According to previous findings (Chapters 10), in online self-assessment conditions, goal-setting and time-management have strong positive effects on autonomous control, effort-regulation moderately positively affects learners' autonomy, while help-seeking has a strong negative effect.

12.4.1.3. Determination of analytics parameters for empirically measuring SRL strategies

Based on these findings, each of these strategies should be mapped to a mechanism for empirically practising the strategy. The learners shall interact with these mechanisms, and the analytics for empirically measuring the strategies will be extracted from the logged interactions. Towards determining the analytics parameters for empirically measuring the SRL strategies, previous approaches from relevant literature are expected to shed light.

In particular, the effort-regulation strategy can be practically measured as the effort-exertion that the learners exhibit during dealing with the self-assessment tasks. In this case, an indicative measure is Response-Time Effort (RTE), that measures the proportion of items which the learner tries to solve (solution behaviour) instead of guessing the answers (Wise & Kong, 2005). This measure has been found to be positively affected by autonomous learning interactions (Chapter 11), indicating that when the learners choose the tasks they consider more appropriate for their learning goals, then they select a task and are consciously involved with it, trying to complete the task, and avoiding to guess the solution.

Furthermore, in self-assessment conditions, the goal-expectancy strategy has been strongly associated with learners' response-times and the correctness of the delivered answer. Specifically, goal-expectancy has been found a strong direct determinant of the Time to answer correctly (TTAC) and Time to answer wrongly (TTAW) factors (Papamitsiou et al., 2014). As such, goal-expectancy can be measured using these analytics.

In addition, the learners can exercise the help-seeking strategy using the previously suggested metacognitive instructional help mechanism, implemented as task-related analytics visualizations (Chapter 7). Relevant research has shown that engaging in sense-making from the task-related analytics visualizations, leads the learners to better regulating their engagement with the self-assessment tasks (Chapter 8). The learners' interactions with the metacognitive help

mechanism are measured in terms of frequencies of requests for the visualizations (FRAV), as well as time-spent on viewing, processing and sense-making (TAVV) (Chapter 7). Although a strong negative effect of help-seeking on autonomous control was found in previous studies (Chapter 11), however, it was also found that using the task-related analytics visualizations, the learners understand deeper the real requirements of the tasks (Chapter 8). As such, it is expected that the more the learners consult the analytics visualizations, and make-sense about how demanding the tasks really are, the more self-directed, and autonomous the selection of the next task will be.

Finally, the learners can improve their time-management skills using a mechanism that keeps them time-aware, i.e., a timer that they can hide and show as many times they want, for as long as they want, at the moments they need to do so. Time itself cannot be managed because it is fixed. However, time-awareness is expected to affect how the learners adjust their strategies and make self-enforced decisions in time-limited self-assessment activities: the learners should select the next task by considering the available time, beyond other factors, as well. In that sense, time-awareness is key for time-management. Indicative analytics for the usage of the timer include the frequency of showing/hiding it (FSHT) and the duration of showing it (TTST).

Table 12-1 synthesizes the parameters of autonomous learning capacity analytics. This table includes the mapping between an SRL strategy and the measured SRL analytics, as well as the autonomy analytics.

Table 12-1. The autonomous learning capacity analytics - ALCA

SRL Strategy	SRL Analytics		
Effort Regulation	Effort exertion	Response Time Effort	RTE
Goal expectancy	Response-time	Time to Answer Correctly	TTAC
		Time to Answer Wrongly	TTAW
Help-Seeking	Metacognitive Help	Frequencies of requests for the visualizations	FAVR
		Time-spent on viewing the visualizations	TAVV
Time-management	Time-Awareness	Frequency of showing/hiding the timer	FSHT
		Duration of showing the timer	TTST
Autonomy	Autonomy Analytics		
	Autonomous Control	frequency of accepting the task	FACC
		frequency of replacing the task	FREP
		Time on Decision Making	TTDM

Based on the above, Figure 12-3 demonstrated how the model for assessing autonomous learning capacity development is particularized in online self-assessment conditions. The analytics about the autonomous learning capacity will next be fed to the learners in order to increase their awareness regarding their progress in achieving autonomy.

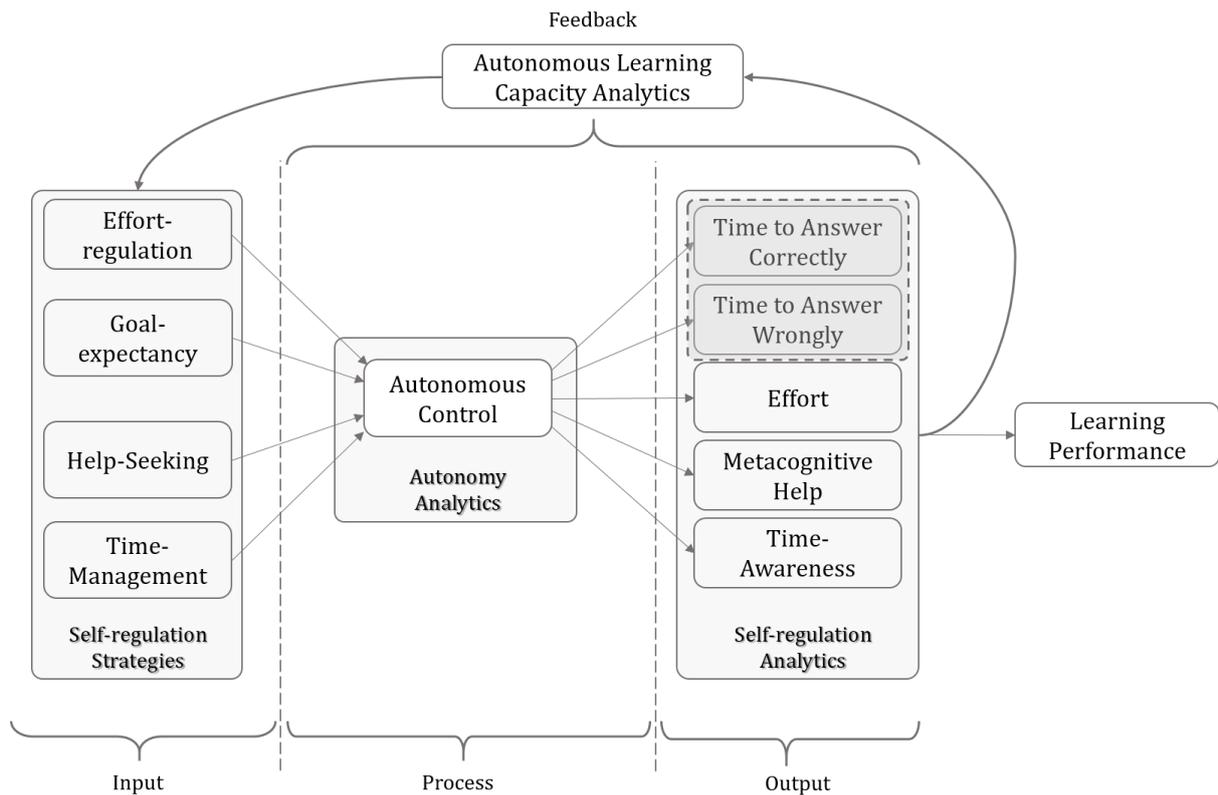


Figure 12-3. A model for assessing autonomous learning capacity development in online self-assessment conditions

As explained in section 12.2, the development of autonomous learning capacity is a long-term process: it requires a relatively long period of time, during which the learners progressively gain their autonomy, by being trained in controlled conditions, until they reach the desired level of freedom to make responsible choices. As such, in order to estimate the development in autonomous capacity, multiple measurements are required. The difference in the analytics parameters between the successive measurements corresponds to the development in the learners' autonomous learning capacity. In order to estimate the progress achieved, panel data analysis methods can be utilized.

The empirical evaluation of the model is necessary. This is within our future work plans.

Chapter 13 : Supporting groups with recommendations

"It is through others that we develop into ourselves."

Lev Vygotsky

Motivating students in collaborative activities with game-theoretic group recommendations

13.1. Introduction

Recommending educational resources to groups of students instead of individuals is a common practice in collaborative learning contexts. However, delivering those resources that will prompt all students' thinking and learning, motivate all group members to actively engage in the collaborative activity and accomplish their learning needs, is a non-trivial task (Berkovsky & Freyne, 2010; Masthoff, 2011). The reason behind this claim is that students in a group may not be fulfilled by the same items, yet wish to meet their own learning goals, making it difficult for the group to reach a consensus regarding the consistency between the expected gain from the resources and the actual experience, as well as their intention to use the resources. When students participate in groups, each student is influenced by the perceptions, decisions and choices of the other students, and they all together try to agree upon what is considered as useful, helpful, and whether it facilitates the activity's requirements and their own learning objectives. The problem this study addresses is how to optimally recommend educational resources to groups of students with respect to each individual member's motivation and intention to use the recommendation.

Inspired from Carvalho & Macedo (2013), we argue that *Game Theory* could efficiently solve "conflicts of interest" between the group members (Nash, 1951) and guide the recommendation of educational resources. In collaborative learning contexts, "conflict of interest" refers to the situation in which the individual students' perceptions regarding how much the resources will facilitate the students' self-determined goals and will contribute to improving their own competences (i.e., personal benefits) is potentially competing to the benefits of the other group members. It implies a diversity in group members' self-motivated considerations regarding the usefulness of the resources, as well as in their behavioral intention to finally use the resources.

Game theory is "a study of mathematical models of conflict and cooperation between intelligent rational decision-makers (players)" (Myerson, 1985). This study demonstrates a method for recommending educational resources to groups of students based on non-cooperative games. *Non-cooperative* is a technical term and not an assessment of the degree of cooperation among players in the game; a non-cooperative game can model cooperation, focusing on predicting individual players' actions and payoffs, but the players make self-enforced decisions independently (Nash, 1951).

13.2. Related Work, Motivation of the Research and Research Question

Recommender systems (RSs) are programs that search in collections of items (e.g. products, services or people) and target at suggesting to their users those items that will best

satisfy their preferences, by inferring the users' potential interest in the items (Resnick & Varian, 1997); the decision is driven by the collected and analyzed information about the items, the users and the user-item interactions (Bobadilla, Ortega, Hernando, & Gutiérrez, 2013). The level of relevance of an item to a user is usually expressed by the degree of user's appreciation on it, i.e., a rating. Literature on the topic is rich (Bobadilla et al., 2013; Lu, Wu, Mao, Wang, & Zhang, 2015; Park, Kim, Choi, & Kim, 2012). Prevalent approaches include collaborative filtering (Sarwar, Karypis, Konstan, & Riedl, 2001), content-based (Pazzani & Billsus, 2007) and knowledge-based (Burke, 2002) techniques. Due to drawbacks and limitations of these techniques (e.g., prediction accuracy, data sparsity, cold-start issues), more sophisticated approaches have been proposed, e.g., fuzzy-logic based (Yager, 2003), social network-based (He & Chu, 2010) and context aware RSs (Adomavicius & Tuzhilin, 2008).

13.2.1. Group recommender systems

In many cases, in different application domains, the users have to carry out an activity together, as a group, resulting in a rapid increase of RSs that cope with the challenge of addressing recommendations for groups of users instead of individuals. In these cases, the goal is to recommend items that would meet all users' preferences as much as possible (Adomavicius & Tuzhilin, 2008). However, recommending to groups is more complex than recommending to individuals (Masthoff, 2011); group members usually do not have the same preferences and interests, making it difficult to reach to an agreement between them and satisfy them all. Dissimilarities among users are apparent in group recommenders (McCarthy et al., 2006), and as a result, how the group members' disagreements on the same items are bridged, is critical for the effectiveness of group recommendations (Amer-Yahia, Roy, Chawlat, Das, & Yu, 2009).

In order to address issues of disagreement between the group members, the researchers from the group RSs domain adopted the semantics originating from the Social Choice Theory (Fishburn, 1973; Pennock, Horvitz, & Giles, 2000). Methods like popularity voting, most respected person, least misery, most pleasure and average (Masthoff, 2011) have been applied to the decision making process to solve conflicting ratings of preferences and to establish an automatic way of how a group of people can reach to consensus (Masthoff & Gatt, 2006). For example, MusicFX, a group RS for selecting a music station, uses a variant of the average without misery strategy for group profile aggregation (McCarthy & Anagnost, 1998). INTRIGUE, a hybrid system for sightseeing destination recommendation to tourists, takes into account sub-groups' characteristics and aggregates the sub-group recommendations according to the created models (Ardissono, Goy, Petrone, Segnan, & Torasso, 2003). PolyLens recommends movies to groups by fusing recommendations using the least misery criterion (O'Connor, Cosley, Konstan, & Riedl, 2001), whereas HappyMovie uses the individuals' "social trust" and personality in an average profile strategy (Quijano-Sanchez, Recio-Garcia, Diaz-Agudo, & Jimenez-Diaz, 2013). Other works focus on integrating the social, expertise and preference dissimilarity in the recommendation process (Gartrell et al., 2010). For a systematic review, see (Masthoff, 2015).

In the educational group RSs research field, and, although still rather sparse, the prevalent approaches include: (a) merging individual preferences in a pseudo group profile according to an aggregation strategy, prior to generating the recommendation (e.g., Dwivedi & Bharadwaj, 2015; Rodríguez, Giraldo, Tabares, Duque, & Ovalle, 2016), (b) constructing groups with high inner member similarity (homogeneous) and recommending resources from a merged list of recommendations, generated for each group member individually (e.g., Kompan & Bielikova, 2016; Yanhui, Dequan, Yongxin, & Lin, 2015), or (c) evaluating aggregation methods and applying classification on meta-data, including the prior evaluation results and a set of learners' characteristics (Zapata, Menéndez, Prieto, & Romero, 2015). In these approaches, the consensus functions aggregate the group members' personalities, interests, and learning styles.

13.2.2. Motivation of the research - research questions

However, these methods share four types of drawbacks related to: (a) the group inner similarity, (b) the aggregation strategies, (c) the number of recommended resources, and (d) the individual members' conformity degree, i.e., the adjustment of one's opinion towards the majority seeking for approval (Asch, 1951). Regarding the first drawback type, homogeneous groups is an unwanted restriction, since homogeneity in group synthesis is not always possible to achieve. Moreover, heterogeneity in groups is considered as more beneficial for learners in collaborative learning contexts (Graf & Bekele, 2006; Slavin, 1987). In addition, studies in cognitive psychology state that in processes that concern judgments (like decision making in a group recommendation process), both concepts of similarity and dissimilarity share equal importance (Medin, Goldstone, & Gentner, 1990; Mussweiler, 2003). Furthermore, the existing methods fail to achieve high quality performance and goodness of recommendation for the majority of students in heterogeneous groups (Masthoff, 2015; Yu, Zhou, Hao, & Gu, 2006). Regarding the second drawback type, not all aggregation strategies work efficiently in all cases (Zapata et al., 2015), whereas evaluating the aggregation strategies prior to applying one of them is time consuming (if not raising a fairness issue in recommendation). Regarding the third drawback type, existing methods recommend only one item at a time, though it is very likely that students want to access multiple learning resources. In this case, they would be more pleased with a sequence of suggested items. Finally, regarding the fourth drawback type, the focus of recommendation is on the overall group satisfaction, bypassing the relevance of the recommended items to the corresponding students, and how beneficial these items are to the learning subjects themselves. To the best of our knowledge, none of the abovementioned approaches consider in the recommendation process the students' behavioral intention to take the recommendation and use the resources. Similarly, measuring students' persistence – indicating the actual engagement with the resources (in a learning analytics fashion) – is missing.

Thus, the emerging research questions are:

RQ1: *Can we accurately and efficiently recommend sequences of educational resources to homogeneous and heterogeneous groups of students, with respect to both the individuals' and the group's motivation, and intention to use the resources?*

RQ2: *What is the impact of a recommendation on individual students' persistence as well as on the groups' learning performance in the collaborative problem-solving activity?*

Towards answering the research questions, we argue that non-cooperative games could efficiently solve conflicts of interest between group members and guide the recommendation of a sequence of learning resources.

13.3. Problem Formulation as Non-Cooperative Game

We examine the case of having students who collaboratively solve problems in groups (at least two members). The group members are students with *potentially conflicting* self-determined learning goals and expectations (i.e., motivation to be engaged in the collaborative problem-based activity and perceive as useful the suggested items). Thus, the groups might be homogeneous (high similarity between the group members), mildly heterogeneous (medium similarity between the group members) or heterogeneous groups (low similarity between the group members). The goal is to recommend to each group those items (single or sequence) that will be beneficial to the group as a whole – supporting the group members to efficiently complete the assigned collaborative activity – and that are expected to maximize each member's motivation to take the recommendation and use the items, and their actual engagement with the items, i.e., persistence.

13.3.1. Problem definition

Consider a set of students $L = \{l_i\}$ and a set of educational resources $R = \{r_j\}$; the indexes i, j refer to an individual student or resource (item), respectively. Let G be a set of all groups that may be formed by L ; then $|G| = 2^n - n - 1$. If $g \in G$, then $|g| = k$, the number k of group members in group g , with $k \geq 2$. For each student i and each item j , the student's motivation to take the recommendation m_{ij} can be estimated from the student's self-enforced evaluation of how much the particular item corresponds to the student's expectations, how much it prompts the student's thinking and how useful the student considers the item to be, with respect to the student's learning goals and competences. Student's motivation triggered by each item is measured with an appropriate questionnaire in a 5-point Likert-like scale (briefly outlined in section 13.3.3); if a student has not yet evaluated an item, then $m_{ij} = 0$. Also, consider \hat{m}_{ij} as the predicted motivational excitement that item j imparts to student i , i.e., a *decision criterion* for selecting the item, approximated with the Matrix Factorization technique (Koren, Bell, & Volinsky, 2009) (briefly explained in 13.3.4).

The items to be recommended to each group should not have been previously seen or evaluated by any of the group members, and are strategically designed to promote the students' interest in a specific learning topic and improve their competences accordingly. We model the group recommendation problem as a non-cooperative game, i.e., a triad (k, Q, f) where:

- The k students (group members) are the *players*.
- The set of unrated items $Q = \{q_z\} = \bigcap_i \{j \mid \hat{m}_{ij} = 0\}$, $Q \subseteq R$, are the available *actions*; a vector $x = (q_1, \dots, q_\mu) \in Q$ is a *strategy profile*.

- The *payoff* function for a student i and a strategy profile x calculates the predicted motivation of student i in the group, resulting from the actions by all group members – including himself – as the average individual predicted motivation from all items in x ;

$$\text{payoff is computed by: } f_i(x) = \frac{\sum \hat{m}_{iz}}{|q|}, \text{ where } |q| \text{ is the total number of items in } x.$$

The items that will be recommended to the group of students are those in the Nash Equilibrium (NE) (single item or sequence of items). A strategy profile $x^* \in Q$ is a NE if: $\forall i, x_i \in Q : f_i((x_i^*, x_{-i}^*)) \geq f_i((x_i, x_{-i}^*))$, where x_i is a strategy profile for student i and x_{-i} is a strategy profile of all students except for student i . In other words, considering that the other students will not modify their own strategy, the student who has the option of deviating should have no benefit by unilaterally changing his own strategy. In the group recommendation problem, in the NE, no student i can further increase their motivation from the recommendation by altering their strategy to $x_i \neq x_i^*$, provided that all other students stay with their selected strategies. The students' strategies converge to the NE after an iterative best-response strategy update.

In case there are more than one strategies that are NE, the recommendation solution for this group is the one that is “*socially optimum*”: no other strategy $x^* \in Q$ has both a weakly better payoff for all students and a strictly better payoff for some student: $\forall i f_i(x^*) \geq f_i(x)$ and $\exists i f_i(x^*) > f_i(x)$. In other words, if the recommendation solution is Pareto optimal, it is impossible to improve the motivation of a student without worsening the motivation of another student, indicating an optimal solution for the group as a whole. However, it is very possible that none of the NE is Pareto optimal. In this case, we calculate the distance between the highest and lowest payoffs in the strategies that are NE, and select the strategy that minimizes this distance, indicating a fair solution for the group. Algorithm 1 presents the algorithm for finding the NE and selecting the best-response strategy.

Algorithm 1: Finding the Nash Equilibrium and selecting the best-response strategy

Input: k, Q, f (The non-cooperative game)

Output: $x^* \in Q$ (The Nash Equilibrium profile strategy)

```

1. for i=1; i ≤ k; i++ do //for all students
2.   assign  $x_i \in Q, x_{-i} \in Q$  //initialize profile strategy
3. repeat
4.   repeat
5.     for i=1; i ≤ k; i++ do //for all students
6.       assign  $x_i^* \in Q, x_{-i}^* \in Q$  //assign another strategy
7.       compute  $f_i(x_i^*, x_{-i}^*)$  //compute the payoff
8.     until  $f_i(x_i^*, x_{-i}^*) \geq f_i(x_i, x_{-i}^*)$  //no student has incentive to change the strategy
9.     if  $f_i(x^*) > f_i(x)$  then //check for Pareto efficiency
10.      Pareto=true
11.    compute  $d_i$  // difference max-min in NE
12.  until Pareto or min  $d_i$ 
13. return  $x^* \in Q$  //the NE

```

Finally, for calculating the predicted group motivation and intention to use the recommendation, a group consensus function $M(g,x)$ computes the average motivation from each item in the recommended strategy x for the group g , where $f_i(x)$ is the payoff for each

member i : $M(g,x) = \frac{\sum_{x \in Q} f_i(x)}{|g|}$. The architecture of the suggested approach for educational group recommendations is illustrated in Figure 13-1.

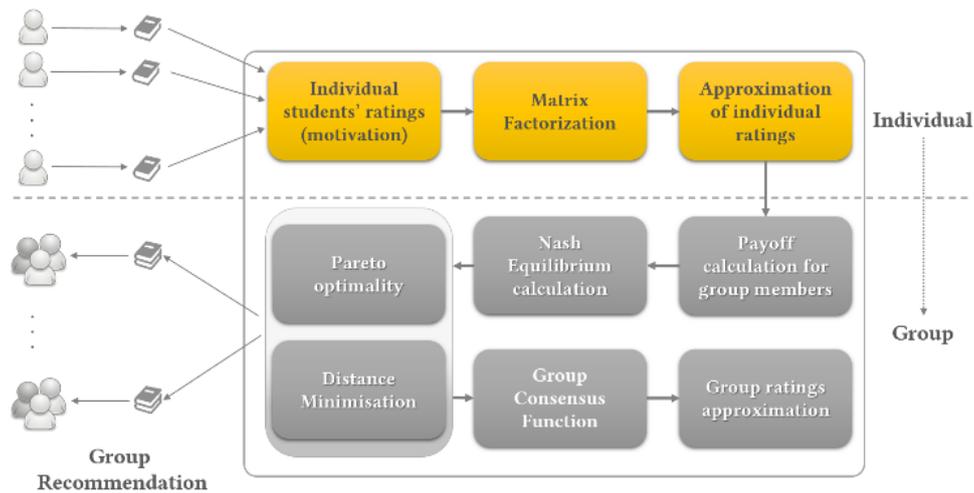


Figure 13-1. Architecture of the non-cooperative game-theoretic group recommender system for educational resources

13.3.2. Illustrative example

Consider a group with two members, i.e., the students A and B. Table 13-1 demonstrates the individual students' predicted motivation \hat{m}_{ij} to be excited from the educational resources r1, r2, r3, r4 and r5, after matrix factorization. None of the students A and B has previously seen or evaluated any of these five items.

Table 13-1. The individual students' predicted motivation from the educational resources

	R1	R2	R3	R4	R5
A	3.6	4.2	1.8	2.6	3.2
B	1.2	3.4	2.4	4.6	4.6

Table 13-2 illustrates the payoff (motivation) for each student from all the possible actions (strategies) taken by himself and the other group member.

Table 13-2. The payoff (motivation) for each student from all the possible actions (strategies)

		B				
		r1	r2	r3	r4	r5
A	r1	(3.6, 1.2)	(3.9, 2.3)	(2.7, 1.8)	(3.1, 2.9)	(3.4, 2.9)
	r2	(3.9, 2.3)	(4.2, 3.4)	(3.0, 2.9)	(3.4, 4.0)	(3.7, 4.0)
	r3	(2.7, 1.8)	(3.0, 2.9)	(1.8, 2.4)	(2.2, 3.5)	(2.5, 3.5)
	r4	(3.1, 2.9)	(3.4, 4.0)	(2.2, 3.5)	(2.6, 4.6)	(2.9, 4.6)
	r5	(3.4, 2.9)	(3.7, 4.0)	(2.5, 3.5)	(2.9, 4.6)	(3.2, 4.6)

From this table, it can be seen that there are two NE, the strategies profiles (r2, r4) and (r2, r5). The reason is that if student A chooses action r2, then student B has the same benefit from actions r4 and r5, and does not benefit in changing his action to r1 or r2 or r3. Likewise, considering that student B chooses action r4 or r5, then student A has no benefit to change the action from r2 to r1 or r3 or r4 or r5. Between these two strategies, (r2, r5) is Pareto optimal. This means that these two items (r2 and r5) should be recommended to the group members in order to optimize the motivation for each individual member, whereas, neither of the students can get more payoff (motivation) without decreasing the payoff of the other student, indicating an optimal solution for the group as a whole.

13.3.3. Measuring motivation excited from resources

Motivation is defined as “the process whereby goal-directed activity is instigated and sustained” (Pintrich & Schunk, 2002, p. 5). In the context of learning, motivation is “a student’s tendency to find academic activities meaningful and worthwhile and to try to derive academic benefits from them” (Brophy, 2004, p. 249). Many theories have been proposed to measure different motivational constructs (e.g., expectancy-value theory (Wigfield & Eccles, 2000), attribution theory of achievement motivation (Weiner, 1985), goal-orientation theory (Dweck, 1986), self-determination theory (Deci & Ryan, 2002)) and explain why it critically affects learning.

The definition of motivation highlights that motivation is a goal-oriented process (Pintrich & Schunk, 2002); it determines the students’ goal-setting process, affecting their choices and decisions, accordingly (Maehr & Meyer, 1997). Moreover, previous research results shown that when a system triggers the students’ intrinsic motivation (i.e., is perceived as challenging and satisfying (Ryan & Deci, 2000)), then, the students’ behavioral intention to use that system increases (e.g., Huang, 2017). In addition, (Davis, 1989) showed that perceived usefulness is an example of extrinsic motivation (i.e., students are doing an activity for its instrumental value (Ryan & Deci, 2000)) for intention to use information services. Furthermore, it was found that the clarity of educational content could affect learner perceptions of usefulness (Terzis & Economides, 2011), and as such, clarity could be indirectly considered as a factor that motivates the e-learner.

According to the above, and to keep the measurement of motivation simple, yet contextualized and coherent to the research, for assessing their appreciation to each item, the students had to rate (a) their own perceived clarity of each item, (b) how much each item fullfield their own learning goals, (c) their own perceived usefulness of each item and (d) their intention to use the item. For this purpose, four questions were delivered to them in a 5-point Likert-like scale (Table 13-3). The average score per student was considered as the student’s perceived motivation from the corresponding item.

Table 13-3. Questions for measuring motivation

Question
Clarity of item's content: <i>The item was clear and understandable</i>
Goal fulfilment from the item: <i>The item met my expectations and covered my learning goals</i>
Usefulness of the item: <i>The item helped me improve my learning</i>
Behavioral intention to use the item: <i>I indent to use the item in the future</i>

13.3.4. Motivation approximation: Matrix Factorization

By definition, in non-cooperative games, the students act rationally (i.e., they would select those items that would increase their own motivation and competences), and know that the other students act rationally as well. Moreover, in games, it is assumed that the students are aware of their own predicted motivation from following each available strategy, as well as of the predicted motivation of the other group members from their choices. In order to suggest items to the group members, this information should become available to the game-theoretic group recommender, to guide decision support.

As stated in sub-section 13.3.1, the predicted motivation \hat{m}_{ij} for student i from item j is approximated with the Matrix Factorization technique (Koren et al., 2009). The basic idea is to view the student-item motivation as a sparse matrix, for which we wish to predict the values of its empty cells, such that they would be consistent with the existing motivation values in the matrix. This is achieved by computing a low-rank approximation of the motivation matrix. As notational convention, bold small letters denote vectors, and bold capital letters denote matrices.

Let \mathbf{M} be the matrix of size $|L| \times |R|$ that contains the motivation that the students get from the items. Each student l_i is associated with an f -dimensional factor vector \mathbf{l}_i , and similarly each item r_j with an f -dimensional factor vector \mathbf{r}_j . To get the predicted (approximated) motivation from an item r_j for student l_i , the inner product of the corresponding factor vectors is computed: $\hat{m}_{ij} = \mathbf{l}_i^T \mathbf{r}_j$. The resulting dot product captures the student's l_i overall motivation from the item r_j , and models this interaction. The major challenge is then to compute the mapping of each item and each student to the factor vectors, \mathbf{r}_j , \mathbf{l}_i , so that they accurately estimate the known motivation values without over-fitting. The simplest approach to learn the factor vectors is to minimize the regularized squared error on the set of known motivation values:

$$\min \sum_{(l,r) \in K} (m_{ij} - \hat{m}_{ij})^2 + \lambda (\|\mathbf{l}_i\|^2 + \|\mathbf{r}_j\|^2),$$

where K is the set of (l_i, r_j) pairs for which m_{ij} is known. The constant λ controls the extent of regularization and is usually determined by cross-validation. To minimize this function and determine the factor vectors, Stochastic Gradient Descent (Bottou, 2010) can be applied.

13.3.5. Student grouping: Fuzzy C-Means

A central topic in group recommender systems is the partition of the users into a number of groups, i.e., the group formation problem. The existing methods in educational group

recommender systems promote shaping homogeneous groups of students (e.g., Kompan & Bielikova, 2016; Yanhui et al., 2015). However, as already stated in the motivation of the research section, having heterogeneous groups of students is considered as more beneficial in collaborative learning contexts, with respect to students' overall learning gain (Graf & Bekele, 2006; Slavin, 1987). Thus, both homogeneous and heterogeneous groups should be considered. Since the groups of students are not already known, clustering techniques are appropriate for detecting similarities (and dissimilarities) within the data.

A simple idea to end-up with homogeneous (heterogeneous) groups is to group the students based on the available individual ratings (the matrix containing the students' motivation from the items they have already rated), in such a way so that students with similar (dissimilar) ratings for the same items are in the same group. The Fuzzy C-Means (FCM) algorithm (the fuzzy version of the k-means algorithm) (Bezdek, 1981) was employed in this process. Unlike k-means, FCM allows data points to obtain fuzzy memberships to all clusters; in FCM a student may belong to more than one group with a different probability. FCM works efficiently even with small groups.

Let $L = \{l_i\}$ be the set of students (i.e., data points) and $C = \{c_j\}$ be the set of clusters centers. Each student l_i is associated with an f -dimensional factor vector \mathbf{l}_i , and similarly each centroid c_j with an f -dimensional factor vector \mathbf{c}_j . For every cluster a membership matrix \mathbf{U} is created to represent the membership probabilities for every student. If u_{ij} is the membership probability of l_i in the cluster j , and k is the fuzziness index ($k > 1$, $k \in \mathbb{R}$), the goal is to minimize (maximize) of the objective function: $J_k = \sum_{i=1}^n \sum_{j=1}^m (u_{ij})^k \|\mathbf{l}_i - \mathbf{c}_j\|^2$, where $\|\cdot\|$ is the Euclidean distance between the i^{th} student and the j^{th} cluster center.

However, FCM takes as input the desired number of clusters and not the number of students per cluster, resulting in significant inequality in the sizes of the clusters (groups of students). During the clustering process, data points with extreme values tend to isolate, resulting in significantly less members in some of the created clusters than others. In order to automatically reform the clusters to become of equal size (preserving homogeneity), FCM's probability matrix \mathbf{U} was utilized for exchanging members among the groups based on their highest probability of belonging to particular clusters, and according to a maximum number of members allowed in each cluster.

For the formation of mildly heterogeneous groups, after classifying students in homogeneous and heterogeneous groups, students can be randomly selected from different clusters, and re-grouped in dyads and (or) triads.

13.4. Experimental Evaluation

13.4.1. Participants and experimental setup

The proposed game-theoretic group-recommendation method was evaluated on a realistic setting with data from an empirical study with 102 students (55 girls [53.9%] and 47

boys [46.1%], aged 16 years old) at a European High School, in November 2017. The activity was about collaboratively writing simple functions in the Python programming language, using the recommended resources, as well as the lectures in the classroom.

The experiment was conducted in three phases. During the first phase, 168 educational resources (i.e., worked examples, solved exercises, etc.), designed to motivate students and increase their interest in Python, were randomly assigned to the individuals. Each student had to study at least 4, but not more than 6 items, within 2 days. After studying each item, the students had to rate it, assessing the overall motivation the item excited to the student, as explained in section 13.3.3. The resulting dataset consisted of $|L|=102$ students, $|R|=148$ items and $|M|=687$ student-item ratings.

For the needs of the second phase, the students were arranged into four general, equivalent groups: one treatment (T - 26 students) and three control groups ($C1, C2, C3$ - 26, 24, 26 students respectively). Each of these general groups was further partitioned in three types of sub-groups with the FCM clustering method (described in section 13.3.5), and with respect to their members' previous ratings: (a) homogeneous (Hom), (b) heterogeneous (Het), and (c) mildly heterogeneous (m. Het). For each general group, $|G|=9-10$ sub-groups were formed including 3 Hom., 3 m.Het. and 3 or 4 Het. (i.e., 39 sub-groups in total), with $|g|$ varying from 2 to 3 students per sub-group.

Next, one (or more) item(s) were delivered to each sub-group regularly (every two days) for two weeks, according to a recommendation strategy: the suggested game-theoretic method (GT) was applied on the sub-groups of T , the Average method (AVG) was used on the sub-groups of $C1$, the recommendations to the sub-groups of $C2$ were generated with Least Misery (LM), and the Most Pleasure method (MP) provided the recommendations to sub-groups of $C3$. AVG , LM and MP are briefly demonstrated in (Masthoff, 2011), and a short description is in the following subsection (i.e., 4.2.2). Table 13-4 summarizes the distribution of the students in groups and sub-groups during the second phase of the activity, and maps the recommendation method to the corresponding group.

Table 13-4. Description of Groups in the Second Phase

GroupID	No. of students	Hom.	m. Het.	Het.	Rec. Method
T	26	3(x3)	3(x3)	4(x2)	GT
$C1$	26	3(x3)	3(x3)	4(x2)	AVG
$C2$	24	3(x3)	3(x3)	3(x2)	LM
$C3$	26	3(x3)	3(x3)	4(x2)	MP

After studying the recommended items for two days, all group members had to rate them both individually, and as a team, using the same questionnaires as in the previous phase. Every second day the \mathbf{M} matrix, containing the real individual ratings, was updated. Another matrix, \mathbf{V} , containing the actual group ratings on the items was also constructed and updated. At the end of the second week, the student-item ratings were $|M|=1464$, and the group-item ratings were

$|M^G|=232$. It should be noted that all sub-groups of the control groups received one item per recommendation cycle, whereas the sub-groups of the treatment group received up-to-three items per cycle. Throughout the experimental process, the items recommended to each sub-group should not have been previously rated by any of the sub-group members.

The third phase of the activity was about collaboratively writing simple functions in Python, using the resources recommended during the previous two phases, as well as their knowledge gained during the course lectures, within one week time. The activity cycle is illustrated and synopsized in Figure 13-2.

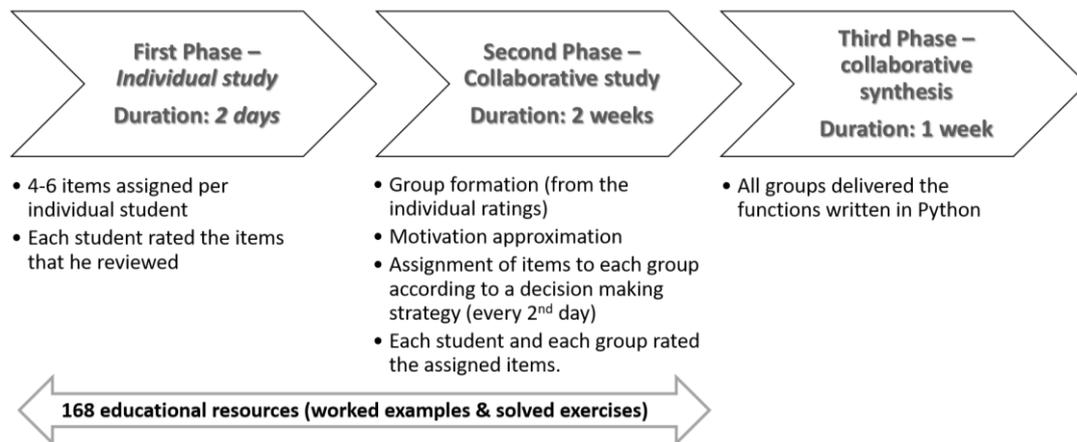


Figure 13-2. The experimental activity process

Throughout the activity, the recommended items were delivered to the participants via a configured version of the Learning Analytics and Educational Recommender System (LAERS) environment (Appendix A). During all three phases, the items were available to the students to open and review, but not to download (in order to force the students' persistence); the students' time-spent on viewing the recommended items, as well as the respective frequencies of reviewing them were being tracked (to measure groups' persistence). The students could assess the motivation that each item excited to them (either individually or as a team) only once, at the end of the respective recommendation cycle, when they were prompted to do so, but could not reconsider or modify their evaluation from that point on.

Finally, the collaborative problem-based learning activity was assessed by the course instructor according to the completeness and correctness of the final deliverable (the source code was produced with the Python 3.4.4 IDLE and submitted as a .py file), and was graded in a scale of [1, 20] as well.

13.4.2. Methods and Evaluation metrics

13.4.2.1. Group decision strategies.

As stated in the previous sub-section, for each one of the control groups (i.e., $C1$, $C2$, $C3$), the expected group motivation from an item was provided by a different group decision (aggregation) method, formulating how the corresponding sub-groups of students reach to a consensus and come up with a decision about that particular item. Let k be the number of students

in a group, \hat{m}_{ij} the predicted motivation of student i from item j , the group motivation ratings were assigned according to the following strategies:

C1 – Average (AVG): A consensus-based approach, where all group members jointly and equally make a decision. The group motivation equals the average motivation ratings across the

group members: $M(k, j) = \frac{\sum_{i \in g} \hat{m}_{ij}}{k}$. In simple terms, *AVG* sets the average rating given by the group members to each item as the predicted rating of target group, and selects as recommendations those items that achieved the highest predicted ratings.

C2 – Least-Misery (LM): A borderline approach that targets to please the least happy member of the group, resulting the group to behave under a least-misery principle. In this case, the group motivation equals the minimum motivation among all group members: $M(k, j) = \min_{i \in g} \hat{m}_{ij}$.

In other words, *LM* considers the rating of each item, assumes that the group's predicted rating on each item is the lowest value from the ratings given by all group members, and recommends these items. Thus, a group is as motivated as its least motivated member.

C3 – Most-Pleasure (MP): Another borderline group decision strategy satisfying the highest rating within the group. The motivation a group of students gets from an item equals to the maximum motivation within the group: $M(k, j) = \max_{i \in g} \hat{m}_{ij}$. Similarly to *LM*, *MP* takes under consideration the ratings of each item and next recommends the item with the maximum motivation among all group members. Thus, a group is as motivated as its most motivated member.

All solutions were implemented in MATLAB. Furthermore, the Gambit tool (Mckelvey, McLennan, & Turocy, 2006) was used to verify the correct identification of Nash Equilibria.

13.4.2.2. Evaluation measures

Our proposed method targets at solving conflicts of interest by minimizing the prediction error of group motivation from the recommended educational resources (items). In the context of prediction accuracy estimation, the *Root Mean Square Error (RMSE)* is generally accepted as a good measure of precision, commonly used as an evaluation metric to compare prediction errors of different models for the same data. It measures the sample standard deviation of the difference between values approximated by an estimator and the values actually observed (Hyndman & Koehler, 2006). In our study, we explore the precision of our prediction with respect to motivation from the recommended items, as it is actually rated by *each* student, and by a given *group* of

students. *RMSE* is computed as: $RMSE = \sqrt{\frac{\sum_{j=1}^n (m_{kj} - \hat{m}_{kj})^2}{n}}$, where n is the number of items rated.

Lower values indicate better predictions, and consequently, better decision strategy.

We also used the maximum *RMSE* for capturing the robustness of the recommender system, as it corresponds to the worst-case accuracy across *any* group. Lower *mRMSE* values indicate that all groups will receive good recommendations. This measure is computed as:

$$mRMSE = \max \sqrt{\frac{\sum_{j=1}^n (m_{kj} - \hat{m}_{kj})^2}{n}}$$

Furthermore, to measure the quality of the ranked list of recommended items delivered to groups of students, i.e., to evaluate its goodness, we used a measure from Information Retrieval, specifically crafted for ranking: the *Normalized Discounted Cumulative Gain (nDCG)* which assumes multiple levels of relevance (Järvelin & Kekäläinen, 2002). In simple terms, *Discounted Cumulative Gain (DCG)* measures the gain of an item (i.e., the relevance score – if rating is missing, zero value is set) based on its position in the resulting list. The gain from the list is accumulated from top to bottom, and more relevant items are preferable to be on the top of the list (i.e., in our case, the most motivating). Thus, prior to accumulation, the scores are divided by the logarithm of the item's position, leading to a discount. *DCG* for a group of *k* students at position *N* (length of recommendation list), is computed as:

$$DCG_k @ N = m_{kj_i} + \sum_{i=2}^N \frac{m_{kj_i}}{\log(i+1)}$$

However, comparing *DCGs* between groups of students is not valid. As such, normalized *DCG (nDCG)* values are computed by arranging all items in an ideal order, and next dividing *DCG* by the ideal one (*IDCG*).

Accordingly, *nDCG* is defined as: $nDCG_k @ N = \frac{DCG_k @ N}{IDCG_k @ N}$, where *IDCG* is the maximum possible

DCG, and *nDCGk@N* getting values between 0 and 1, with 0 indicating the worst ranking and 1 representing the ideal ranking of items. In our study, due to limitations in available educational resources to be used as the recommendation items set, we only used short lists of up-to five items per group. Thus, we calculated *nDCG* with *N=3* and *N=5*. We compared the effectiveness of both the *group* and *individual* recommendations when varying the aggregation method.

Finally, we measured the diversity of recommendation lists between different groups, by employing the *Hamming Distance (HD)* (Zhou et al., 2008) metric. *HD* estimates if the recommendations to all groups make full use of all items, leaving only a few items without being recommended. If Q_{g,g^*} is the “overlapped” number of items recommended to both groups *g* and *g**

respectively, then the *HD* between group *g* and group *g**, is defined as: $HD(g, g^*) = 1 - \frac{Q_{g,g^*}}{|z|}$,

where *z* is the length of the recommendation list. High *HD* means high diversity, making full use of all items and leaving out of recommendation only a few items; a highly personalized recommendation list should have higher *HD* to other lists.

13.5. Results

13.5.1. Prediction accuracy, effectiveness and diversity of recommendations and students' persistence for groups

Tables 13-5(a), 13-5(b) and 13-5(c) demonstrate the results for the evaluation metrics (average values) for all decision support strategies compared in this study, i.e., the currently proposed game-theoretic method (*GT*) applied on the treatment group, and the Average (*AVG*), Least-Misery (*LM*), and Most-Pleasure (*MP*) methods applied on each one of the control groups, for homogeneous (high inner member similarity), mildly heterogeneous (medium inner member similarity), as well as heterogeneous (low inner member similarity) synthesis of the sub-groups respectively. The sub-groups sizes was firm, varying from two to three students, as explained in section 13.4.1.

Table 13-5. (a) Metrics for Homogeneous Groups

	RMSE	mRMSE	nDCG@3	nDCG@5	HD
GT	0.342	0.475	0.954	0.961	0.871
AVG	0.351	0.448	0.959	0.958	0.834
LM	0.386	0.647	0.883	0.884	0.715
MP	0.392	0.724	0.877	0.875	0.686

(b) Metrics for Mildly Heterogeneous Groups

	RMSE	mRMSE	nDCG@3	nDCG@5	HD
GT	0.366	0.515	0.949	0.950	0.843
AVG	0.564	0.738	0.912	0.913	0.754
LM	0.785	1.233	0.806	0.805	0.622
MP	0.800	1.839	0.774	0.772	0.604

(c) Metrics for Heterogeneous Groups

	RMSE	mRMSE	nDCG@3	nDCG@5	HD
GT	0.416	0.524	0.942	0.940	0.832
AVG	0.773	0.958	0.852	0.851	0.674
LM	1.092	1.645	0.716	0.716	0.504
MP	1.521	2.132	0.667	0.668	0.498

According to these results, all decision support methods achieve low approximation error in prediction of motivation ratings for the homogeneous students' groups, as expected. On the contrary, for mildly heterogeneous groups, accuracy is high for the *GT* and satisfactory for the *AVG* method, but the prediction error significantly increases when the aggregation strategy is *LM* or *MP*. For heterogeneous groups, the approximation error in prediction of motivation is low only when the recommendation strategy is the *GT*, whereas it is high for all the other cases.

Furthermore, the group recommendation effectiveness tends to decrease only for the heterogeneous sub-groups. Figure 13-3 illustrates the average goodness of the ranked list of recommended items delivered to the sub-groups of students when the top ranked items are 5 (*nDCG@5*) and when the top ranked items are 3 (*nDCG@3*), according to the inner similarity of

the sub-groups, and by considering the decision support strategy.

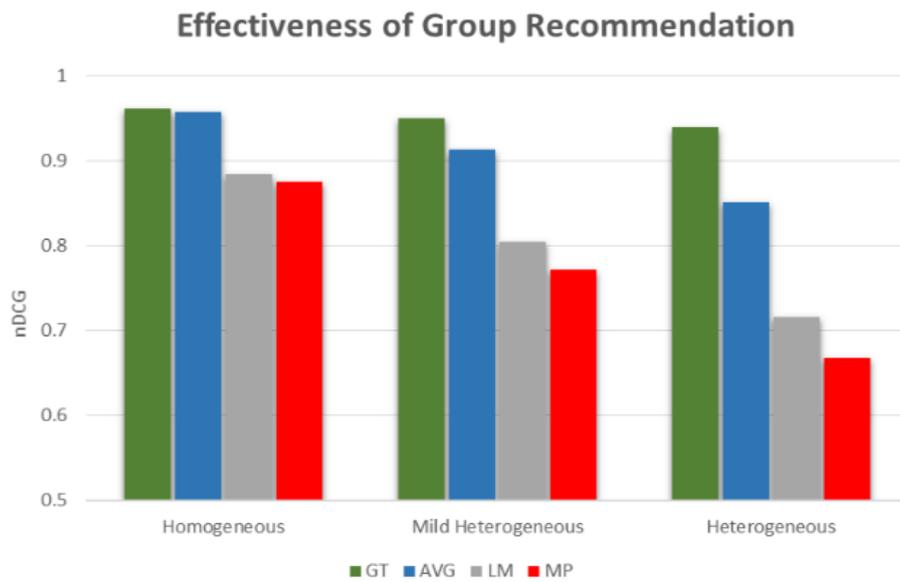


Figure 13-3. (average) Effectiveness of group recommendations with respect to the aggregation strategy and the group inner similarity.

Diversity in recommendations, reflected in the HD values, indicates that the recommendation to all groups – the homogeneous and the (mildly) heterogeneous – make sufficient use of all items and few items will be left without being recommended, when the recommendation method is the proposed GT method. Figure 13-4 illustrates the diversity in recommendation according to the different decision support strategy for either the homogeneous or the (mildly) heterogeneous groups.

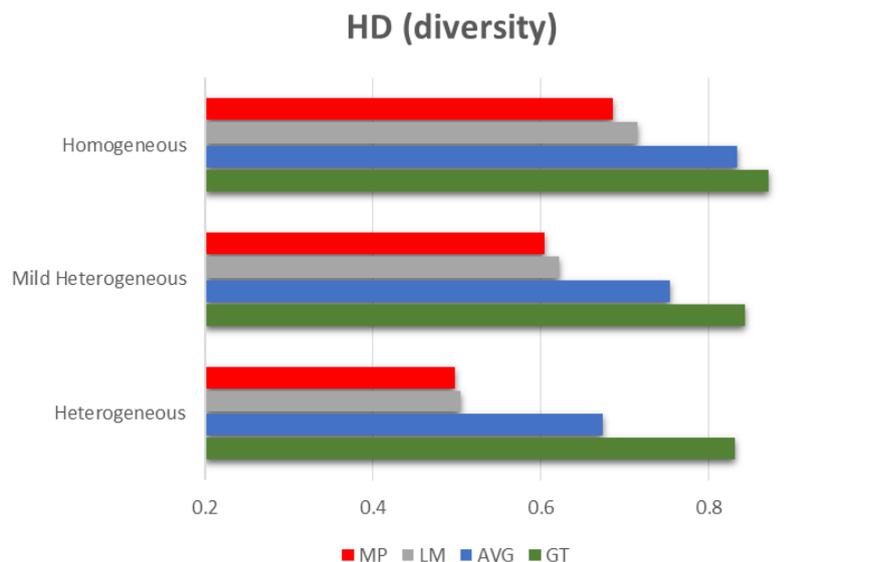


Figure 13-4. Diversity of group recommendations with respect to the aggregation strategy and the group inner similarity.

Finally, Figure 13-5 illustrates the groups' average time-spent on reviewing the items (based on the total time-spent to review the items and the respective frequencies of reviewing)

throughout the second and third phase of the activity, for all decision support strategies compared in this study. As stated in section 13.4.1, this measure is indicative of students' actual engagement with the items, and codes their persistence.

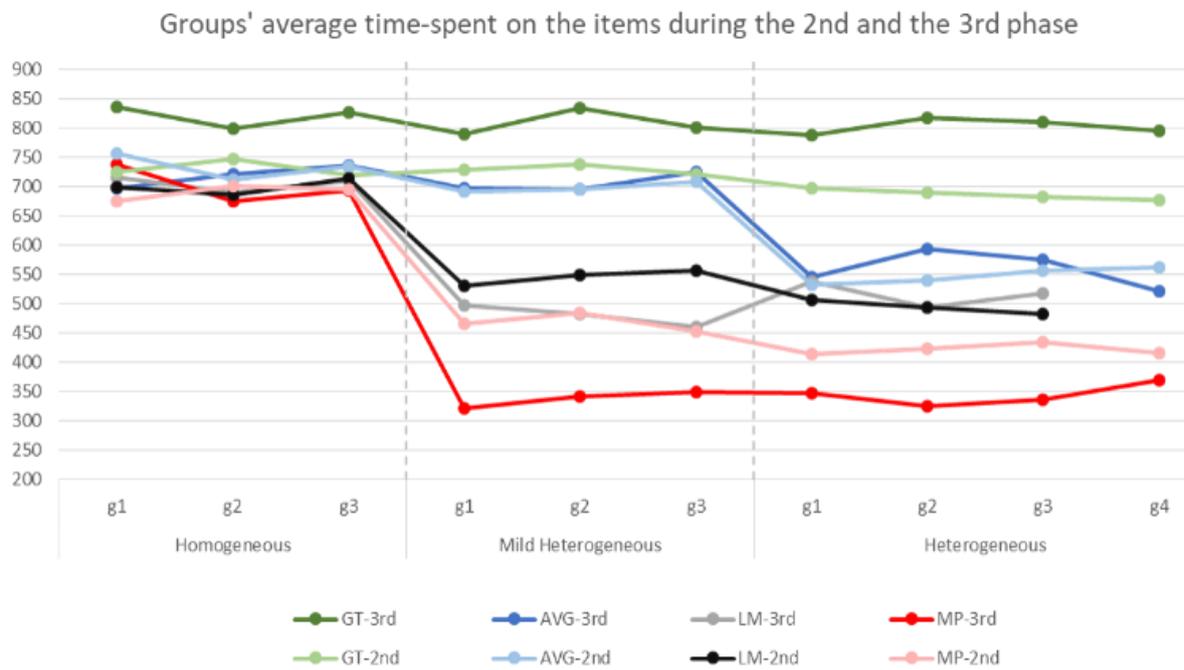


Figure 13-5. Groups' actual engagement with the items during the collaborative phases of the activity, with respect to the aggregation strategy and the group similarity.

13.5.2. Effectiveness of recommendations and students' conformity for individual students within groups

Furthermore, in order to understand *when* the group recommendations are better or worse (ranked) for each individual within the sub-groups, we measured the difference between the effectiveness of the individual and the group recommendations' lists. This difference is indicative of the individuals' degree of conformity, regarding the adjustment of their motivation from the recommendation with respect to the motivation of the group they are members of. A positive difference means that the group recommendations are better ranked than the individual recommendations. Figure 13-6 shows a scatter plot where each student, in a group, is represented by a point, for the two better performing methods, i.e., the GT and the AVG method. Here, the x axis measures nDCG@3 for the individual recommendation list, while the y axis shows the distance of the individual's from the respective group's nDCG of this group recommendation list for the same student. In this figure, the green trendline corresponds to the GT method, whereas the red trendline corresponds to the AVG method, respectively.

Please, note that during the two weeks of experimentation, each student and each group received recommendations every second day, resulting to a total of more than one recommendations, and hence a student may be represented by several points.

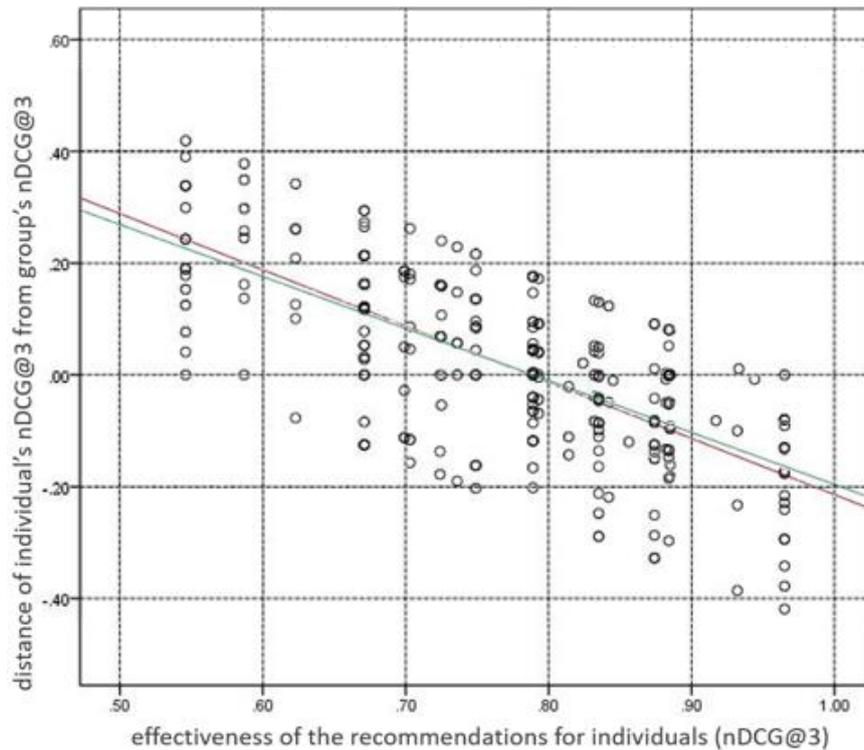


Figure 13-6. Distance of group motivation with respect to individual motivation.

13.5.3. Effect of recommendations on group learning performance

Regarding the effect of the recommendation strategy on the groups' performance in the collaborative learning activity, as it is reflected on their grades, *independent sample t-tests* were conducted and the *Cohen's d effect size* was calculated. Results indicate that taking the recommended resources has a differentiated impact on performance depending on the groups' achievement level in the collaborative activity. Table 13-6 compares the average learning performance for all four recommendation strategies for all types of group inner homogeneity level. Finally, Figure 13-7 plots the mean scores per group inner homogeneity type per recommendation strategy.

Table 13-6. Effect of Recommendation strategy on Performance

	Mean (SD)	Mean (SD)	Mean Diff.	t-value	p-value	Cohen's d
GT - AVG		15.20 (2.150)	1.80	2.377*	0.029	1.063
GT - LM	17.00 (1.054)	14.78 (1.922)	2.22	3.171*	0.006	1.432
GT - MP		13.90 (1.524)	3.10	5.291*	0.000	2.365
AVG - LM		14.78 (1.922)	0.42	0.449	0.659	0.205
AVG - MP	15.20 (2.150)	13.90 (1.524)	1.30	0.208	0.136	0.697
LM - MP	14.78 (1.922)	13.90 (1.524)	0.88	1.109	0.283	0.507

*p<0.05

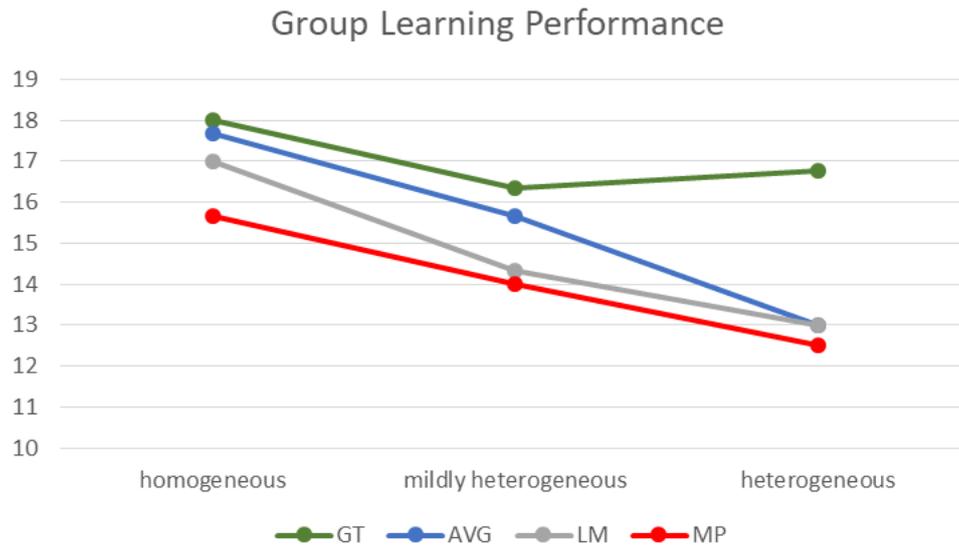


Figure 13-7. Means of learning performance per recommendation strategy per group homogeneity type.

13.6. Discussion

Recommending educational resources to groups of students, targeting at optimizing all students' motivation, both individually and as a group, is a complicated task. The core issue is to determine how a group of students reaches to a consensus about their degree of appreciation for each item, in such a way that reflects the self-determined interests and motivation of *each and all* group members. This study focuses on addressing diversity in group members' self-motivated considerations regarding the usefulness of the resources, as well as in their behavioral intention to finally use the resources (conflict of interest).

The review of related research identified four types of drawbacks related to: (a) the promotion of high group inner similarity, (b) the effectiveness of aggregation strategies, (c) the number of recommended resources, and (d) the omission of the individual members' conformity degree (Dwivedi & Bharadwaj, 2015; Kompan & Bielikova, 2016; Rodríguez et al., 2016; Yanhui et al., 2015; Zapata et al., 2015). Moreover, none of these approaches – to the best of our knowledge – explored the actual engagement of the group members with the recommended items, in a learning analytics fashion, explaining the students' persistence, and reasoning the final learning performance in the collaborative activity. As such, the emerging research questions were:

RQ1: *Can we accurately and efficiently recommend sequences of educational resources to homogeneous and heterogeneous groups of students, with respect to both the individuals' and the group's motivation, and intention to use the resources?*

RQ2: *What is the impact of a recommendation on individual students' persistence as well as on the groups' learning performance in the collaborative problem-solving activity?*

In order to address the abovementioned issues and answer the research questions, this study suggested and evaluated a non-cooperative game-theoretic perspective for solving conflict of interest between the group members and guiding the recommendation process.

For the empirical evaluation of the approach in a realistic setting, data were collected during a collaborative problem-oriented learning activity with 102 students from a European High School. The objectives were twofold: (a) to compare the accuracy and the effectiveness of ranked lists of recommended items delivered to groups of students by the suggested method to other state-of-the-art decision support methods, with respect to the individual's motivation from the recommended items, and (b) to explore the impact of recommendation on individual members' level of conformity, as well as on the overall groups' persistence and learning performance. The following novel facts and important observations have arisen.

13.6.1. Accuracy and effectiveness of the recommendation for different levels of group inner similarity

Firstly, from table 13-5 it becomes apparent that all decision support methods achieve low approximation error in prediction of motivation ratings for the homogeneous students' sub-groups. However, the proposed game-theoretic strategy minimizes the prediction error of the sub-group motivation ratings, as, by far, it scores the lowest RMSE values for all categories of inner sub-group similarity. Especially for the highly heterogeneous sub-groups, the other aggregation methods combine potentially conflicting rankings that could create a group recommendation which might not be motivating for the group members; accuracy is high for the GT method ($RMSE=0.416$), but the prediction error significantly increases when the aggregation strategy is AVG ($RMSE=0.773$), LM ($RMSE=1.092$) or MP ($RMSE=1.521$). As such, in this case, the GT decision strategy resolves sufficiently the conflict of interest and delivers the most appropriate items to the students.

Still, mRMSE demonstrates some variance in the prediction error across sub-groups. In particular, for the homogeneous sub-groups, the GT method did not have the lowest prediction error ($mRMSE=0.475$); in this case, it turns out that the AVG strategy was a better approach, although only slightly ($mRMSE=0.448$). Yet, this was an expected finding, since the average method works well in most cases of homogeneous groups (Kompan & Bielikova, 2016; Rodríguez et al., 2016; Yanhui et al., 2015).

Secondly, from the same table of results, one can observe that the suggested GT method has a good overall performance (i.e., the nDCG values reflecting the effectiveness of ranked list or recommendations), although not always the best; in one case of homogeneous groups, the AVG method provided slightly more effective recommendations ($nDCG@3=0.959$) compared to the list of GT ($nDCG@3=0.954$). However, it is important to notice that, compared to the other methods, the performance of the proposed GT seems to be stable and robust, regardless of the inner sub-group similarity, targeting ranking quality and demonstrating only small variations. From the evaluation results it was found that nDCG for the GT method is close to 1.0 (higher than 0.9) in all cases of sub-group homogeneity, whereas the respective values for the other methods decrease as the inner group similarity decreases.

Thirdly, another finding concerns the plurality of the recommended lists of items (diversity of items). It is very possible that the deficiency in capturing the whole range of students' needs and preferences could lead to poor motivation. Improving the diversity of recommendation results is expected to increase within group motivation from the recommended resources. As seen in Table 13-5, the values of the Hamming Distance (HD) metric reflect that the personalization of recommendation is better for homogeneous groups, regardless of the method employed, whereas, the GT method provides satisfactory personalization even for heterogeneous groups ($HD=0.832$) compared to the other methods (HD varying from 0.498 to 0.674). In addition, this finding implies that the GT method makes sufficient use of all items and only few items will be left without being recommended.

13.6.2. Impact of recommendation on conformity, persistence and learning performance

The second research question concerned the impact of the recommendation on the individuals' gain as well as on the overall groups' performance. For this purpose, three indices were employed: (a) the individuals' conformity degree, expressed as the distance between the individual nDCG and the respective group's nDCG, (b) the actual aggregated time-spent on the recommended items and the frequencies of reviewing them, as measures of students' persistence, in a learning analytics fashion, and (c) the groups' grades (scores) in the collaborative activity.

Firstly, in order to understand how much the individuals adjusted their personal evaluation of motivation from the recommended items compared to the group's they belong to (conformity degree), we plotted the best fit line through the points in the scatter plot (in Fig. 13-6). One can observe that when the employed method is the GT, the gradient of the curve is smaller compared to the next best performing method, i.e., the AVG. This means that the recommendations are highly ranked for the individuals as well as for the group. In other words, the individuals within the groups don't have to highly adjust their personal consideration about their motivation gained from the recommended items, to be approved by the other group members. The measure employed, i.e., the distance between the individual nDCG and the respective group's nDCG, is a topic that deserves further analysis and could be explored as a measure of the individual's degree of conformity.

Secondly, as seen from the curves in Fig. 13-5, the students' engagement with the resources during both collaborative phases of the activity, and regardless of the inner group similarity, was stable and high when the recommendations were generated with the GT method. The same fact is also true for the mildly heterogeneous groups when the recommendation method is AVG, as well. When the inner group member similarity is high (homogeneous groups), all methods deliver items that motivate students to actually participate in the problem-oriented collaborative activity. Yet, for heterogeneous groups, the students' disengagement from the resources successively increases for the AVG, LM and MP methods. The disengagement is even

higher for these methods during the third phase, when the groups had to complete the collaborative assignment by using the recommended resources. This fact is illustrated on the average time-spent on the resources and on the frequency of reviewing them to complete the task.

Finally, the independent samples t-tests for comparing the differences in groups' final grades with respect to the recommendation method verified that the resources delivered to the groups affected their learning performance, as expected.

13.7. Conclusions

In this study, we aimed at recommending educational resources to groups of students with diverse levels of inner member similarity. The goal was to decide upon those resources that would best support *each* and *all* group members to efficiently complete the assigned collaborative activity.

Inspired from Carvalho & Macedo (2013), we argue that *Game Theory* could efficiently solve "conflicts of interest" between the group members (Nash, 1951) and guide the recommendation of educational resources. Game theory is about social situations, providing solid recommendations to the players regarding their own optimal strategy, as well as administering an external observer that predicts the outcome of interactions (i.e., in our approach, the decision support system).

The proposed solution models the recommendation strategy as a problem of finding the Nash Equilibrium, i.e., a state in which no student can be benefited more in terms of further improving their own motivation by unilaterally deviating from the Nash Equilibrium. However, the best collective result does not always come from each individuals following their own interest, but rather from reaching the group's consensus; whereas a Nash Equilibrium does not correspond to a socially optimal outcome, a Pareto optimal equilibrium describes a social optimum in the sense that no individual player can improve their payoff without making at least one other player worse off. Pareto optimality is not a solution concept, but is used to evaluate the overall gain.

To this end, we developed a mathematical formulation (i.e., Algorithm1) for group-recommendation of educational resources as a non-cooperative game, as well as an architecture for building such group recommender systems (i.e., Fig. 13-1). From the evaluation of the approach with a realistic dataset, results revealed a tendency that the accuracy of the predicted group motivation, the goodness of the ranked list of recommendations, and the problem-solving performance for the treatment group were significantly higher compared to other state-of-the-art methods. The diversity of the items in the recommendation, indicating a satisfactory personalization even for heterogeneous groups, was higher with the GT method, as well.

We also explicitly compared the differences in individual evaluation of motivation from the recommendation, to the group's perception; we introduced the distance between the individual nDCG and the respective group's nDCG, as a measure of students' degree of conformity.

Additional research is required. Moreover, we adopted a learning analytics view point to explore the groups' actual engagement with the recommended resources and evaluated the effect of their persistence on the final learning performance. It was confirmed that the more motivating the recommended items, and the higher the students' persistence, the better the learning outcome.

However, there are some limitations. Firstly, the samples of the 168 educational resources and 102 students considered in the evaluation process are small and potentially biased; bigger datasets should be analyzed. Secondly, we investigated only groups of two to three students; the behavior of GT with larger groups of students (e.g., 4 to 5 members) should be explored as well. Lastly, we assumed that the group formation method used in this study would not raise issues of uncertainty; other methods for group formation should be explored as well.

Furthermore, a number of challenges for future work has emerged. For example, more sophisticated measures of motivation and persistence could be applied (e.g., incorporating the students' affective states, perceived enjoyment, challenge). The learning analytics research could contribute towards this direction. Yet, another challenging issue is focusing on the transparency of the group recommendation: showing each individual's payoff and eventually, how motivated the other group members are, could improve the particular student's understanding of the recommendation process, and perhaps make it easier to accept the educational resources that initially he/she did not like.

To conclude, the core contribution of this study is that the proposed game-theoretic solution demonstrates a socially and individually optimum group recommendation method, based on the motivational effect of the resources and the students' intention to use them, and yields statistically significant results even for highly heterogeneous groups of students.

Chapter 14 : Conclusions and future directions

“No book can ever be finished. While working on it we learn just enough to find it immature the moment we turn away from it.”

Karl R. Popper

Overall Discussion of contributions, implications and future research directions

The overarching idea of this thesis was to use data collected by learning environments to provide a better understanding on how learners’ develop their autonomous learning capacity in online self-assessment conditions, that allow for autonomous choices. In this thesis, four stages of analytics development were demonstrated, and their appropriateness was empirically assessed through a series of evaluation studies. In this chapter, the main findings and contributions of the presented work are briefly summarized. We elaborate on these findings with respect to the research objectives and research questions, as they were stated in Section 1.3 of the present thesis. A discussion regarding their implications for future theory, research and practice concludes this dissertation.

14.1. Addressing the research questions and elaborating on the findings – Impact of the research

Research Question 1: *(a) which learning analytics factors explain sufficiently the multiple aspects of learners’ interactions with the assessment tasks? (b) How can we model the cause-effect relationships between these factors towards explaining the variance in the learners’ performance?*

Overview: During the first stage of the learning analytics development, we sought for the factors that explain learning performance in self-assessment procedures. Two studies were conducted and three different data-analysis approaches were employed for addressing the research question. In both cases, Temporal and Behavioral learning/learner data were collected, either as tracked interactions or as self-reported perceptions, and analytics measures were constructed and evaluated. The decision on which data to gather and explore was driven by previous studies that identified significant determinants of learners’ achievement in self-assessment procedures.

Study1 – Applying Machine Learning (Supervised Classification) on Temporal and Behavioral Data for explaining learning achievement.

Findings: The results indicated that the time to answer correctly and the time to answer wrongly in combination with the goal expectancy could satisfactorily be used for classification of learners in computer-based testing procedures. The low misclassification rates are indicative of the accuracy of the proposed method. Further to that, the ensemble learning method provided the most accurate classification results compared to the other methods. These findings verified formerly reported results regarding the appropriateness of response-times to be used for

prediction purposes. Moreover, these findings confirmed the outcome of previous studies that employed the same analytics parameters, but analysed the data using variance-based methods.

Contributions & Implications: Discriminating response-time according to the correctness of the answer instead of using it as a single parameter has the potential to discriminate the learners' achievement level. When combined with the learners' goal expectations (i.e., how prepared they believe they are, and which are your achievement goals), the two distinct parameters of response-time can better justify the assessment outcome. These factors (learner features) can next be incorporated in the learner profile to model the learner's proficiency, and how it changes (varies) over time (*practical implication*). However, an interesting finding that requires more investigation was that most algorithms perform worse when two additional features were included in the analysis (*implication for research*).

Study2 – Applying fuzzy set quantitative comparative analysis (fsQCA) on Temporal and Behavioral Data for explaining learning achievement.

Findings: The results revealed five solutions that explain high achievement and five solutions that explain medium/low achievement, with respect to the different combinations of the considered analytics (response-time, self-regulation and satisfaction from content). The significant role of the distinctive response-times and goal-orientation was confirmed once again. An interesting finding highlighted the role of time-management in achieving high performance. To the best of our knowledge, this was the first study to explore the combinational/conditional interrelationship between time-management and content comprehensibility, and extends previous work, which focuses solely on the effect of content on performance. Another interesting finding was that all of the developed analytics parameters had strong participation in at least four of the identified solutions. This finding further confirmed the appropriateness of the extracted analytics measures to sufficiently explain learning performance.

Contributions & Implications: The approach offered an alternative view on how response-times, self-regulation and satisfaction from content combine to predict high and medium/low performance in self-assessment tests, complementing previous research in the area. It differs from previous research by performing configuration analysis and examining asymmetric relationships among the factors: multiple configurations between the same factors revealed different solutions to justify high or medium/low performance. These solutions were not possible to be identified by the commonly employed variance-based techniques that detect symmetrical relations only (*implication for research*). This study showcased how learning analytics researchers can utilize fsQCA to make sense of diverse analytics and take design decisions for various user groups (*methodological implication*). When applied together with complexity theory and configuration theory, fsQCA can contribute to the creation of new hypotheses, models, and theories (*theoretical implication*).

Generalized study – Analyzing Temporal and Behavioral data from Computer-based fixed/adaptive self-assessment using Partial Least Squares–Structural Equation Modeling

Findings: The analysis revealed that the learners’ manipulations of the self-assessment items and their achievements are rationalized in terms of response-times, effort (guessing), time-allocation according to the items’ difficulty, motivation and satisfaction. The use of these analytics factors contributed to reasoning learners’ performance in online self-assessment tests, either fixed or adaptive, ($R^2 > 0.68$), and consolidate the premise of “testing analytics”. Another interesting finding was that the features that worsened the classification performance in the first study were explored again, from a different perspective, and while the effect of effort was found statistically significant in this cases, the contribution of level of certainty was statistically insignificant.

Contributions & Implications: The developed causal and measurement models diagnosed satisfactorily the obtained self-assessment score and contributes sufficiently to revealing the most informative factors that rationalize learners’ actions in online summative fixed and adaptive self-assessment testing conditions. The set of developed analytics includes measures that were validated in multiple procedures with multiple analysis techniques. The extracted model demonstrates the cause-effect relationship between these analytics and provides the opportunity to make-sense about the learners’ activity, accordingly (**theoretical implication**). Based on the findings, testing analytics could be infused into a self-assessment system in order to facilitate the interpretation of the achieved scores (**practical implication**). These analytics shall next be used to measure learners’ behavior, towards understanding how they develop their autonomous capacity (**practical implication**). Building analytics-enhanced systems that would (a) allow learners to monitor their own progress, (b) help them evaluate and adjust their strategies to increase goal achievement and, (c) provide the autonomy to choose between appropriately suggested items to solve next, we can train the learners to self-regulate and determine a unique, personal “path” to successfully accomplish their goals.

Research Question 2:

How feasible is it to maintain “currently-aware” learner models in real-time, by refitting their parameters in run-time, and how accurately can these models approximate the learners’ next cognitive states from the current ones?

Overview: During the second stage of the learning analytics development, the previously identified factors were infused into the learner models and were combined with additional learner/learning factors to holistically describe the learners and to better and timely support them. Two studies were conducted and three different data analysis approaches were employed. Two generic modeling approaches were explored: constructing the models after the learners have completed the self-assessment, in a batch processing fashion, and constructing the models in real-time, granting “current-awareness”.

Study 1 – Considering Personality traits with Temporal and Behavioral factors and shaping the Learner Models with Supervised Classification Techniques.

Findings: The results of the study indicated a positive effect of extraversion and agreeableness on goal-expectancy, a positive effect of conscientiousness on both goal-expectancy and level of certainty, and a negative effect of neuroticism and openness on level of certainty. Further, extraversion, agreeableness and conscientiousness have statistically significant indirect impact on students' response-times and level of achievement. Like in a previous study, during the previous stage of analytics development, the ensemble RandomForest method provided accurate classification results (>80% accuracy), indicating that a time-spent driven description of students' behavior could have added value towards dynamically reshaping the respective models.

Contributions & Implications: The findings confirmed and complied with previous research results that suggested the use of time-oriented factors for enhancing the learner models. This study also showcased how the inconclusive findings about the role of certainty – measured with analytics in terms of time-spent – on explaining performance can turn into meaningful guidelines when this factor is associated with the learners' personality traits. A contribution of this study was the detection of temporal and behavioural patterns according to the learners' personality traits, with respect to their achievement level. Moreover, this study goes one step beyond by introducing the characteristics of each one of the five identified classes. The overall findings that associate the learners' personality traits with temporal and behavioural analytics, open the discussion of considering other factors, beyond cognitive, behavioral or motivational, to understand the learners' achievement during online self-assessment activities (*theoretical implications*). Specifically, the findings open a window towards adopting a learning analytics perspective for exploring the affective factors that are associated with personality, and which have been found to influence learners' actions in these conditions (*implications for research*).

Study 2 – Enhancing the Learner Models with Temporal Dynamics and updating them using Adaptive Data-Stream Classification.

Findings: Exploiting the analytics from the first stage of their development, this study shown that the non-cognitive traits (i.e., goal-expectations and self-efficacy) used for initializing the models provided sufficient information about the classes that the learners should be assigned to. Moreover, additional analytics were developed and explored. During the evaluation of these analytics, it was shown that the introduced learners' time-spent according to the items' difficulty in combination with their evolving mastery of skill/knowledge, the aggregated response-times on answering the self-assessment items, their time-varying level of certainty and their effort could efficiently be used for classifying learners in real-time (average HAT accuracy > 82%). In addition, multiple cognitive skills were jointly represented in a single learner profile, and the progress gained for each of them was accurately estimated. An important finding was that the changes in the distributions of the values of the time-varying features caused class drifting and

indicated changes in learners' progress; as such, class drift detection asserts an urgency for adaptation or other intervention to further support the learner in need.

Contributions & Implications: These findings contribute in many different ways. The models successfully capture the temporal dimension of online learning, signal changes in learners' behavior, adapt to these changes, and predict the future states of the learners. The final learner models are easy to interpret and to make sense about which are the exact learner traits at each point in time (for each one of the multiple skills measured), and how they have evolved to this point. It seems that the time-varying factors allow for gaining fruitful insight into learners' current cognitive state ("current-awareness"), and should be incorporated for updating the states of the learner models accordingly. Foremost, this combination of features and analysis methods make it feasible to detect differences between learners, discriminate and categorize them, and to detect differences in the behavior and knowledge of the same learner and forecast the learner's next state (*practical implication*). Our findings open the discussion towards "currently-aware" learner models in real-time: enhancing the models with temporal information and dynamics, and processing the collected interaction data in real-time, in a way that is both satisfactorily accurate and easy to make-sense (*implication for research*).

Research Question 3.1:

(a) Are there any differences in the usage of metacognitive help with respect to the learners' motivational profiles? If yes, how significant are these differences? (b) Which are the emerging help-seeking strategies? (c) How are these strategies associated with the motivational profiles?

Research Question 3.2:

(a) Can learners make-sense from the task-related analytics visualizations? If yes, how the actual usage of visualizations is related to the learners' perceptions of visualizations' usefulness? (b) Are there any differences in the usage of task-related analytics visualizations with respect to the learners' level of performance? If yes, how significant are these differences? (c) Which is the effect of metacognitive help on learners' performance? (d) Does the exploitation of task-related metacognitive information enhance the learners' performance? If yes, how significant is its effect on learning performance? (e) Do learners' interpretations of the metacognitive help actually help them to deeper engage with the task? How significant is this effect?

Research Question 3.3:

Are there any changes in learners' engagement and performance due to receiving metacognitive help, over time? If yes, how significant are these changes?

Overview: During the previous stages, the major role of motivational factors in the beginning of the self-assessment procedures was highlighted. At the third stage of the analytics development, the motivational profiles were associated with temporal patterns of help-seeking strategies, and the actual usage of help-seeking was associated to learners' on-task engagement and performance. Additional analytics were developed for capturing the learners' interactions

with a help feature that delivered task-related analytics visualizations to the learners, as on-demand data-driven metacognitive instrumental help. The instrumental, metacognitive information about the tasks was extracted from the logged learners' interactions with the tasks (i.e., it is learner-centered), and was delivered to the learners on-demand, as simple bar/column charts. Help-seeking was modeled in terms of frequencies of requests for this additional information and time-spent on viewing it and sense-making. Three studies were conducted: an exploratory, an experimental and a longitudinal study, and different data-analysis methods were utilized to evaluate the added value of the extracted analytics.

Study 1 – Applying Variance-based and Pattern-based approaches for associating Motivational Profiles with Help-seeking Strategies.

Findings: The patterns-based analysis revealed three distinctive help-seeking strategies and three configurations of the motivational factors that are associated with the same or different help-seeking strategies. The insight gained from the constructed help-seeking analytics indices is as follows: extending previous findings, the results highlighted that highly motivated learners consistently sought for metacognitive help when the learning task had increased complexity. Moreover, although less motivated learners seemed to request for help, they were less likely to process this information. The most interesting finding was that medium motivated learners were less likely to persistently exhibit low help-seeking behavior: these learners seemed aware of their need for help, and they tended to externalize this need instead of avoiding help-seeking. However, their help-seeking requests decreased for the more difficult items, implying a scepticism on whether these learners finally decided to use the help or not. The lack of previous research – to the best of our knowledge – on explaining the medium motivated learners' help-seeking strategies, accounts for the added value of the present study.

Contributions & Implications: Beyond confirming, contradicting or extending previous results, this study was the first one – to the best of our knowledge – that dived into the learners' interactions with metacognitive support and associated the behavioral patterns (strategies) of this type of help-seeking with the learner's motivational profiles. Thus, methodologically, this study could guide researchers on how to utilize pattern-based methods to make sense of diverse analytics and take design decisions for various user groups. Moreover, discriminating help-seeking strategies according to the learners' motivational profiles could contribute to better adapting the delivery of help to facilitate the learners' goals, abilities and expectations. This means, that *researchers and practitioners* could work together towards designing learning environments enhanced with adaptive, learner-centric help-seeking features. Moreover, training the learners to read and understand the task-related analytics visualizations (i.e., the data-driven metacognitive instrumental help) and to make sense from this information is expected to help these learners to better regulate their effort, and in longer term, to improve their engagement and performance (*practical implication*).

Study 2 – Analyzing the impact of task-related metacognitive help on learners’ on-task engagement and learning performance in terms of appropriately designed analytics.

Findings: The overall results of this study demonstrated a coherent relationship between the actual usage of the task-related analytics visualizations, the learners’ engagement with the task, their learning performance, and their perceptions about the comprehensibility and helpfulness of the metacognitive help. In particular, it was found that high performing students used the analytics visualizations more often and they allocated considerable time to think and reflect about the received information, and elicit its implications. On the contrary, low-performing students rarely sought for assistance and requested for metacognitive hints about the tasks (probably because they felt uncomfortable with this type of information and they did not know how to use it). This finding provided additional empirical evidence to previously reported results that associated higher learning gains with time allocated on help-seeking and hint reasoning. Furthermore, this finding is in line with other research works that argue that learners in need usually do not seek help, while learners who can achieve higher – even without additional support – tend to ask for complementary hints and resources. Moreover, a consistent pattern of help-seeking and perceptions about the usefulness was revealed, as well. Learners who critically assessed the visualized information and inferred the actual difficulty/requirements of the tasks, regulated their time-allocation and effort expenditure accordingly, and perceived the delivered help as informative and actionable, as well. The added value of this finding is that it provided a preliminary insight on how learners could transfer the knowledge inferred from metacognitive information into practice and convert it into action, in response to previous claims: using the visualizations, they identify critical tasks and regulate their engagement accordingly.

Contributions & Implications: Methodologically, this study exploited learning analytics from two perspectives: (a) their direct feeding to the learners as metacognitive information about the tasks, and (b) their adoption as a consolidated research method to study the actual usage of metacognitive help, its effect on engagement and performance, and the learners’ perceptions about its usefulness. As such, this study contributes by employing analytics both for the measurement of help-seeking, as well as for the analysis and reporting of its effect on learning (*methodological implication*). Furthermore, the study proposed and evaluated a structural model for explaining engagement and performance by associating help-seeking with response-times and effort. The model explained almost 80% of the variance in performance, whilst the large effect size of help-seeking on performance highlights the need to further investigate the role of task-related metacognitive help in the learning process (*implication for research*). Moreover, the study adds to the help-seeking literature by introducing and evaluating a task-related analytics approach as on-demand metacognitive help. In other words, it suggested delivering to the learners metacognitive information about the tasks instead of their own behavior, in order to support them dealing with the tasks (*practical implication*). The added value of the shift from self-related to task-related analytics derives from the multiple benefits the latter have for the

learners. Specifically, the task-related analytics seemed to boost the learners to practice higher-order critical thinking abilities for sense-making and decision-making: the learners evaluated, assessed and mostly contributed to generating peripheral information about the task (i.e., that is not directly related to the content of the task), and they used this information to adjust their actions accordingly (*practical implication*). In addition, the fact that the information is about the task and not the learner herself adds in drawing attention on the task rather than the self. This type of information can increase the learners' awareness about the actual requirements of the learning task because the analytics are measurements from actual interactions with the task. Consequently, the learners may develop time-regulation competences (i.e., regulation of time-allocation and on-task effort expenditure) according to the actual needs of the task (*practical implication*). The most interesting finding of this study concerned the capacity that task-related metacognitive help has to promote "responsible learning" from a social-interactive perspective (*theoretical implication*). Elaborating further on that claim, the task-related analytics were collectively generated and extracted from the actual interactions of all learners. The results of this study shown that the more the learners become aware that their engagement with the task affects the analytics the other learners receive, the more careful they might become, and thus, the more responsible for their choices and actions (as seen from the reduced guessing and increased effort). This opens new research directions towards integrating the "collective intelligence" factor in supporting the help-seeking process (*implication for research*). Foremost, instead of requesting for immediate assistance or getting informed about their own behavior – that is directly depending on the (peer/teacher/system) tutor – through the task-related analytics the learners elicit help from a self-reflection, task-oriented process. Thus, they maintain their sense of independence from the tutor, which could potentially contribute to enhancing their autonomous learning capacity (*implication for research*).

Study 3 – A longitudinal approach on how learners' behavior changes over time and how help-seeking affects these changes.

Findings: The results of the study confirmed once again the statistically significant effect of the distinct response-times on learning performance, even in a setting of a longitudinal study. In addition, it was found that between the first measurements of analytics parameters, prior to exposing learners to the treatment (i.e., the task-related analytics visualizations), and the second measurement, when the metacognitive help was available, the difference in the response-times was statistically significant and the effect of the difference on explaining the difference on learning performance was statistically medium. Similarly, the difference in response-times explains satisfactorily the difference in performance between the first and the third phase of the repeated measurements, whilst the effects of these differences are statistically small between the second and third phases. What mediated and caused these differences was the usage of the analytics visualizations. Thus, what these findings imply is that the intervention employed, i.e., the usage of the available metacognitive help, strongly contributes in increasing learners' on-task

engagement, which in turn, results in improved performance. No statistically significant difference on learners' effort between the three phases was found.

Contributions & Implications: The core contributions of this study were threefold:

(a) methodologically it was one of the very limited in number studies in the field of learning analytics that implemented a longitudinal research design. This study showcased how the time metric for describing change can be coded to facilitate the research design. Time of measurements is frequently a predictor variable in longitudinal research. As shown from the findings, this factor was included in the final fixed effect model, and it indeed was one of the strong determinants of the change in learning performance. As such, this study provided the description of a coherent longitudinal study in the area of learning analytics (*methodological implication*).

(b) from the findings became apparent that the metacognitive help seeking caused significant changes in learners' behaviour in terms of response-times, which in turn resulted in changes in performance. Investigating the effects of difference in the usage of the on-demand task-related visualizations on the changes in performance is necessary to be clarified, as well. Designing and implementing longer longitudinal studies, with more phases of exposing the sample to the treatments (i.e., more points in time) would facilitated that objective (*implication for research*). In addition, providing alternative forms of assistance (e.g., executive help formats, explicit hints, etc.), measuring the effects of the differences in response-times and effort, and comparing these differences to the ones estimated in this study is expected to shed light to the effect size of the employed intervention (*implication for research*).

(c) combining the findings of this study with previous results that indicated an alignment of using the metacognitive help with perceiving them as useful and helpful, further justified the role and significance of the intervention. As such, it provided a strong indication that training learners to use, read, and make-sense from learning analytics fosters their metacognition and assists them to ask for assistance at the moment they actually need it. In a sense, the findings provided empirical evidence on the added-value of enhancing the learning environments with interaction features that facilitate the learners' self-directed decisions. Accordingly, further training the learners on how to efficiently use such features is expected to build their capacity for autonomous learning (*practical implication*).

Research Question 4.1:

(a) Which is the effect of self-regulated learning strategies on the learners' control of autonomous learning? (b) Which is the impact of autonomous control on learners' performance, response-times and effort? (c) Does the exploitation of autonomous control (measured with utilized analytics) contribute to enhancing the learners' performance? If yes, how significant is its effect on learning performance?

Research Question 4.2:

(a) Are there any differences in the analytics parameters of autonomous control, with respect to the learners' level of performance? If yes, how significant is the effect of each one of these parameters? (b) Are there any differences in the learners' engagement with the self-assessment task, with respect to their level of performance? If yes, how significant is the effect of each one of these parameters?

Research Question 4.3:

To what extent can we exploit learning analytics to assess learners' autonomous capacity development?

Overview: At the fourth and last stage of analytics development, the focus was shifted on measuring and coding, as well as assessing the learners' autonomous choices during the online self-assessment. The self-assessment environment was adjusted in order to support autonomous interactions and to allow the learners to select on their own the next self-assessment task, based on the difficulty of the items and the self-set self-assessment goals. Three levels of autonomous control were defined: full-autonomy, "semi-"autonomy and no-autonomy. Overall, autonomy was modeled in terms of frequencies of interaction types and time-spent on decision making. Two studies were conducted, i.e., an exploratory and an experimental. In the exploratory study, the impact of specific self-regulated learning strategies on autonomous choices was investigated in conditions that the learners had full-autonomy of choice. The selection of the self-regulated learning strategies was based on the findings from previous related work. In the experimental study, the effects of the different degrees of autonomous control on learners' on-task engagement and learning performance (measured with the analytics defined during the previous analytics development stages) were examined. Based on the overall findings, a holistic, analytics-driven model for assessing autonomous learning capacity development was proposed.

Study 1 - Applying Variance-based Structural Equation Modeling for explaining autonomous interactions and decisions according to self-regulated learning strategies.

Findings: The results showcased that goal-setting and time-management have strong positive effects on autonomous control, effort-regulation moderately positively affects learners' autonomy, while help-seeking has a strong negative effect. The overall prediction accuracy of the model proposed in this study was 33.2%, which is moderate. Although this seems an intuitive finding, however, one would expect a higher prediction accuracy, due to the high similarity between the concepts of autonomy and self-regulation, acknowledged in literature. The most intriguing finding of this study concerned the strong direct negative effect of the help-seeking factor on autonomous interactions, resulting in supporting the opposite of initial hypothesis, and contradicting previous claims that this self-initiated strategy also reflects learning autonomy.

Contributions & Implications: The study took us a step ahead on the understanding of autonomous capacity, by considering the learners' autonomous choices, using a learning analytics perspective. In the demonstrated approach, autonomy was modeled in terms of frequencies of interaction types and time-spent on decision making. Both analytics measures are in line with

other similar parameters developed through-out this doctoral research, or commonly used in the learning analytics research practise. The moderate prediction accuracy achieved in this study contributed to understanding the diversity between the concepts of self-regulation and autonomy, which are extensively used interchangeably in literature, and provided empirical evidence that justifies their discrimination; additional factors should be considered as well, as moderators that catalyse the relationship between self-regulation and autonomy (e.g., affective states) (*theoretical implication*). Furthermore, the strong positive effects of goal-expectancy and time-management on autonomy extended previous findings by adding empirical evidence, and imply that learners want to be consistent with their choices and to take responsibility for their learning (*theoretical implication*). These findings also contributed to understanding the role of time-management as a strategy towards achievement, further confirming the findings from previous stages of this research (see research question 1). Yet, the divergence of our finding from previous theoretical claims regarding the role of help-seeking indicated that the relationship between both self-initiated behaviors (i.e., autonomous choices and help-seeking) should be further explored (*implication for research*). Clarifying this relationship shall next open new directions towards making decisions on how to efficiently guide the learners to feel free to seek help (with all the learning gains that help-seeking brings in) (*practical implication*). Overall, understanding how learners' self-regulation contributes to and influences their autonomy can provide insight on how to plan the learners' SRL support in online learning environments (*implications for research*). This means that, in the next step, practitioners shall be able to integrate specific features into online learning environment, in order to train learners on effectively using the SRL strategies, accordingly (*practical implication*).

Study 2 - Analyzing the impact of autonomous interactions on learners' on-task engagement and learning performance.

Findings: The results of the study revealed that autonomous control has a strong positive effect on performance in both full-autonomous and "semi-"autonomous self-assessment conditions. Differences between groups were also identified with respect to the analytics parameters that reflect engagement, i.e., their effort expenditure and response-times. In other words, when autonomous interactions are available, learners take control of their choices and increase their effort to achieve the best possible outcome, avoiding guessing. Moreover, it was also found that the effects are stronger in adaptive, "semi-" autonomous self-assessment setting, compared to the full-autonomous conditions. The learners who received guided autonomy (in terms of proposing to them an appropriate item, as well as providing them the option to reject the suggestion and replace the item) exhibited higher engagement and increased performance.

Contributions & Implications: This study added to the autonomy literature by presenting a learning analytics-driven research method that revealed strong association between learners' interactions within controlled autonomous and semi-autonomous learning

environments and the learners' achievement (**methodological implication**). The strong direct effect of autonomous control on both types of response-times (i.e., to answer correctly/wrongly) and on effort-expenditure, indicated a relationship between autonomously choosing a learning item and time-spent on processing it, and constitutes empirical evidence that when learners make their own self-enforced choices of learning items, they are likely to treat these items more responsibly, by devoting more time on trying to deal with them. As such, it verified the respective assumptions in former studies and explained why learners who feel autonomous are likely to engage in deeper-processing mode and to invest more time on hands-on activities (**theoretical implication**). Due to the very high positive impact of autonomy on effort, the latest could be perceived as an *indicator of autonomous capacity development*. Furthermore, the higher effort exertion that learners exhibit in the "semi-" autonomous conditions might be due to the effect of the guided autonomy: learners are not by default experienced on how to be autonomous. In the adaptive conditions, the assessment itself guides the learners to achieve the best possible score. Enhancing the adaptive procedure with autonomous options, the assessment environment offers to the learner the opportunity to practice autonomy in guided conditions. As a result, the learner progressively engages more in the self-enforced choices, makes better decisions, needs less guidance on how to become autonomous (i.e., increases autonomous capacity), and finally, performs better (**practical implication**). The sense of autonomy triggered by the option of individually choosing the next item, is in fact a "pseudo-autonomous" control that increases the students' degree on engagement in the test (expressed as higher effort and response-times, as explained above). This can be interpreted as a "placebo" effect. Within guided conditions, the "pseudo-autonomous" control increased students' autonomous capacity: they perceived the process as autonomous, they believed that they were making their self-enforced choices and they were consciously involved in the self-assessment, yet they were seamlessly guided by the system to increase their effort and improve their performance (**implication for research**). Finally, by considering all the identified measures and effects of the explored factors, this thesis proposed a theoretical, yet analytics-driven model for assessing autonomous learning capacity development. The empirical evaluation of this model is expected to shed light to how practising self-regulation can lead to advancing autonomy, as well as to set the foundations for a framework that will guide decisions on designing autonomous learning environments (**implication for research**).

Research Question 5:

(a) *Can we accurately and efficiently recommend sequences of educational resources to homogeneous and heterogeneous groups of students, with respect to both the individuals' and the group's motivation, and intention to use the resources?* **(b)** *What is the impact of a recommendation on individual students' persistence as well as on the groups' learning performance in the collaborative problem-solving activity?*

Overview: Another objective of this thesis was to explore autonomous capacity within collaborative, group-learning conditions. Specifically, the goal was to consider the individuals'

self-enforced preferences and decisions within the group they belong to, in order to motivate their participation and to support them with appropriate group-recommendations of educational resources. Taking or not the recommendation is a self-directed decision, guided by how much the recommendation corresponds to and facilitates the individual's self-set learning goals. Game-theory was utilized for guiding the group-recommendations and an experimental study was conducted, and heuristics were used to evaluate the goodness of recommendation as well as the impact of recommendation on conformity, persistence and learning gain.

Study – Utilizing Game-Theory to guide the Recommendation of Educational Resources to Groups of learners with low, medium or high inner Degree of Similarity of the group-members Motivation.

Findings: According to the results of the study, all decision support methods achieved low approximation error in prediction of motivation ratings for the homogeneous students' sub-groups. However, the proposed game-theoretic (GT) strategy minimized the prediction error of the sub-group motivation ratings for all categories of inner sub-group similarity. Moreover, the suggested method had a good overall performance (i.e., the nDCG values reflecting the effectiveness of ranked list or recommendations), although not always the best; in one case of homogeneous groups, the "average" method provided slightly more effective recommendations compared to the list of GT method. However, it is important to notice that, compared to the other methods, the performance of the proposed GT was stable and robust, regardless of the inner sub-group similarity, targeting ranking quality and demonstrating only small variations. Another finding concerns the plurality of the recommended lists of items (diversity of items). Improving the diversity of recommendation results was expected to increase within-group motivation from the recommended resources. Furthermore, the recommendations were highly ranked for the individuals as well as for the group. In other words, the individuals within the groups did not have to highly adjust their personal consideration about their motivation gained from the recommended items, to be approved by the other group members. Finally, the learners' engagement with the resources throughout the collaborative activity, and regardless of the inner group similarity, was stable and high when the recommendations were generated with GT.

Contributions & Implications: The core contribution of this study was the development of a mathematical formulation for group-recommendation of educational resources as a non-cooperative game, as well as an architecture for building such group recommender systems. Introducing and configuring Game-Theory for guiding group-recommendation in the educational domain is by itself a totally new approach on the topic (*theoretical & methodological implication*). From the evaluation of the approach with a realistic dataset, and with heuristics that are commonly employed in the domain of recommender systems, the results revealed a tendency that the accuracy of the predicted group motivation, the goodness of the ranked list of recommendations, and the problem-solving performance for the experimental group were

significantly higher compared to other state-of-the-art methods. The diversity of the items in the recommendation, indicating a satisfactory personalization even for heterogeneous groups, was higher with the GT method, as well (*practical implication*). The measure employed, i.e., the distance between the individual nDCG and the respective group's nDCG, is a topic that deserves further analysis and could be explored as a measure of the individual's degree of conformity (*implication for research*).

14.2. Instead of Epilogue – Lesson learnt

Autonomous learning, as a self-enforced initiative, is essential because it gradually leads to efficient learning. And the more the learning turns online, the more the opportunities for autonomous learning are. However, developing capacity for autonomous learning is not trivial. The work presented in this thesis provides one example of how new developments in learning analytics can contribute to understanding the underlying mechanisms, by extracting measures of learners' autonomous interactions, as well as by constructing meaningful indices for assessing capacity development, in a productive manner.

Investing on both the individuals' capacity for self-development – through autonomous learning – as well as on the current advancements in data science, has great potential towards supporting decision-making: wrong decisions do not constitute failure, until we refuse to correct them; the most long-term failures are the outcome of making excuses instead of decisions. As Elon Musk stated, “failure is an option here; if you don't fail enough, you don't innovate enough”.

Bibliography

- Abdous, M., He, W., & Yen, C.-J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Educational Technology & Society, 15*(3), 77–88.
- Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research, 87*(3), 659–701. <http://doi.org/10.3102/0034654316689306>
- Adomavicius, G., & Tuzhilin, A. (2008). Context-aware Recommender Systems. In *Proceedings of the 2008 ACM Conference on Recommender Systems* (pp. 335–336). New York, NY, USA: ACM. <http://doi.org/10.1145/1454008.1454068>
- Aggarwal, C. C. (2006). *Data Streams: Models and Algorithms (Advances in Database Systems)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Ala-Mutka, K. M. (2005). A Survey of Automated Assessment Approaches for Programming Assignments. *Computer Science Education, 15*(2), 83–102. <http://doi.org/10.1080/08993400500150747>
- Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward Meta-cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor. *Int. J. Artif. Intell. Ed., 16*(2), 101–128.
- Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help Seeking and Help Design in Interactive Learning Environments. *Review of Educational Research, 73*(3), 277–320. <http://doi.org/10.3102/00346543073003277>
- Ali, L., Hatala, M., Gašević, D., & Jovanović, J. (2012). A Qualitative Evaluation of Evolution of a Learning Analytics Tool. *Comput. Educ., 58*(1), 470–489. <http://doi.org/10.1016/j.compedu.2011.08.030>
- Alpaydin, E. (2010). *Introduction to Machine Learning* (2nd ed.). The MIT Press.
- Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician, 46*(3), 175–185. <http://doi.org/10.1080/00031305.1992.10475879>
- Amer-Yahia, S., Roy, S. B., Chawlat, A., Das, G., & Yu, C. (2009). Group Recommendation: Semantics and Efficiency. *Proc. VLDB Endow., 2*(1), 754–765. <http://doi.org/10.14778/1687627.1687713>
- Anaya, A. R., Luque, M., & Peinado, M. (2016). A visual recommender tool in a collaborative learning experience. *Expert Systems with Applications, 45*, 248–259. <http://doi.org/https://doi.org/10.1016/j.eswa.2015.01.071>
- Anderson (Ed.), L. W., Krathwohl (Ed.), D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., ... Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives*. New York : Longman.
- Andrade, M. S. (2014). Dialogue and Structure: Enabling Learner Self-Regulation in Technology-Enhanced Learning Environments. *European Educational Research Journal, 13*(5), 563–574. <http://doi.org/10.2304/eeerj.2014.13.5.563>
- Andrade, M. S., & Bunker, E. L. (2009). A model for self-regulated distance language learning. *Distance Education, 30*(1), 47–61. <http://doi.org/10.1080/01587910902845956>
- Ardissono, L., Goy, A., Petrone, G., Segnan, M., & Torasso, P. (2003). INTRIGUE: Personalized recommendation of tourist attractions for desktop and hand held devices. *Applied Artificial Intelligence, 17*(8), 687–714. <http://doi.org/10.1080/713827254>
- Arnold, I. J. M. (2016). Cheating at online formative tests: Does it pay off? *The Internet and Higher Education, 29*, 98–106. <http://doi.org/http://dx.doi.org/10.1016/j.iheduc.2016.02.001>

- Arroyo, I., & Woolf, B. P. (2005). Inferring Learning and Attitudes from a Bayesian Network of Log File Data. In *Proceedings of the 2005 Conference on Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology* (pp. 33–40). Amsterdam, The Netherlands, The Netherlands: IOS Press.
- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In *Groups, leadership and men; research in human relations*. (pp. 177–190). Oxford, England: Carnegie Press.
- Azevedo, R. (2014). Issues in dealing with sequential and temporal characteristics of self- and socially-regulated learning. *Metacognition and Learning*, 9(2), 217–228. <http://doi.org/10.1007/s11409-014-9123-1>
- Baard, P. P., Deci, E. L., & Ryan, R. M. (2004). Intrinsic Need Satisfaction: A Motivational Basis of Performance and Well-Being in Two Work Settings. *Journal of Applied Social Psychology*, 34(10), 2045–2068. <http://doi.org/10.1111/j.1559-1816.2004.tb02690.x>
- Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). Models and Issues in Data Stream Systems. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 1–16). New York, NY, USA: ACM. <http://doi.org/10.1145/543613.543615>
- Baker, B. M. (2007). *A conceptual framework for making knowledge actionable through capital formation*. University of Maryland University College, United States -- Maryland.
- Baker, R., Corbett, A., Roll, I., & Koedinger, K. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 18(3), 287–314. <http://doi.org/10.1007/s11257-007-9045-6>
- Baker, R., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task Behavior in the Cognitive Tutor Classroom: When Students “Game the System.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 383–390). New York, NY, USA: ACM. <http://doi.org/10.1145/985692.985741>
- Baker, R. S. J. d, Corbett, A. T., & Aleven, V. (2008a). Improving contextual models of guessing and slipping with a truncated training set. In *Proceedings of the 1st International Conference on Educational Data Mining, EDM 2008*, (pp. 67–76). Montreal, Canada.
- Baker, R. S. J. d, Corbett, A. T., & Aleven, V. (2008b). More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In B. P. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings* (pp. 406–415). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-69132-7_44
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (Vol. 5). Greenwich, CT: Information Age Publishing.
- Barclay, D., Higgins, C., & Thompson, R. (1995). The partial least squares (PLS) approach to causal modeling: Personal computer adoption and use as an illustration. *Technology Studies*, 2(2), 285–309.
- Barla, M., Bieliková, M., Ezzeddinne, A. B., Kramár, T., Šimko, M., & Vozár, O. (2010). On the Impact of Adaptive Test Question Selection for Learning Efficiency. *Comput. Educ.*, 55(2), 846–857. <http://doi.org/10.1016/j.compedu.2010.03.016>
- Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S.-L. (2009). Measuring self-regulation in online and blended learning environments. *The Internet and Higher Education*, 12(1), 1–6. <http://doi.org/http://dx.doi.org/10.1016/j.iheduc.2008.10.005>
- Barnes, T. (2005). Q-matrix Method: Mining Student Response Data for Knowledge. In *In the*

Technical Report (WS-05-02) of the AAAI-05 Workshop on Educational Data Mining.

- Barrick, R. M., Mount, K. M., & Judge, A. T. (2001). Personality and Performance at the Beginning of the New Millennium: What Do We Know and Where Do We Go Next? *International Journal of Selection and Assessment*, 9(1-2), 9–30. <http://doi.org/10.1111/1468-2389.00160>
- Beck, J. E. (2005). Engagement Tracing: Using Response Times to Model Student Disengagement. In *Proceedings of the 2005 Conference on Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology* (pp. 88–95). Amsterdam, The Netherlands, The Netherlands: IOS Press.
- Beck, J. E., & Chang, K. (2007). Identifiability: A Fundamental Problem of Student Modeling. In C. Conati, K. McCoy, & G. Paliouras (Eds.), *User Modeling 2007* (pp. 137–146). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Beck, J. E., & Gong, Y. (2013). Wheel-Spinning: Students Who Fail to Master a Skill. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 431–440). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Beck, J. E., & Woolf, B. P. (2000). High-Level Student Modeling with Machine Learning. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent Tutoring Systems* (pp. 584–593). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Belk, M., Germanakos, P., Fidas, C., & Samaras, G. (2014). A Personalization Method Based on Human Factors for Improving Usability of User Authentication Tasks. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), *User Modeling, Adaptation, and Personalization* (pp. 13–24). Cham: Springer International Publishing.
- Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large scale educational testing*. Princeton, NJ: Educational Testing Service, Policy Information Center.
- Benson, P. (2001). *Teaching and Researching Autonomy in Language Learning*. London: Longman.
- Berkovsky, S., & Freyne, J. (2010). Group-based Recipe Recommendations: Analysis of Data Aggregation Strategies. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 111–118). New York, NY, USA: ACM. <http://doi.org/10.1145/1864708.1864732>
- Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers.
- Biderman, M. D., Nguyen, N. T., & Sebren, J. (2008). Time-on-task mediates the conscientiousness–performance relationship. *Personality and Individual Differences*, 44(4), 887–897. <http://doi.org/https://doi.org/10.1016/j.paid.2007.10.022>
- Bidjerano, T., & Dai, D. Y. (2007). The relationship between the big-five model of personality and self-regulated learning strategies. *Learning and Individual Differences*, 17(1), 69–81. <http://doi.org/https://doi.org/10.1016/j.lindif.2007.02.001>
- Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *US Department of Education, Office of Educational Technology*, 1, 1–57.
- Bifet, A., & Gavaldà, R. (2009). Adaptive Learning from Evolving Data Streams. In N. M. Adams, C. Robardet, A. Siebes, & J.-F. Boulicaut (Eds.), *Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, August 31 - September 2, 2009. Proceedings* (pp. 249–260). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-03915-7_22
- Bifet, A., Holmes, G., Kirkby, R., & Pfahringer, B. (2010). MOA: Massive Online Analysis. *J. Mach.*

Learn. Res., 11, 1601–1604.

- Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., & Gavaldà, R. (2009). New Ensemble Methods for Evolving Data Streams. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 139–148). New York, NY, USA: ACM.
<http://doi.org/10.1145/1557019.1557041>
- Bifet, A., Read, J., Žliobaite, I., Pfahringer, B., & Holmes, G. (2013). Pitfalls in Benchmarking Data Stream Classification and How to Avoid Them. In H. Blockeel, K. Kersting, S. Nijssen, & F. Železný (Eds.), *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part I* (pp. 465–479). Berlin, Heidelberg: Springer Berlin Heidelberg.
http://doi.org/10.1007/978-3-642-40988-2_30
- Bipp, T., Steinmayr, R., & Spinath, B. (2008). Personality and achievement motivation: Relationship among Big Five domain and facet scales, achievement goals, and intelligence. *Personality and Individual Differences*, 44(7), 1454–1464.
<http://doi.org/https://doi.org/10.1016/j.paid.2008.01.001>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. L. & M. R. Novick (Ed.), *Statistical Theories of Mental Test Scores* (pp. 397–472).
- Blikstein, P. (2011). Using Learning Analytics to Assess Students' Behavior in Open-ended Programming Tasks. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 110–116). New York, NY, USA: ACM.
<http://doi.org/10.1145/2090116.2090132>
- Blikstein, P., & Worsley, M. A. B. (2016). Multimodal Learning Analytics and Education Data Mining: Using computational technologies to measure complex learning tasks. *The Journal of Learning Analytics*, 3(2), 220–238.
- Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender Systems Survey. *Know.-Based Syst.*, 46, 109–132. <http://doi.org/10.1016/j.knosys.2013.03.012>
- Boroujeni, A., Roohani, A., & Hasanimanesh, A. (2015). The impact of extroversion and introversion personality types on EFL learners' writing ability. *Theory & Practice In Language Studies*, 5(1), 212–218.
- Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In Y. Lechevallier & G. Saporta (Eds.), *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers* (pp. 177–186). Heidelberg: Physica-Verlag HD.
http://doi.org/10.1007/978-3-7908-2604-3_16
- Brawner, K. W., & Gonzalez, A. J. (2016). Modelling a learner's affective state in real time to improve intelligent tutoring effectiveness. *Theoretical Issues in Ergonomics Science*, 17(2), 183–210. <http://doi.org/10.1080/1463922X.2015.1111463>
- Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2), 123–140.
<http://doi.org/10.1023/A:1018054314350>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<http://doi.org/10.1023/A:1010933404324>
- Broadbent, J. (2017). Comparing online and blended learner's self-regulated learning strategies and academic performance. *The Internet and Higher Education*, 33, 24–32.
<http://doi.org/https://doi.org/10.1016/j.iheduc.2017.01.004>
- Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, 27, 1–13. <http://dx.doi.org/10.1016/j.iheduc.2015.04.007>

- Brophy, J. (2004). *Motivating students to learn, 2nd ed. Motivating students to learn, 2nd ed.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Brusilovsky, P. (2001). Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 11(1), 87–110. <http://doi.org/10.1023/A:1011143116306>
- Brusilovsky, P., Karagiannidis, C., & Sampson, D. (2004). Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Lifelong Learning*, 14, 2004.
- Brusilovsky, P., & Peylo, C. (2003). Adaptive and Intelligent Web-based Educational Systems. *Int. J. Artif. Intell. Ed.*, 13(2–4), 159–172.
- Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R., Zadorozhny, V., & Durlach, P. J. (2016). Open Social Student Modeling for Personalized Learning. *IEEE Transactions on Emerging Topics in Computing*, 4(3), 450–461. <http://doi.org/10.1109/TETC.2015.2501243>
- Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331–370. <http://doi.org/10.1023/A:1021240730564>
- Burrus, J., Jackson, T., Holtzman, S., Roberts, R. D., & Mandigo, T. (2013). EXAMINING THE EFFICACY OF A TIME MANAGEMENT INTERVENTION FOR HIGH SCHOOL STUDENTS. *ETS Research Report Series*, 2013(2), i--35. <http://doi.org/10.1002/j.2333-8504.2013.tb02332.x>
- Calvo, R. A., & D’Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing*, 1(1), 18–37. <http://doi.org/10.1109/T-AFFC.2010.1>
- Candy, P. C. (1991). *Self-direction for lifelong learning*. San Francisco: Jossey-Bass.
- Cao, J., & Stokes, S. L. (2007). Bayesian IRT Guessing Models for Partial Guessing Behaviors. *Psychometrika*, 73(2), 209. <http://doi.org/10.1007/s11336-007-9045-9>
- Capa, R. L., Audiffren, M., & Ragot, S. (2008). The effects of achievement motivation, task difficulty, and goal difficulty on physiological, behavioral, and subjective effort. *Psychophysiology*, 45(5), 859–868. <http://doi.org/10.1111/j.1469-8986.2008.00675.x>
- Carvalho, L. A. M. C., & Macedo, H. T. (2013). Users’ Satisfaction in Recommendation Systems for Groups: An Approach Based on Noncooperative Games. In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 951–958). New York, NY, USA: ACM. <http://doi.org/10.1145/2487788.2488090>
- Casal, G. B., Caballo, V. E., Cueto, E. G., & Cubos, P. F. (1990). Attention and reaction time differences in introversion-extraversion. *Personality and Individual Differences*, 11(2), 195–197. [http://doi.org/https://doi.org/10.1016/0191-8869\(90\)90015-j](http://doi.org/https://doi.org/10.1016/0191-8869(90)90015-j)
- Catmur, C. (2013). Sensorimotor learning and the ontogeny of the mirror neuron system. *Neuroscience Letters*, 540, 21–27. <http://doi.org/10.1016/j.neulet.2012.10.001>
- Cen, H., Koedinger, K., & Junker, B. (2006). Learning Factors Analysis -- A General Method for Cognitive Model Evaluation and Improvement. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems* (pp. 164–175). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Challis, D. (2005). Committing to quality learning through adaptive online assessment. *Assessment & Evaluation in Higher Education*, 30(5), 519–527. <http://doi.org/10.1080/02602930500187030>
- Chamorro-Premuzic, T., & Furnham, A. (2005). *Personality and intellectual competence*. London: Lawrence Erlbaum.
- Chamorro-Premuzic, T., Furnham, A., & Lewis, M. (2007). Personality and approaches to

- learning predict preference for different teaching methods. *Learning and Individual Differences*, 17(3), 241–250. <http://doi.org/https://doi.org/10.1016/j.lindif.2006.12.001>
- Chang, S.-R., Plake, B. S., Kramer, G. A., & Lien, S.-M. (2011). Development and Application of Detection Indices for Measuring Guessing Behaviors and Test-Taking Effort in Computerized Adaptive Testing. *Educational and Psychological Measurement*, 71(3), 437–459. <http://doi.org/10.1177/0013164410385110>
- Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A Reference Model for Learning Analytics. *Int. J. Technol. Enhanc. Learn.*, 4(5/6), 318–331. <http://doi.org/10.1504/IJTEL.2012.051815>
- Chatzopoulou, D. I., & Economides, A. A. (2010). Adaptive assessment of student's knowledge in programming courses. *Journal of Computer Assisted Learning*, 26(4), 258–269. <http://doi.org/10.1111/j.1365-2729.2010.00363.x>
- Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, 54(8), 88. <http://doi.org/10.1145/1978542.1978562>
- Chen, C.-M., & Chen, M.-C. (2009). Mobile Formative Assessment Tool Based on Data Mining Techniques for Supporting Web-based Learning. *Comput. Educ.*, 52(1), 256–273. <http://doi.org/10.1016/j.compedu.2008.08.005>
- Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Recommendation System for Adaptive Learning. *Applied Psychological Measurement*, 42(1), 24–41. <http://doi.org/10.1177/0146621617697959>
- Cheng, K.-H., & Tsai, C.-C. (2011). An investigation of Taiwan University students' perceptions of online academic help seeking, and their web-based learning self-efficacy. *The Internet and Higher Education*, 14(3), 150–157. <http://doi.org/10.1016/J.IHEDUC.2011.04.002>
- Chi, M., Koedinger, K., Gordon, G., Jordan, P., & Van Lehn, K. (2011). Instructional Factors Analysis: A Cognitive Model For Multiple Instructional Interventions. In *Proceedings of Educational Data Mining*.
- Chin, W. W. (1998). The partial least squares approach for structural equation modeling. In *Modern methods for business research*. (pp. 295–336). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Chin, W. W., & Dibbern, J. (2010). An Introduction to a Permutation Based Procedure for Multi-Group PLS Analysis: Results of Tests of Differences on Simulated Data and a Cross Cultural Analysis of the Sourcing of Information System Services Between Germany and the USA. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications* (pp. 171–193). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-32827-8_8
- Chrysafiadi, K., & Virvou, M. (2013a). PeRSIVA: An Empirical Evaluation Method of a Student Model of an Intelligent e-Learning Environment for Computer Programming. *Comput. Educ.*, 68, 322–333. <http://doi.org/10.1016/j.compedu.2013.05.020>
- Chrysafiadi, K., & Virvou, M. (2013b). Review: Student Modeling Approaches: A Literature Review for the Last Decade. *Expert Syst. Appl.*, 40(11), 4715–4729. <http://doi.org/10.1016/j.eswa.2013.02.007>
- Chua, Y. P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior*, 28(5), 1580–1586. <http://doi.org/http://dx.doi.org/10.1016/j.chb.2012.03.020>
- Claessens, B. J. C., van Eerde, W., Rutte, C. G., & Roe, R. A. (2007). A review of the time management literature. *Personnel Review*, 36(2), 255–276. <http://doi.org/10.1108/00483480710726136>

- Clarebout, G., & Elen, J. (2009). Benefits of inserting support devices in electronic learning environments. *Computers in Human Behavior*, 25(4), 804–810.
<http://doi.org/10.1016/j.chb.2008.07.006>
- Clarebout, G., Horz, H., & Elen, J. (2009). The use of support devices in electronic learning environments. *Computers in Human Behavior*, 25(4), 793–794.
<http://doi.org/10.1016/j.chb.2008.07.004>
- Clow, D., & Makriyannis, E. (2011). iSpot Analysed: Participatory Learning and Reputation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 34–43). New York, NY, USA: ACM. <http://doi.org/10.1145/2090116.2090121>
- Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design Experiments in Educational Research. *Educational Researcher*, 32(1), 9–13.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.
- Conard, M. A. (2006). Aptitude is not enough: How personality and behavior predict academic performance. *Journal of Research in Personality*, 40(3), 339–346.
<http://doi.org/10.1016/j.jrp.2004.10.003>
- Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian Networks to Manage Uncertainty in Student Modeling. *User Modeling and User-Adapted Interaction*, 12(4), 371–417.
<http://doi.org/10.1023/A:1021258506583>
- Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19(3), 267–303.
<http://doi.org/10.1007/s11257-009-9062-8>
- Conejo, R., Guzmán, E., Perez-de-la-Cruz, J.-L., & Barros, B. (2014). An empirical study on the quantitative notion of task difficulty. *Expert Systems with Applications*, 41(2), 594–606.
<http://doi.org/http://dx.doi.org/10.1016/j.eswa.2013.07.084>
- Confessore G., & P. E. (2004). Factor validation of the learner autonomy profile (Version 3.0) and Extraction of the Short Form. *International Journal of Self Directed Learning*, 1, 39–58.
- Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, 59(2), 661–686.
<http://doi.org/https://doi.org/10.1016/j.compedu.2012.03.004>
- Cook, D. J., & Das, S. K. (2012). Review: Pervasive Computing at Scale: Transforming the State of the Art. *Pervasive Mob. Comput.*, 8(1), 22–35. <http://doi.org/10.1016/j.pmcj.2011.10.004>
- Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.
<http://doi.org/10.1007/BF01099821>
- Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In *Handbook of Human-Computer Interaction (Second Edition)* (pp. 849–874).
- Corrin, L., & de Barba, P. (2015). How Do Students Interpret Feedback Delivered via Dashboards? In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 430–431). New York, NY, USA: ACM.
<http://doi.org/10.1145/2723576.2723662>
- Corrin, L., de Barba, P. G., & Bakharia, A. (2017). Using Learning Analytics to Explore Help-seeking Learner Profiles in MOOCs. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 424–428). New York, NY, USA: ACM.
<http://doi.org/10.1145/3027385.3027448>

- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <http://doi.org/10.1007/BF00994018>
- Costa, P. T. J., & McCrae, R. R. (1992). *NEO-PI-R: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Cotterall, S. (1995). Readiness for autonomy: Investigating learner beliefs. *System*, 23(2), 195–205. [http://doi.org/10.1016/0346-251X\(95\)00008-8](http://doi.org/10.1016/0346-251X(95)00008-8)
- Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti, A. E., Sterbinsky, A., Cheney, K. R., & Okwumabua, T. (2013). The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Computers and Education*, 68, 495–504. <http://doi.org/10.1016/j.compedu.2013.06.010>
- Cross, S., Waters, Z., Kitto, K., & Zuccon, G. (2017). Classifying Help Seeking Behaviour in Online Communities. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 419–423). New York, NY, USA: ACM. <http://doi.org/10.1145/3027385.3027442>
- Daley, S. G., Hillaire, G., & Sutherland, L. M. (2016). Beyond performance data: Improving student help seeking by collecting and displaying influential data in an online middle-school science curriculum. *British Journal of Educational Technology*, 47(1), 121–134. <http://doi.org/10.1111/bjet.12221>
- Daniels, L. M., & Gierl, M. J. (2017). The impact of immediate test score reporting on university students' achievement emotions in the context of computer-based multiple-choice exams. *Learning and Instruction*. <http://doi.org/10.1016/j.learninstruc.2017.04.001>
- Davis, A. (1999). *The Limits of Educational Assessment*. Hoboken, NJ: Wiley-Blackwell.
- Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.*, 13(3), 319–340. <http://doi.org/10.2307/249008>
- Dawson, S., Joksimović, S., Kovanović, V., Gašević, D., & Siemens, G. (2015). Recognising learner autonomy: {Lessons} and reflections from a joint x/c {MOOC}. In *Proceedings of 2015 {HERDSA} conference*. Melbourne, AU: HERDSA.
- de Amorim, R. C., & Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, 324, 126–145. <http://doi.org/10.1016/j.ins.2015.06.039>
- De Bra, P. (2002). Adaptive Educational Hypermedia on the Web. *Commun. ACM*, 45(5), 60–61. <http://doi.org/10.1145/506218.506247>
- De Raad, B., & Schouwenburg, H. C. (1996). Personality traits in learning and education. *European Journal of Personality*, 10, 185–200.
- Deci, E. L., Eghrari, H., Patrick, B. C., & Leone, D. R. (1994). Facilitating Internalization: The Self-Determination Theory Perspective. *Journal of Personality*, 62(1), 119–142. <http://doi.org/10.1111/j.1467-6494.1994.tb00797.x>
- Deci, E. L., & Ryan, R. M. (Eds.). (2002). *Handbook of self-determination research. Handbook of self-determination research*. Rochester, NY, US: University of Rochester Press.
- Dejaeger, K., Goethals, F., Giangreco, A., Mola, L., & Baesens, B. (2012). Gaining insight into student satisfaction using comprehensible data mining techniques. *European Journal of Operational Research*, 218(2), 548–562. <http://dx.doi.org/10.1016/j.ejor.2011.11.022>
- Dekker G.W., P. M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In D. M. R. C. & V. S. Barnes T. (Ed.), *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009* (pp. 41–50). Cordoba, Spain.
- Desmarais, M. C., & Baker, R. S. (2012). A Review of Recent Advances in Learner and Skill

- Modeling in Intelligent Learning Environments. *User Modeling and User-Adapted Interaction*, 22(1–2), 9–38. <http://doi.org/10.1007/s11257-011-9106-8>
- Devaraj, S., Easley, R. F., & Crant, J. M. (2008). How does personality matter? Relating the five-factor model to technology acceptance and use. *Information Systems Research*, 19(1), 93–105. <http://doi.org/10.1287/isre.1070.0153>
- Dickinson, L. (1995). Autonomy and motivation a literature review. *System*, 23(2), 165–174. [http://doi.org/10.1016/0346-251X\(95\)00005-5](http://doi.org/10.1016/0346-251X(95)00005-5)
- Dickman, S. J., & Meyer, D. E. (1988). Impulsivity and speed-accuracy tradeoffs in information processing. *Journal of Personality and Social Psychology*. US: American Psychological Association. <http://doi.org/10.1037/0022-3514.54.2.274>
- Digman, J. M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, 41(1), 417–440. <http://doi.org/10.1146/annurev.ps.41.020190.002221>
- Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*. US: American Psychological Association. <http://doi.org/10.1037/0022-3514.73.6.1246>
- Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence*, 41(1), 1–10. <http://doi.org/http://dx.doi.org/10.1016/j.intell.2012.10.003>
- Doll, W. J., & Torkzadeh, G. (1988). The Measurement of End-user Computing Satisfaction. *MIS Q.*, 12(2), 259–274. <http://doi.org/10.2307/248851>
- Doty, D. H., Glick, W. H., & Huber, G. P. (1993). Fit, equifinality, and organizational effectiveness: A test of two configurational theories. *Academy of Management Journal*, 36(6), 1196–1250.
- Duckworth, A. L., & Yeager, D. S. (2015). Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. *Educational Researcher*, 44(4), 237–251. <http://doi.org/10.3102/0013189X15584327>
- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification (2Nd Edition)*. Wiley-Interscience.
- Dul, J. (2016). Identifying single necessary conditions with NCA and fsQCA. *Journal of Business Research*, 69(4), 1516–1523.
- Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293), 52–64. <http://doi.org/10.1080/01621459.1961.10482090>
- Durall, E., & Gros, B. (2014). Learning Analytics as a Metacognitive Tool. In *CSEdu* (pp. 380–384).
- Dweck, C. S. (1986). Motivational Processes Affecting Learning. *American Psychologist*, 41(10), 1040–1048.
- Dwivedi, P., & Bharadwaj, K. K. (2015). e-Learning recommender system for a group of learners based on the unified learner profile approach. *Expert Systems*, 32(2), 264–276. <http://doi.org/10.1111/exsy.12061>
- Eckerson, W. W. (2006). *Performance dashboards: Measuring, monitoring, and managing your business*. Hoboken, New Jersey: John Wiley & Sons.
- Economides, A. A. (2005). Personalized feedback in CAT. *WSEAS Transactions on Advances in Engineering Education*, 3(2), 174–181.
- Economides, A. A. (2009a). Adaptive Context-Aware Pervasive and Ubiquitous Learning. *Int. J. Technol. Enhanc. Learn.*, 1(3), 169–192. <http://doi.org/10.1504/IJTEL.2009.024865>
- Economides, A. A. (2009b). Conative Feedback in Computer-Based Assessment. *Computers in the*

- Schools*, 26(3), 207–223. <http://doi.org/10.1080/07380560903095188>
- Eilam, B., & Aharon, I. (2003). Students' planning in the process of self-regulated learning. *Contemporary Educational Psychology*, 28(3), 304–334. [http://doi.org/https://doi.org/10.1016/S0361-476X\(02\)00042-5](http://doi.org/https://doi.org/10.1016/S0361-476X(02)00042-5)
- Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*. US: American Psychological Association. <http://doi.org/10.1037/0022-3514.80.3.501>
- Ellis, C. (2013). Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics. *British Journal of Educational Technology*, 44(4), 662–664. <http://doi.org/10.1111/bjet.12028>
- Elo, A. E. (1978). *The rating of chessplayers, past and present* (Vol. 3). Batsford London.
- Essalmi, F., Ayed, L. J. Ben, Jemni, M., Kinshuk, & Graf, S. (2010). A fully personalization strategy of E-learning scenarios. *Computers in Human Behavior*, 26(4), 581–591. <http://doi.org/10.1016/j.chb.2009.12.010>
- Farrell, T., & Rushby, N. (2016, January 1). Assessment and learning technologies: An overview. *British Journal of Educational Technology*. <http://doi.org/10.1111/bjet.12348>
- Ferguson, R. (2012). *The state of learning analytics in 2012: A review and future challenges*. Knowledge Media Institute, Technical Report KMI-2012, 1, 2012.
- Finkelman, M. D., Kim, W., Weissman, A., & Cook, R. J. (2014). Cognitive Diagnostic Models and Computerized Adaptive Testing: Two New Item-Selection Methods That Incorporate Response Times. *Journal of Computerized Adaptive Testing*, 2, 59–76. <http://doi.org/10.7333/1412-0204059>
- Finney, S. J., Barry, C. L., Jeanne Horst, S., & Johnston, M. M. (2018). Exploring profiles of academic help seeking: A mixture modeling approach. *Learning and Individual Differences*, 61, 158–171. <http://doi.org/10.1016/j.lindif.2017.11.011>
- Fishburn, P. C. (1973). *The Theory of Social Choice*. Princeton University Press.
- Fiss, P. C. (2011). Building better causal theories: A fuzzy set approach to typologies in organization research. *Academy of Management Journal*, 54(2), 393–420.
- Fitzpatrick, A. R. (1983). The Meaning of Content Validity. *Applied Psychological Measurement*, 7(1), 3–13. <http://doi.org/10.1177/014662168300700102>
- Fletcher, A., & Shaw, G. (2012). How does student-directed assessment affect learning? Using assessment as a learning process. *International Journal of Multiple Research Approaches*, 6(3), 245–263. <http://doi.org/10.5172/mra.2012.6.3.245>
- Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, 18(1), 39–50.
- Fournier, H., Kop, R., & Sitlia, H. (2011). The Value of Learning Analytics to Networked Learning on a Personal Learning Environment. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 104–109). New York, NY, USA: ACM. <http://doi.org/10.1145/2090116.2090131>
- Furnham, A., Christopher, A., Garwood, J., & Martin, N. G. (2008). Ability, demography, learning style, and personality trait correlates of student preference for assessment method. *Educational Psychology*, 28(1), 15–27. <http://doi.org/10.1080/01443410701369138>
- Gama, J., Sebastião, R., & Rodrigues, P. P. (2009). Issues in Evaluation of Stream Learning Algorithms. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 329–338). New York, NY, USA: ACM. <http://doi.org/10.1145/1557019.1557060>

- Gartrell, M., Xing, X., Lv, Q., Beach, A., Han, R., Mishra, S., & Seada, K. (2010). Enhancing Group Recommendation by Incorporating Social Relationship Interactions. In *Proceedings of the 16th ACM International Conference on Supporting Group Work* (pp. 97–106). New York, NY, USA: ACM. <http://doi.org/10.1145/1880071.1880087>
- Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71. <http://doi.org/10.1007/s11528-014-0822-x>
- Geisser, S. (1974). A Predictive Approach to the Random Effect Model. *Biometrika*, *61*(1), 101–107.
- Gellatly, I. R. (1996). Conscientiousness and task performance: Test of cognitive process model. *Journal of Applied Psychology*, *81*(5), 474–482. <http://doi.org/10.1037/0021-9010.81.5.474>
- Ghorbandordinejad, F., & Ahmadabad, R. M. (2016). Examination of the Relationship Between Autonomy and English Achievement as Mediated by Foreign Language Classroom Anxiety. *Journal of Psycholinguistic Research*, *45*(3), 739–752. <http://doi.org/10.1007/s10936-015-9371-5>
- Giesbers, B., Rienties, B., Tempelaar, D., & Gijsselaers, W. (2013). Investigating the relations between motivation, tool use, participation, and performance in an e-learning course using web-videoconferencing. *Computers in Human Behavior*, *29*(1), 285–292. <http://doi.org/10.1016/j.chb.2012.09.005>
- Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351. <http://doi.org/http://dx.doi.org/10.1016/j.compedu.2011.06.004>
- Gollwitzer, P. M., Sheeran, P., Michalski, V., & Seifert, A. E. (2009). When Intentions Go Public: Does Social Reality Widen the Intention-Behavior Gap? *Psychological Science*, *20*(5), 612–618. <http://doi.org/10.1111/j.1467-9280.2009.02336.x>
- Gong, Y., Beck, J. E., & Heffernan, N. T. (2011). How to Construct More Accurate Student Models: Comparing and Optimizing Knowledge Tracing and Performance Factor Analysis. *Int. J. Artif. Intell. Ed.*, *21*(1–2), 27–45. <http://doi.org/10.3233/JAI-2011-016>
- Gong, Y., Beck, J. E., & Ruiz, C. (2012). Modeling Multiple Distributions of Student Performances to Improve Predictive Accuracy. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *User Modeling, Adaptation, and Personalization* (pp. 102–113). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Gonzalez-Brenes, J., Huang, Y., & Brusilovsky, P. (2014). General Features in Knowledge Tracing to Model Multiple Subskills, Temporal Item Response Theory, and Expert Knowledge. In *The 7th International Conference on Educational Data Mining* (pp. 84–91).
- Gowda, S. M., Rowe, J. P., de Baker, R. S. J., Chi, M., & Koedinger, K. R. (2011). Improving Models of Slipping, Guessing, and Moment-By-Moment Learning with Estimates of Skill Difficulty. In *Educational Data Mining*.
- Graf, S., & Bekele, R. (2006). Forming Heterogeneous Groups for Intelligent Collaborative Learning Systems with Ant Colony Optimization. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006. Proceedings* (pp. 217–226). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/11774303_22
- Graziano, G. W., & Eisenberg, H. N. (1997). Agreeableness: a dimension of personality. In R. Hogan, J. Johnston, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 795–824). San Diego, CA: Academic Press.
- Guay, F., Boggiano, A. K., & Vallerand, R. J. (2001). Autonomy Support, Intrinsic Motivation, and

- Perceived Competence: Conceptual and Empirical Linkages. *Personality and Social Psychology Bulletin*, 27(6), 643–650. <http://doi.org/10.1177/0146167201276001>
- Guo, W. W. (2010). Incorporating statistical and neural network approaches for student course satisfaction analysis and prediction. *Expert Systems with Applications*, 37(4), 3358–3365. <http://doi.org/https://doi.org/10.1016/j.eswa.2009.10.014>
- Guruler, H., Istanbulu, A., & Karahasan, M. (2010). A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education*, 55(1), 247–254. <http://doi.org/https://doi.org/10.1016/j.compedu.2010.01.010>
- Gutman, L. M., & Schoon, I. (2013). *The impact of non-cognitive skills on outcomes for young people: Literature review*. London: University of London, Institute of Education.
- Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, 3, 1157–1182.
- Gvozdenko, E., & Chambers, D. (2007). Beyond test accuracy: Benefits of measuring response time in computerised testing. *Australasian Journal of Educational Technology*, 23(4), 542–558.
- Halisch, F., & Heckhausen, H. (1977). Search for feedback information and effort regulation during task performance. *Journal of Personality and Social Psychology*, 35(10), 724–733. <http://doi.org/10.1037/0022-3514.35.10.724>
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <http://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hamdan, A., Nasir, R., Rozainee, W., & Sulaiman, W. S. (2013). Time management does not matter for academic achievement unless you can cope. In *International Proceedings of Economics Development and Research* (Vol. 78, pp. 22–26).
- Hao, Q., Barnes, B., Wright, E., & Branch, R. M. (2016). The influence of achievement goals on online help seeking of computer science students. *British Journal of Educational Technology*, 48(6), 1273–1283. <http://doi.org/10.1111/bjet.12499>
- Hao, Q., Wright, E., Barnes, B., & Branch, R. M. (2016). What are the most important prediction of computer science students' online help-seeking behaviors? *Computers in Human Behavior*, 62, 467–474. <http://doi.org/10.1016/j.chb.2016.04.016>
- Harlen, W., & James, M. (1997). Assessment and Learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4(3), 365–379. <http://doi.org/10.1080/0969594970040304>
- Hartnett, M., George, A. St., & Dron, J. (2011). Examining motivation in online distance learning environments: Complex, multifaceted and situation-dependent. *The International Review of Research in Open and Distributed Learning*, 12(6), 20–38. Retrieved from <http://www.irrodl.org/index.php/irrodl/article/view/1030>
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <http://doi.org/10.3102/003465430298487>
- Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6), 1251–1271.
- Hawkins, W. J., Heffernan, N. T., & Baker, R. S. J. D. (2014). Learning Bayesian Knowledge Tracing Parameters with a Knowledge Heuristic and Empirical Probabilities. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 150–155). Cham: Springer International Publishing.
- Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation* (2nd ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

- He, J., & Chu, W. W. (2010). A Social Network-Based Recommender System (SNRS). In N. Memon, J. J. Xu, D. L. Hicks, & H. Chen (Eds.), *Data Mining for Social Network Data* (pp. 47–74). Boston, MA: Springer US. http://doi.org/10.1007/978-1-4419-6287-4_4
- He, W. (2013). Examining Students' Online Interaction in a Live Video Streaming Environment Using Data Mining and Text Mining. *Comput. Hum. Behav.*, *29*(1), 90–102. <http://doi.org/10.1016/j.chb.2012.07.020>
- He, W., Diao, Q., & Hauser, C. (2014). A Comparison of Four Item-Selection Methods for Severely Constrained CATs. *Educational and Psychological Measurement*, *74*(4), 677–696. <http://doi.org/10.1177/0013164413517503>
- Heerde, J. A., & Hemphill, S. A. (2018). Examination of associations between informal help-seeking behavior, social support, and adolescent psychosocial outcomes: A meta-analysis. *Developmental Review*, *47*, 44–62. <http://doi.org/10.1016/j.dr.2017.10.001>
- Hodges, C. B., & Kim, C. (2010). Email, Self-Regulation, Self-Efficacy, and Achievement in a College Online Mathematics Course. *Journal of Educational Computing Research*, *43*(2), 207–223. <http://doi.org/10.2190/EC.43.2.d>
- Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, *58*(301), 13–30. <http://doi.org/10.1080/01621459.1963.10500830>
- Holec, H. (1981). *Autonomy in foreign language learning*. Oxford: Pergamon.
- Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicológica*, *21*(1–2), 175–189.
- Hoye, G. Van, & Lootens, H. (2013). Coping with unemployment: Personality, role demands, and time structure. *Journal of Vocational Behavior*, *82*(2), 85–95. <http://doi.org/https://doi.org/10.1016/j.jvb.2013.01.004>
- Hsiao, I.-H., Bakalov, F., Brusilovsky, P., & König-Ries, B. (2011). Open Social Student Modeling: Visualizing Student Models with Parallel Introspective Views. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), *User Modeling, Adaption and Personalization: 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings* (pp. 171–182). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-22362-4_15
- Hu, P., & Zhang, J. (2017). A pathway to learner autonomy: a self-determination theory perspective. *Asia Pacific Education Review*, *18*(1), 147–157. <http://doi.org/10.1007/s12564-016-9468-z>
- Hu, X., Craig, S. D., Bargagliotti, A. E., Graesser, A. C., Okwumabua, T., Anderson, C., ... Sterbinsky, A. (2012). The Effects of a Traditional and Technology-based After-school Setting on 6th Grade Student's Mathematics Skills. *Journal of Computers in Mathematics and Science Teaching*, *31*(1), 17–38.
- Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, 133–145. <http://doi.org/10.1016/j.compedu.2012.08.015>
- Huang, Y.-M. (2017). Exploring students' acceptance of team messaging services: The roles of social presence and motivation. *British Journal of Educational Technology*, *48*(4), 1047–1061. <http://doi.org/10.1111/bjet.12468>
- Huet, N., Escribe, C., Dupeyrat, C., & Sakdavong, J.-C. (2011). The influence of achievement goals and perceptions of online help on its actual use in an interactive learning environment. *Computers in Human Behavior*, *27*(1), 413–420. <http://doi.org/10.1016/J.CHB.2010.09.003>
- Huet, N., Moták, L., & Sakdavong, J. C. (2016). Motivation to seek help and help efficiency in students who failed in an initial task. *Computers in Human Behavior*, *63*, 584–593.

<http://doi.org/10.1016/j.CHB.2016.05.059>

- Hulten, G., Spencer, L., & Domingos, P. (2001). Mining Time-changing Data Streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 97–106). New York, NY, USA: ACM.
<http://doi.org/10.1145/502512.502529>
- Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review*. US: American Psychological Association. <http://doi.org/10.1037/0033-295X.91.2.153>
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297–307. <http://doi.org/10.1093/biomet/76.2.297>
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.
<http://doi.org/10.1016/j.ijforecast.2006.03.001>
- Ilias O Pappas, Patrick Mikalef, M. N. G., & Pavlou, P. A. (2017). Value co-creation and trust in social commerce: An fsQCA approach. In *Proceedings of the 25th European Conference on Information Systems (ECIS)* (pp. 2153–2168).
- Jalava, N., Joensen, J. S., & Pellas, E. (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, 115, 161–196.
<http://doi.org/https://doi.org/10.1016/j.jebo.2014.12.004>
- Järvelin, K., & Kekäläinen, J. (2002). Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.*, 20(4), 422–446. <http://doi.org/10.1145/582415.582418>
- Jeong, H., & Biswas, G. (2008). Mining student behavior models in Learning-by-teaching environments. In *Proceedings of the 1st International Conference on Educational Data Mining, EDM 2008*, (pp. 67–76). Montreal, Canada.
- Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). Awareness Is Not Enough: Pitfalls of Learning Analytics Dashboards in the Educational Practice. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data Driven Approaches in Digital Education* (pp. 82–96). Cham: Springer International Publishing.
- Jo, I.-H., Kim, D., & Yoon, M. (2015). Constructing Proxy Variables to Measure Adult Learners' Time Management Strategies in LMS. *Journal of Educational Technology & Society*, 18(3), 214–225.
- John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research*, 2nd ed. (pp. 102–138). New York, NY, US: Guilford Press.
- Johns, J., Mahadevan, S., & Woolf, B. (2006). Estimating Student Proficiency Using an Item Response Theory Model. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems* (pp. 473–480). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Johnson, L., Adams, S., & Cummins, M. (2012). *The NMC Horizon Report: 2012 Higher Education Edition*.
- Joksimović, S., Gašević, D., Loughin, T. M., Kovanović, V., & Hatala, M. (2015). Learning at distance: Effects of interaction traces on academic achievement. *Computers & Education*, 87, 204–217. <http://doi.org/https://doi.org/10.1016/j.compedu.2015.07.002>
- Joosten-ten Brinke, D., van Bruggen, J., Hermans, H., Burgers, J., Giesbers, B., Koper, R., & Latour, I. (2007). Modeling Assessment for Re-use of Traditional and New Types of Assessment. *Comput. Hum. Behav.*, 23(6), 2721–2741. <http://doi.org/10.1016/j.chb.2006.08.009>

- Jourdan, Z., Rainer, R. K., & Marshall, T. E. (2008). Business Intelligence: An Analysis of the Literature. *Information Systems Management*, 25(2), 121–131. <http://doi.org/10.1080/10580530801941512>
- Judge, T. A., & Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology*. Judge, Timothy A.: U Florida, Warrington Coll of Business, Dept of Management, 211 D Stuzin Hall, Gainesville, FL, US, 32611-7165, tjudge@ufl.edu: American Psychological Association. <http://doi.org/10.1037/0021-9010.87.4.797>
- Kahraman, N., Cuddy, M. M., & Clauser, B. E. (2013). Modeling Pacing Behavior and Test Speededness Using Latent Growth Curve Models. *Applied Psychological Measurement*, 37(5), 343–360. <http://doi.org/10.1177/0146621613477236>
- Kanfer, R., & Heggestad, E. D. (1997). Motivational traits and skills: A person-centered approach to work motivation. *Research in Organizational Behavior*, 19, 1–56.
- Karabenick, S. A. (2003, January 1). Seeking help in large college classes: A person-centered approach. *Contemporary Educational Psychology*. Academic Press. [http://doi.org/10.1016/S0361-476X\(02\)00012-7](http://doi.org/10.1016/S0361-476X(02)00012-7)
- Karabenick, S. A. (2011). Classroom and technology-supported help seeking: The need for converging research paradigms. *Learning and Instruction*, 21(2), 290–296. <http://doi.org/10.1016/J.LEARNINSTRUC.2010.07.007>
- Karabenick, S. A., & Berger, J.-L. (2013). Help seeking as a self-regulated learning strategy. In *Applications of self-regulated learning across diverse disciplines: A tribute to Barry J. Zimmerman*. (pp. 237–261). Charlotte, NC, US: IAP Information Age Publishing.
- Käser, T., Klingler, S., Schwing, A. G., & Gross, M. (2017). Dynamic Bayesian Networks for Student Modeling. *IEEE Transactions on Learning Technologies*, 10(4), 450–462. <http://doi.org/10.1109/TLT.2017.2689017>
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons.
- Kelly, W. E., & Johnson, J. L. (2005). Time Use Efficiency and the Five-Factor Model of Personality. *Education*, 125(3), 511–515.
- Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How deep is knowledge tracing? In *Educational Data Mining*.
- Khajah, M. M., Huang, Y., González-Brenes, J. P., Mozer, M. C., & Brusilovsky, P. (2014). Integrating knowledge tracing and item response theory: A tale of two frameworks. In *CEUR Workshop Proceedings* (Vol. 1181, pp. 7–15).
- Khribi, M. K., Jemni, M., & Nasraoui, O. (2009). Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. *Educational Technology & Society*, 12(4), 30–42.
- Kim, N., Smith, M. J., & Maeng, K. (2008). Assessment in online distance education: a comparison of three online programs at a university. *Online Journal of Distance Learning Administration*, 11(1). Kitsantas, A. (2002). Test Preparation and Performance: A Self-Regulatory Analysis. *The Journal of Experimental Education*, 70(2), 101–113. <http://doi.org/10.1080/00220970209599501>
- Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, 104, 18–33. <http://doi.org/http://dx.doi.org/10.1016/j.compedu.2016.10.001>
- Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In *Proceedings of the Third*

- International Conference on Learning Analytics and Knowledge* (pp. 170–179). New York, NY, USA: ACM. <http://doi.org/10.1145/2460296.2460330>
- Klašnja-Milićević, A., Vesin, B., Ivanović, M., & Budimac, Z. (2011). E-Learning Personalization Based on Hybrid Recommendation Strategy and Learning Style Identification. *Comput. Educ.*, *56*(3), 885–899. <http://doi.org/10.1016/j.compedu.2010.11.001>
- Kleinberg, J. (2016). Temporal Dynamics of On-Line Information Streams. In M. Garofalakis, J. Gehrke, & R. Rastogi (Eds.), *Data Stream Management: Processing High-Speed Data Streams* (pp. 221–238). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-28608-0_11
- Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. <http://doi.org/https://doi.org/10.1016/j.compedu.2011.02.003>
- Knight, D., Brozina, C., & Novoselich, B. (2016). An Investigation of First-Year Engineering Student and Instructor Perspectives of Learning Analytics Approaches. *Journal of Learning Analytics*, *3*(3), 215–238.
- Knight, S., Shum, S. B., & Littleton, K. (2014). Epistemology, assessment, pedagogy: where learning meets analytics in the middle space. *Journal of Learning Analytics*, *1*(2), 23–47.
- Kompan, M., & Bielikova, M. (2016). Enhancing existing e-learning systems by single and group recommendations. *International Journal of Continuing Engineering Education and Life Long Learning*, *26*(4), 386–404. <http://doi.org/10.1504/IJCEELL.2016.080980>
- Kop, R., Fournier, H., & Mak, J. (2011). A pedagogy of abundance or a pedagogy to support human beings? Participant support on massive open online courses. *The International Review of Research in Open and Distributed Learning*, *12*(7), 74–93.
- Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, *42*(8), 30–37. <http://doi.org/10.1109/MC.2009.263>
- Kormos, J., & Csizér, K. (2014). The Interaction of Motivation, Self-Regulatory Strategies, and Autonomous Learning Behavior in Different Learner Groups. *TESOL Quarterly*, *48*(2), 275–299. <http://doi.org/10.1002/tesq.129>
- Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 47–58.
- Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education*, *27*, 74–89. <http://doi.org/https://doi.org/10.1016/j.iheduc.2015.06.002>
- Krakauer, J. W., & Mazzoni, P. (2011). Human sensorimotor learning: adaptation, skill, and beyond. *Current Opinion in Neurobiology*, *21*(4), 636–644. <http://doi.org/10.1016/j.conb.2011.06.012>
- Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in Human Behavior*, *19*(3), 335–353. [http://doi.org/10.1016/S0747-5632\(02\)00057-2](http://doi.org/10.1016/S0747-5632(02)00057-2)
- Kurucay, M., & Inan, F. A. (2017). Examining the effects of learner-learner interactions on satisfaction and learning in an online undergraduate course. *Computers & Education*, *115*(Supplement C), 20–37. <http://doi.org/10.1016/j.compedu.2017.06.010>
- Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40*(5), 1847–1857.

- <http://doi.org/https://doi.org/10.1016/j.eswa.2012.09.017>
- Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement, 72*(1), 159–175. <http://doi.org/10.1177/0013164411411296>
- Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359–379.
- Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education, 2*(1), 8. <http://doi.org/10.1186/s40536-014-0008-1>
- Lee, Y. H., & Haberman, S. J. (2016). Investigating Test-Taking Behaviors Using Timing and Process Data. *International Journal of Testing, 16*(3), 240–267. <http://doi.org/10.1080/15305058.2015.1085385>
- León, J., Núñez, J. L., & Liew, J. (2015). Self-determination and STEM education: Effects of autonomy, motivation, and self-regulated learning on high school math achievement. *Learning and Individual Differences, 43*, 156–163. <http://doi.org/10.1016/j.lindif.2015.08.017>
- Leong, C. K., Lee, Y. H., & Mak, W. K. (2012). Mining sentiments in SMS texts for teaching evaluation. *Expert Systems with Applications, 39*(3), 2584–2589. <http://doi.org/https://doi.org/10.1016/j.eswa.2011.08.113>
- Leony, D., Muñoz-Merino, P. J., Pardo, A., & Kloos, C. D. (2013). Provision of awareness of learners' emotions through visualizations in a computer interaction-based environment. *Expert Systems with Applications, 40*(13), 5093–5100. <http://doi.org/https://doi.org/10.1016/j.eswa.2013.03.030>
- Levy, S. T., & Wilensky, U. (2011). Mining Students' Inquiry Actions for Understanding of Complex Systems. *Comput. Educ., 56*(3), 556–573. <http://doi.org/10.1016/j.compedu.2010.09.015>
- Lewis, T., & Vialleton, E. (2011). The notions of control and consciousness in learner autonomy and self-regulated learning: a comparison and critique. *Innovation in Language Learning and Teaching, 5*(2), 205–219. <http://doi.org/10.1080/17501229.2011.577535>
- Li, N., Cohen, W. W., Koedinger, K. R., & Matsuda, N. (2011). A machine learning approach for automatic student model discovery. In *Proceedings of the 4th International Conference on Educational Data Mining, EDM 2011* (pp. 31–40). Eindhoven, the Netherlands.
- Lin, C. F., Yeh, Y., Hung, Y. H., & Chang, R. I. (2013). Data mining for providing a personalized learning path in creativity: An application of decision trees. *Computers & Education, 68*, 199–210. <http://doi.org/https://doi.org/10.1016/j.compedu.2013.05.009>
- Lin, C., Shen, S., & Chi, M. (2016). Incorporating Student Response Time and Tutor Instructional Interventions into Student Modeling. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (pp. 157–161). New York, NY, USA: ACM. <http://doi.org/10.1145/2930238.2930291>
- Lin, F.-R., Hsieh, L.-S., & Chuang, F.-T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education, 52*(2), 481–495. <http://doi.org/10.1016/j.compedu.2008.10.005>
- Little, D. (1991). *Learner autonomy 1: Definitions, issues and problems*. Dublin: Authentik.
- Little, D. (1995). Learning as dialogue: The dependence of learner autonomy on teacher autonomy. *System, 23*(2), 175–181. [http://doi.org/10.1016/0346-251X\(95\)00006-6](http://doi.org/10.1016/0346-251X(95)00006-6)
- Littlewood, W. (1996). "Autonomy": An anatomy and a framework. *System, 24*(4), 427–435.

[http://doi.org/10.1016/S0346-251X\(96\)00039-5](http://doi.org/10.1016/S0346-251X(96)00039-5)

- Littlewood, W. (1999). Defining and developing autonomy in East Asian contexts. *Applied Linguistics*, 20(1), 71–94. <http://doi.org/10.1093/applin/20.1.71>
- Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. In *Proceedings of the 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD-98)* (pp. 80–86). AAAI Press.
- Liu, O. L., Rios, J. A., & Borden, V. (2015). The Effects of Motivational Instruction on College Students' Performance on Low-Stakes Assessment. *Educational Assessment*, 20(2), 79–94. <http://doi.org/10.1080/10627197.2015.1028618>
- Lo, C. C. (2010). How Student Satisfaction Factors Affect Perceived Learning. *Journal of Scholarship of Teaching and Learning*, 10(1), 47–54.
- Loyens, S. M. M., Magda, J., & Rikers, R. M. J. P. (2008). Self-Directed Learning in Problem-Based Learning and its Relationships with Self-Regulated Learning. *Educational Psychology Review*, 20(4), 411–427. <http://doi.org/10.1007/s10648-008-9082-7>
- Lu, H., Hu, Y., Gao, J., & Kinshuk. (2016). The effects of computer self-efficacy, training satisfaction and test anxiety on attitude and performance in computerized adaptive testing. *Computers & Education*, 100, 45–55. <http://dx.doi.org/10.1016/j.compedu.2016.04.012>
- Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender System Application Developments. *Decis. Support Syst.*, 74(C), 12–32. <http://doi.org/10.1016/j.dss.2015.03.008>
- Luke, C. L. (2006). Fostering Learner Autonomy in a Technology-Enhanced, Inquiry-Based Foreign Language Classroom. *Foreign Language Annals*, 39(1), 71–86. <http://doi.org/10.1111/j.1944-9720.2006.tb02250.x>
- Lykourantzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., & Loumos, V. (2009). Early and Dynamic Student Achievement Prediction in e-Learning Courses Using Neural Networks. *J. Am. Soc. Inf. Sci. Technol.*, 60(2), 372–380. <http://doi.org/10.1002/asi.v60:2>
- Lykourantzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout Prediction in e-Learning Courses Through the Combination of Machine Learning Techniques. *Comput. Educ.*, 53(3), 950–965. <http://doi.org/10.1016/j.compedu.2009.05.010>
- Macan, T. H., Shahani, C., Dipboye, R. L., & Phillips, A. P. (1990). College students' time management: Correlations with academic performance and stress. *Journal of Educational Psychology*, 82(4), 760–768.
- MacCann, C., Fogarty, G. J., & Roberts, R. D. (2012). Strategies for success in education: Time management is more important for part-time than full-time community college students. *Learning and Individual Differences*, 22(5), 618–623. <http://doi.org/10.1016/j.lindif.2011.09.015>
- Macfadyen, L. P., & Dawson, S. (2010). Mining LMS Data to Develop an “Early Warning System” for Educators: A Proof of Concept. *Comput. Educ.*, 54(2), 588–599. <http://doi.org/10.1016/j.compedu.2009.09.008>
- MacNeill, S., Campbell, L. M., & Hawksey, M. (2014). Analytics for Education. *Journal of Interactive Media in Education*, 1–12.
- Maehr, M. L., & Meyer, H. A. (1997). Understanding Motivation and Schooling: Where We've Been, Where We Are, and Where We Need to Go. *Educational Psychology Review*, 9(4), 371–409. <http://doi.org/10.1023/A:1024750807365>
- Magnusson, D. (1998). The logic and implications of a person-oriented approach. In *Methods and*

- models for studying the individual.* (pp. 33–64). Thousand Oaks, CA, US: Sage Publications, Inc.
- Mäkitalo-Siegl, K., Kohnle, C., & Fischer, F. (2011). Computer-supported collaborative inquiry learning and classroom scripts: Effects on help-seeking processes and learning outcomes. *Learning and Instruction, 21*(2), 257–266. <http://doi.org/10.1016/j.learninstruc.2010.07.001>
- Manouselis, N., Drachsler, H., Verbert, K., & Duval, E. (2013). *Recommender Systems for Learning*. New York, NY: Springer.
- Martinez-Maldonado, R., Schneider, B., Charleer, S., Shum, S. B., Klerkx, J., & Duval, E. (2016). Interactive Surfaces and Learning Analytics: Data, Orchestration Aspects, Pedagogical Uses and Challenges. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 124–133). New York, NY, USA: ACM. <http://doi.org/10.1145/2883851.2883873>
- Masthoff, J. (2011). Group Recommender Systems: Combining Individual Models. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 677–702). Boston, MA: Springer US. http://doi.org/10.1007/978-0-387-85820-3_21
- Masthoff, J. (2015). Group Recommender Systems: Aggregation, Satisfaction and Group Attributes. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 743–776). Boston, MA: Springer US. http://doi.org/10.1007/978-1-4899-7637-6_22
- Masthoff, J., & Gatt, A. (2006). In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems. *User Modeling and User-Adapted Interaction, 16*(3), 281–319. <http://doi.org/10.1007/s11257-006-9008-3>
- Mathews, M., Mitrović, T., & Thomson, D. (2008). Analysing High-Level Help-Seeking Behaviour in ITSs. In W. Nejdl, J. Kay, P. Pu, & E. Herder (Eds.), *Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 312–315). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Mavroudi, A., Giannakos, M., & Krogstie, J. (2018). Supporting adaptive learning pathways through the use of learning analytics: developments, challenges and future opportunities. *Interactive Learning Environments, 26*(2), 206–220. <http://doi.org/10.1080/10494820.2017.1292531>
- McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis. *Research Quarterly for Exercise and Sport, 60*(1), 48–58. <http://doi.org/10.1080/02701367.1989.10607413>
- McCalla, G. I. (1992). The Central Importance of Student Modelling to Intelligent Tutoring. In E. Costa (Ed.), *New Directions for Intelligent Tutoring Systems* (pp. 107–131). Berlin, Heidelberg: Springer Berlin Heidelberg.
- McCardle, L., & Hadwin, A. F. (2015). Using multiple, contextualized data sources to measure learners' perceptions of their self-regulated learning. *Metacognition and Learning, 10*(1), 43–75. <http://doi.org/10.1007/s11409-014-9132-0>
- McCarthy, J. F., & Anagnost, T. D. (1998). MusicFX: An Arbiter of Group Preferences for Computer Supported Collaborative Workouts. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (pp. 363–372). New York, NY, USA: ACM. <http://doi.org/10.1145/289444.289511>
- McCarthy, K., Salamó, M., Coyle, L., McGinty, L., Smyth, B., & Nixon, P. (2006). Group Recommender Systems: A Critiquing Based Approach. In *Proceedings of the 11th International Conference on Intelligent User Interfaces* (pp. 267–269). New York, NY, USA: ACM. <http://doi.org/10.1145/1111449.1111506>

- McCrae, R. R. (1996). Social consequences of experiential openness. *Psychological Bulletin*. US: American Psychological Association. <http://doi.org/10.1037/0033-2909.120.3.323>
- McCrae, R. R., & John, P. O. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, *60*(2), 175–215. <http://doi.org/10.1111/j.1467-6494.1992.tb00970.x>
- McDevitt, B. (1997). Learner autonomy and the need for learner training. *The Language Learning Journal*, *16*(1), 34–39. <http://doi.org/10.1080/09571739785200251>
- McKelvey, R. D., McLennan, A. M., & Turocy, T. L. (2006). *Gambit: Software Tools for Game Theory*.
- McMillan, J. H., & Hearn, J. (2008). Student Self-Assessment: The Key to Stronger Student Motivation and Higher Achievement. *Educational Horizons*, *87*(1), 40–49.
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, *1*(1), 64–69. <http://doi.org/10.1111/j.1467-9280.1990.tb00069.x>
- Mendel, J. M., & Korjani, M. M. (2012). Charles Ragin's fuzzy set qualitative comparative analysis (fsQCA) used for linguistic summarizations. *Information Sciences*, *202*, 1–23.
- Merceron, A., & Yacef, K. (2008). Interestingness measures for association rules in educational data. In J. B. de Baker T. Barnes (Ed.), *Proceedings of the 1st International Conference on Educational Data Mining, EDM 2008*, (pp. 57–66). Montreal, Canada.
- Michinov, N., Brunot, S., Bohec, O. Le, Juhel, J., & Delaval, M. (2011). Procrastination, participation, and performance in online learning environments. *Computers & Education*, *56*(1), 243–252. <http://doi.org/https://doi.org/10.1016/j.compedu.2010.07.025>
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29. <http://doi.org/10.1002/j.2333-8504.2003.tb01908.x>
- Mitchell, T. M. (1997). *Machine Learning* (1st ed.). New York, NY, USA: McGraw-Hill, Inc.
- Mitrovic, A. (2012). Fifteen years of constraint-based tutors: what we have achieved and where we are going. *User Modeling and User-Adapted Interaction*, *22*(1), 39–72. <http://doi.org/10.1007/s11257-011-9105-9>
- Mitrovic, A., & Martin, B. (2002). Evaluating the Effects of Open Student Models on Learning. In P. De Bra, P. Brusilovsky, & R. Conejo (Eds.), *Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 296–305). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Mitrovic, A., Ohlsson, S., & Barrow, D. K. (2013). The effect of positive feedback in a constraint-based intelligent tutoring system. *Computers & Education*, *60*(1), 264–272. <http://doi.org/https://doi.org/10.1016/j.compedu.2012.07.002>
- Moore, J. L., Dickson-Deane, C., & Galyen, K. (2011). e-Learning, online learning, and distance learning environments: Are they the same? *The Internet and Higher Education*, *14*(2), 129–135. <http://doi.org/https://doi.org/10.1016/j.iheduc.2010.10.001>
- Moridis, C. N., & Economides, A. A. (2009a). Mood recognition during online self-assessment tests. *IEEE Transactions on Learning Technologies*, *2*(1), 50–61.
- Moridis, C. N., & Economides, A. A. (2009b). Prediction of Student's Mood During an Online Test Using Formula-based and Neural Network-based Method. *Comput. Educ.*, *53*(3), 644–652. <http://doi.org/10.1016/j.compedu.2009.04.002>
- Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L. (2018). Modeling Learners' Cognitive and Affective States to Scaffold SRL in Open-Ended Learning Environments. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (pp. 131–138). New York, NY, USA: ACM.

<http://doi.org/10.1145/3209219.3209241>

- Murphy, K., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational and goal-oriented perspectives*. Newbury Park, CA: Sage.
- Mussweiler, T. (2003). Comparison processes in social judgment: mechanisms and consequences. *Psychological Review*, *110*(3), 472–489.
- Myerson, R. B. (1985). *Negotiation in Games: A Theoretical Overview*.
- Nagin, D. S. (2005). *Group-Based Modeling of Development*. Cambridge, MA: Harvard Press.
- Narciss, S., Proske, A., & Koerndle, H. (2007). Promoting self-regulated learning in web-based learning environments. *Computers in Human Behavior*, *23*(3), 1126–1144.
<http://doi.org/10.1016/j.chb.2006.10.006>
- Nash, J. (1951). Non-Cooperative Games. *Annals of Mathematics*, *54*(2), 286–295.
- Nedungadi, P., & Remya, M. S. (2015). Incorporating forgetting in the Personalized, Clustered, Bayesian Knowledge Tracing (PC-BKT) model. In *2015 International Conference on Cognitive Computing and Information Processing (CCIP)* (pp. 1–5).
<http://doi.org/10.1109/CCIP.2015.7100688>
- Nelson-Le Gall, S. (1985). Help-Seeking Behavior in Learning. *Review of Research in Education*, *12*, 55–90.
- New Media Consortium. (2018). *2018 Horizon Report Preview*.
- Newby, L., & Winterbottom, M. (2011). Can research homework provide a vehicle for assessment for learning in science lessons? *Educational Review*, *63*(3), 275–290.
<http://doi.org/10.1080/00131911.2011.560247>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. <http://doi.org/10.1080/03075070600572090>
- Nikou, S. A., & Economides, A. A. (2017a). Mobile-Based Assessment: Integrating acceptance and motivational factors into a combined model of Self-Determination Theory and Technology Acceptance. *Computers in Human Behavior*, *68*, 83–95.
<http://doi.org/10.1016/j.chb.2016.11.020>
- Nikou, S. A., & Economides, A. A. (2017b). Mobile-based assessment: Investigating the factors that influence behavioral intention to use. *Computers & Education*, *109*, 56–73.
<http://doi.org/http://dx.doi.org/10.1016/j.compedu.2017.02.005>
- Nonis, S. A., & Hudson, G. I. (2006). Academic Performance of College Students: Influence of Time Spent Studying and Working. *Journal of Education for Business*, *81*(3), 151–159.
<http://doi.org/10.3200/JOEB.81.3.151-159>
- Nortvedt, G. A. (2014). Assessment and learning. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 125–128. <http://doi.org/10.1080/0969594X.2013.852512>
- O'Connor, M., Cosley, D., Konstan, J. A., & Riedl, J. (2001). PolyLens: A Recommender System for Groups of Users. In W. Prinz, M. Jarke, Y. Rogers, K. Schmidt, & V. Wulf (Eds.), *ECSCW 2001: Proceedings of the Seventh European Conference on Computer Supported Cooperative Work 16--20 September 2001, Bonn, Germany* (pp. 199–218). Dordrecht: Springer Netherlands.
http://doi.org/10.1007/0-306-48019-0_11
- Ohlsson, S. (1994). Constraint-Based Student Modeling. In J. E. Greer & G. I. McCalla (Eds.), *Student Modelling: The Key to Individualized Knowledge-Based Instruction* (pp. 167–189). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Okoli, C., & Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of

- Information Systems Research. *Working Papers on Information Systems*, 10(26), 1–51.
- Ordanini, A., Parasuraman, A., & Rubera, G. (2014). When the recipe is more important than the ingredients a Qualitative Comparative Analysis (QCA) of service innovation configurations. *Journal of Service Research*, 17(2), 134–149. <http://doi.org/10.1177/1094670513513337>
- Ortner, M. T., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, 27(3), 157–163.
- Oxford, R. L. (2008). Hero with a thousand faces: Learning autonomy, learning strategies and learning tactics in independent language learning. In S. Hurd & T. Lewis (Eds.), *Language learning strategies in independent settings* (pp. 41–63). Clevedon: Multilingual Matters.
- Oxford, R. L. (2015). Expanded perspectives on autonomous learners. *Innovation in Language Learning and Teaching*, 9(1), 58–71. <http://doi.org/10.1080/17501229.2014.995765>
- Pajares, F., Cheong, Y. F., & Oberman, P. (2004). Psychometric Analysis of Computer Science Help-Seeking Scales. *Educational and Psychological Measurement*, 64(3), 496–513. <http://doi.org/10.1177/0013164403258447>
- Papamitsiou, Z., & Economides, A. A. (2013). Towards the alignment of computer-based assessment outcome with learning goals: The LAERS architecture. In *2013 IEEE Conference on e-Learning, e-Management and e-Services, IC3e 2013* (pp. 13–17). <http://doi.org/10.1109/IC3e.2013.6735958>
- Papamitsiou, Z., & Economides, A. A. (2014a). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society*, 17(4), 49–64.
- Papamitsiou, Z., & Economides, A. A. (2014b). Students' perception of performance vs. actual performance during computer-based testing: a temporal approach. In *8th International Technology, Education & Development Conference* (pp. 401–411). IATED.
- Papamitsiou, Z., & Economides, A. A. (2014c). Temporal Learning Analytics for Adaptive Assessment. *Journal of Learning Analytics*, 1(3), 165–168.
- Papamitsiou, Z., & Economides, A. A. (2014d). The effect of personality traits on students' performance during computer-based testing: A study of the big five inventory with temporal learning analytics. In *Proceedings - IEEE 14th International Conference on Advanced Learning Technologies, ICALT 2014* (pp. 378–382). <http://doi.org/10.1109/ICALT.2014.113>
- Papamitsiou, Z., & Economides, A. A. (2015). A temporal estimation of students' on-task mental effort and its effect on students' performance during computer based testing. In *Proceedings of 2015 International Conference on Interactive Collaborative Learning, ICL 2015*. <http://doi.org/10.1109/ICL.2015.7318194>
- Papamitsiou, Z., & Economides, A. A. (2016). *An Assessment Analytics Framework (AAF) for Enhancing Students' Progress. Formative Assessment, Learning Data Analytics and Gamification: In ICT Education*. <http://doi.org/10.1016/B978-0-12-803637-2.00007-5>
- Papamitsiou, Z., & Economides, A. A. (2017). Exhibiting achievement behavior during computer-based testing: What temporal trace data and personality traits tell us? *Computers in Human Behavior*, 75, 423–438. <http://doi.org/10.1016/j.chb.2017.05.036>
- Papamitsiou, Z., & Economides, A. A. (2018). Can'T Get More Satisfaction?: Game-theoretic Group-recommendation of Educational Resources. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 409–416). New York, NY, USA: ACM. <http://doi.org/10.1145/3170358.3170371>
- Papamitsiou, Z., Economides, A. A., Pappas, I. O., & Giannakos, M. N. (2018). Explaining Learning Performance Using Response-time, Self-regulation and Satisfaction from Content: An fsQCA

- Approach. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 181–190). New York, NY, USA: ACM.
<http://doi.org/10.1145/3170358.3170397>
- Papamitsiou, Z., Karapistoli, E., & Economides, A. A. (2016). Applying classification techniques on temporal trace data for shaping student behavior models. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* (Vol. 25–29–Apr, pp. 299–303). <http://doi.org/10.1145/2883851.2883926>
- Papamitsiou, Z., Terzis, V., & Economides, A. A. (2014). Temporal learning analytics for computer based testing. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14* (pp. 31–35). <http://doi.org/10.1145/2567574.2567609>
- Pappas, I. O., Giannakos, M. N., Jaccheri, L., & Sampson, D. G. (2017). Assessing Student Behavior in Computer Science Education with an fsQCA Approach: The Role of Gains and Barriers. *ACM Transactions on Computing Education (TOCE)*, 17(2), Article No. 10.
- Pappas, I. O., Giannakos, M. N., & Sampson, D. G. (2016). Making Sense of Learning Analytics with a Configurational Approach. In *Proceedings of the workshop on Smart Environments and Analytics in Video-Based Learning (SE@ VBL)* (pp. 42–52).
- Pappas, I. O., Giannakos, M. N., & Sampson, D. G. (2017). Fuzzy set analysis as a means to understand users of 21st-century learning systems: The case of mobile learning and reflections on learning analytics research. *Computers in Human Behavior*.
<http://doi.org/https://doi.org/10.1016/j.chb.2017.10.010>
- Pappas, I. O., Papavlasopoulou, S., Giannakos, M. N., & Sampson, D. G. (2017). An Exploratory Study on the Influence of Cognitive and Affective Characteristics in Programming-Based Making Activities. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)* (pp. 507–511). <http://doi.org/10.1109/ICALT.2017.78>
- Pardos, Z. A., Baker, R. S. J. D., San Pedro, M. O. C. Z., Gowda, S. M., & Gowda, S. M. (2013). Affective States and State Tests: Investigating How Affect Throughout the School Year Predicts End of Year Learning Outcomes. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 117–124). New York, NY, USA: ACM.
<http://doi.org/10.1145/2460296.2460320>
- Pardos, Z. A., Gowda, S. M., Baker, R. S. J. d., & Heffernan, N. T. (2012). The Sum is Greater Than the Parts: Ensembling Models of Student Knowledge in Educational Software. *SIGKDD Explor. Newsl.*, 13(2), 37–44. <http://doi.org/10.1145/2207243.2207249>
- Pardos, Z. A., & Heffernan, N. T. (2010). Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In P. De Bra, A. Kobsa, & D. Chin (Eds.), *User Modeling, Adaptation, and Personalization: 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010. Proceedings* (pp. 255–266). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-13470-8_24
- Pardos, Z. A., & Heffernan, N. T. (2011). KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), *User Modeling, Adaption and Personalization: 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings* (pp. 243–254). Berlin, Heidelberg: Springer Berlin Heidelberg.
http://doi.org/10.1007/978-3-642-22362-4_21
- Park, D. H., Kim, H. K., Choi, I. Y., & Kim, J. K. (2012). A Literature Review and Classification of Recommender Systems Research. *Expert Syst. Appl.*, 39(11), 10059–10072.
<http://doi.org/10.1016/j.eswa.2012.02.038>
- Pavlik, P. I., Brawner, K., Olney, A., & Mitrovic, A. (2013). A review of student models used in intelligent tutoring systems. In R. A. Sottolare, A. Graesser, X. Hu, & H. Holden (Eds.), *Design Recommendations for Intelligent Tutoring Systems* (Vol. 1, pp. 39–68). US Army Research

Laboratory, Orlando, FL.

- Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis --A New Alternative to Knowledge Tracing. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling* (pp. 531–538). Amsterdam, The Netherlands, The Netherlands: IOS Press.
- Pazzani, M. J., & Billsus, D. (2007). The Adaptive Web. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), (pp. 325–341). Berlin, Heidelberg: Springer-Verlag.
- Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education, 98*, 169–179. <http://doi.org/10.1016/j.compedu.2016.03.017>
- Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction, 27*(3), 313–350. <http://doi.org/10.1007/s11257-017-9193-2>
- Pelánek, R., & Jarušek, P. (2015). Student Modeling Based on Problem Solving Times. *International Journal of Artificial Intelligence in Education, 25*(4), 493–519. <http://doi.org/10.1007/s40593-015-0048-x>
- Peña-Ayala, A. (2014). Review: Educational Data Mining: A Survey and a Data Mining-based Analysis of Recent Works. *Expert Syst. Appl., 41*(4), 1432–1462. <http://doi.org/10.1016/j.eswa.2013.08.042>
- Peña, A., Kayashima, M., Mizoguchi, R., & Dominguez, R. (2011). Improving Students' Meta-cognitive Skills Within Intelligent Educational Systems: A Review. In *Proceedings of the 6th International Conference on Foundations of Augmented Cognition: Directing the Future of Adaptive Systems* (pp. 442–451). Berlin, Heidelberg: Springer-Verlag.
- Pennock, D. M., Horvitz, E., & Giles, C. L. (2000). Social Choice Theory and Recommender Systems: Analysis of the Axiomatic Foundations of Collaborative Filtering. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence* (pp. 729–734). AAAI Press.
- Perry, M. B., Cochran, J. J., Cox, L. A., Keskinocak, P., Kharoufeh, J. P., & Smith, J. C. (2010). The Exponentially Weighted Moving Average. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc. <http://doi.org/10.1002/9780470400531.eorms0314>
- Pervin, A. L., & John, O. P. (2001). *Personality theory and research (8th ed.)*. New York: John Wiley & Sons Inc.
- Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep Knowledge Tracing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 505–513). Curran Associates, Inc.
- Pintrich, P. R. (2000). The Role of Goal Orientation in Self-Regulated Learning. In *Handbook of Self-Regulation* (pp. 451–502). Elsevier. <http://doi.org/10.1016/B978-012109890-2/50043-3>
- Pintrich, P. R. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review, 16*(4), 385–407. <http://doi.org/10.1007/s10648-004-0006-x>
- Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education : theory, research, and applications* (2nd ed). Englewood Cliffs, N.J. : Merrill, Prentice-Hall International.

- Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. (1991). A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ). *Ann Arbor, MI: University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning.*
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educational and Psychological Measurement, 53*(3), 801–813.
<http://doi.org/10.1177/0013164493053003024>
- Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal Research: The Theory, Design, and Analysis of Change. *Journal of Management, 36*(1), 94–120.
<http://doi.org/10.1177/0149206309352110>
- Putwain, D. W., Daly, A. L., Chamberlain, S., & Sadreddini, S. (2016). “Sink or swim’: buoyancy and coping in the cognitive test anxiety – academic performance relationship.” *Educational Psychology, 36*(10), 1807–1825. <http://doi.org/10.1080/01443410.2015.1066493>
- Puustinen, M., & Rouet, J. F. (2009). Learning with new technologies: Help seeking and information searching revisited. *Computers and Education, 53*(4), 1014–1019.
<http://doi.org/10.1016/j.compedu.2008.07.002>
- Puzziferro, M. (2008). Online Technologies Self-Efficacy and Self-Regulated Learning as Predictors of Final Grade and Satisfaction in College-Level Online Courses. *American Journal of Distance Education, 22*(2), 72–89. <http://doi.org/10.1080/08923640802039024>
- Qiu, Y., Qi, Y., Lu, H., Pardos, Z. A., & Heffernan, N. T. (2011). Does Time Matter? Modeling the Effect of Time with Bayesian Knowledge Tracing. In *Educational Data Mining.*
- Quijano-Sanchez, L., Recio-Garcia, J. A., Diaz-Agudo, B., & Jimenez-Diaz, G. (2013). Social Factors in Group Recommender Systems. *ACM Trans. Intell. Syst. Technol., 4*(1), 8:1--8:30.
<http://doi.org/10.1145/2414425.2414433>
- Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond* (Vol. 240). Wiley Online Library.
- Ragin, C. C., Drass, K. A., & Davey, S. (2006). Fuzzy-set/qualitative comparative analysis 2.0. *Tucson, Arizona: Department of Sociology, University of Arizona.*
- Resnick, P., & Varian, H. R. (1997). Recommender Systems. *Commun. ACM, 40*(3), 56–58.
<http://doi.org/10.1145/245108.245121>
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students’ academic performance: a systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353–387. <http://doi.org/10.1037/a0026838>
- Rienties, B., Tempelaar, D., Giesbers, B., Segers, M., & Gijssels, W. (2014). A dynamic analysis of why learners develop a preference for autonomous learners in computer-mediated communication. *Interactive Learning Environments, 22*(5), 631–648.
<http://doi.org/10.1080/10494820.2012.707127>
- Rihoux, B., & Ragin, C. C. (2009). *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques* (Vol. 51). Sage Publications, Thousand Oaks, CA.
- Robinson, T. N., & Zahn, T. P. (1988). Preparatory interval effects on the reaction time performance of introverts and extraverts. *Personality and Individual Differences, 9*(4), 749–761. [http://doi.org/https://doi.org/10.1016/0191-8869\(88\)90064-5](http://doi.org/https://doi.org/10.1016/0191-8869(88)90064-5)
- Rodrigo, M. M. T., Baker, R. S. J. d., Lagud, M. C. V., Lim, S. A. L., Macapanpan, A. F., Pascua, S. A. M. S., ... Viehland, N. J. B. (2007). Affect and Usage Choices in Simulation Problem-Solving Environments. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work* (pp. 145–152). Amsterdam, The Netherlands, The Netherlands: IOS Press.

- Rodríguez-Triana, M. J., Martínez-Monés, A., Asensio-Pérez, J. I., & Dimitriadis, Y. (2014). Scripting and monitoring meet each other: Aligning learning analytics and learning design to support teachers in orchestrating CSCL situations. *British Journal of Educational Technology*, 46(2), 330–343. <http://doi.org/10.1111/bjet.12198>
- Rodríguez, P., Giraldo, M., Tabares, V., Duque, N., & Ovalle, D. (2016). Recommendation System of Educational Resources for a Student Group. In J. Bajo, M. J. Escalona, S. Giroux, P. Hoffa-D\kabrowska, V. Julián, P. Novais, ... R. Azambuja-Silveira (Eds.), *Highlights of Practical Applications of Scalable Multi-Agent Systems. The PAAMS Collection: International Workshops of PAAMS 2016, Sevilla, Spain, June 1-3, 2016. Proceedings* (pp. 419–427). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-39387-2_35
- Roll, I., Alevén, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2), 267–280. <http://doi.org/10.1016/J.LEARNINSTRUC.2010.07.004>
- Roll, I., Baker, R. S. J. d, Alevén, V., & Koedinger, K. R. (2014). On the Benefits of Seeking (and Avoiding) Help in Online Problem-Solving Environments. *Journal of the Learning Sciences*, 23(4), 537–560. <http://doi.org/10.1080/10508406.2014.883977>
- Romero-Zaldivar, V.-A., Pardo, A., Burgos, D., & Delgado Kloos, C. (2012). Monitoring Student Progress Using Virtual Appliances: A Case Study. *Comput. Educ.*, 58(4), 1058–1067. <http://doi.org/10.1016/j.compedu.2011.12.003>
- Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting Students' Final Performance from Participation in On-line Discussion Forums. *Comput. Educ.*, 68(C), 458–472. <http://doi.org/10.1016/j.compedu.2013.06.009>
- Romero, C., & Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. *Expert Syst. Appl.*, 33(1), 135–146. <http://doi.org/10.1016/j.eswa.2006.04.005>
- Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601–618. <http://doi.org/10.1109/TSMCC.2010.2053532>
- Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12–27. <http://doi.org/10.1002/widm.1075>
- Romero, C., Ventura, S., Espejo, G. P., & Hervás, C. (2008). Data mining algorithms to classify students. In J. B. de Baker T. Barnes (Ed.), *Proceedings of the 1st International Conference on Educational Data Mining, EDM 2008*, (pp. 8–17). Montreal, Canada.
- Romero, C., Ventura, S., Zafra, A., & Bra, P. de. (2009). Applying Web Usage Mining for Personalizing Hyperlinks in Web-based Adaptive Educational Systems. *Comput. Educ.*, 53(3), 828–840. <http://doi.org/10.1016/j.compedu.2009.05.003>
- Roschelle, J., & Teasley, S. D. (1995). The Construction of Shared Knowledge in Collaborative Problem Solving. In C. O'Malley (Ed.), *Computer Supported Collaborative Learning* (pp. 69–97). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Rosé, C. P., Ferschke, O., Tomar, G., Yang, D., Howley, I., Alevén, V., ... Baker, R. (2015). Challenges and Opportunities of Dual-Layer MOOCs: Reflections from an edX Deployment Study. In *Proceedings of the 11th International Conference on Computer Supported Collaborative Learning (CSCL 2015)* (pp. 848–851).
- Roussel, P., Elliot, A. J., & Feltman, R. (2011). The influence of achievement goals and social goals on help-seeking from peers in an academic context. *Learning and Instruction*, 21(3), 394–402. <http://doi.org/10.1016/J.LEARNINSTRUC.2010.05.003>
- Rudner, L. (2003). The Classification Accuracy of Measurement Decision Theory. In *Paper presented at the annual meeting of the National Council on Measurement in Education*.

- Ryan, A. M., & Shin, H. (2011). Help-seeking tendencies during early adolescence: An examination of motivational correlates and consequences for achievement. *Learning and Instruction, 21*(2), 247–256. <http://doi.org/10.1016/J.LEARNINSTRUC.2010.07.003>
- Ryan, & Deci. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology, 25*(1), 54–67. <http://doi.org/10.1006/ceps.1999.1020>
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*. US: American Psychological Association. <http://doi.org/10.1037/0022-3514.57.5.749>
- Ryan, R. M., & Deci, E. L. (2009). Promoting self-determined school engagement: Motivation, learning, and well-being. In *Handbook of motivation at school*. (pp. 171–195). New York, NY, US: Routledge/Taylor & Francis Group.
- Santos, J. L., Govaerts, S., Verbert, K., & Duval, E. (2012). Goal-oriented Visualizations of Activity Tracking: A Case Study with Engineering Students. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge* (pp. 143–152). New York, NY, USA: ACM. <http://doi.org/10.1145/2330601.2330639>
- Santos, O. C., & Boticario, J. G. (2012). Affective issues in semantic educational recommender systems. In D. H. V. K. Manouselis N. & O. C. Santos (Eds.), *Proceedings of the 2nd Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2012)* (Vol. 896, pp. 71–82). Cordoba, Spain: Published by CEUR Workshop Proceedings.
- Sarstedt, M., Henseler, J., & Ringle, C. M. (2011). Multigroup Analysis in Partial Least Squares (PLS) Path Modeling: Alternative Methods and Empirical Results. In *Advances in International Marketing* (pp. 195–218). [http://doi.org/10.1108/S1474-7979\(2011\)0000022012](http://doi.org/10.1108/S1474-7979(2011)0000022012)
- Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 285–295). New York, NY, USA: ACM. <http://doi.org/10.1145/371920.372071>
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling Item Response Times With a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement, 34*(3), 213–232. <http://doi.org/10.1111/j.1745-3984.1997.tb00516.x>
- Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In *Computer-based testing: Building the foundation for future assessments*. (pp. 237–266). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Schuitema, J., Peetsma, T., & van der Veen, I. (2016). Longitudinal relations between perceived autonomy and social support from teachers and students' self-regulated learning and achievement. *Learning and Individual Differences, 49*, 32–45. <http://doi.org/https://doi.org/10.1016/j.lindif.2016.05.006>
- Schunk, D. H. (1991). Self-Efficacy and Academic Motivation. *Educational Psychologist, 26*(3–4), 207–231. <http://doi.org/10.1080/00461520.1991.9653133>
- Schunk, D. H. (1995). Self-efficacy, motivation, and performance. *Journal of Applied Sport Psychology, 7*(2), 112–137. <http://doi.org/10.1080/10413209508406961>
- Schwendimann, B. A., Rodríguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., ... Dillenbourg, P. (2017). Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. *IEEE Transactions on Learning Technologies, 10*(1), 30–41. <http://doi.org/10.1109/TLT.2016.2599522>
- Segedy, J., Kinnebrew, J. S., & Biswas, G. (2011). Modeling Learner's Cognitive and Metacognitive

- Strategies in an Open-Ended Learning Environment. In *AAAI Fall Symposium: Advances in Cognitive Systems*.
- Self, J. A. (1990). Bypassing the intractable problem of student modeling. In C. Frasson & G. Gauthier (Eds.), *Intelligent-tutoring systems: At the crossroads of AI and education* (pp. 107–123). Norwood, NJ: Ablex.
- Sergis, S., Sampson, D. G., & Giannakos, M. (2017). Enhancing Student Digital Skills: Adopting an Ecosystemic School Analytics Approach. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)* (pp. 21–25). <http://doi.org/10.1109/ICALT.2017.87>
- Sergis, S., Sampson, D. G., & Pelliccione, L. (2018). Investigating the impact of Flipped Classroom on students' learning experiences: A Self-Determination Theory approach. *Computers in Human Behavior*, *78*, 368–378. <http://doi.org/10.1016/J.CHB.2017.08.011>
- Setia, M. S. (2016). Methodology Series Module 3: Cross-sectional Studies. *Indian Journal of Dermatology*, *61*(3), 261–264. <http://doi.org/10.4103/0019-5154.182410>
- Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An Investigation of Examinee Test-Taking Effort on a Large-Scale Assessment. *Applied Measurement in Education*, *26*(1), 34–49. <http://doi.org/10.1080/08957347.2013.739453>
- Shee, D. Y., & Wang, Y.-S. (2008). Multi-criteria Evaluation of the Web-based e-Learning System: A Methodology Based on Learner Satisfaction and Its Applications. *Comput. Educ.*, *50*(3), 894–905. <http://doi.org/10.1016/j.compedu.2006.09.005>
- Shih, B., Koedinger, K. R., & Scheines, R. (2008). A response time model for bottom-out hints as worked examples. In R. de Baker, T. Barnes, & J. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 117–126).
- Shute, V. J., & Rahimi, S. (2017, February 1). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*. <http://doi.org/10.1111/jcal.12172>
- Siemens, G., & Baker, R. S. J. d. (2012). Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge* (pp. 252–254). New York, NY, USA: ACM. <http://doi.org/10.1145/2330601.2330661>
- Sierens, E., Vansteenkiste, M., Goossens, L., Soenens, B., & Dochy, F. (2009). The synergistic relationship of perceived autonomy support and structure in the prediction of self-regulated learning. *British Journal of Educational Psychology*, *79*(1), 57–68. <http://doi.org/10.1348/000709908X304398>
- Silm, G., Must, O., & Täht, K. (2013). Test-taking effort as a predictor of performance in low-stakes tests. *Trames Journal of the Humanities and Social Sciences*, *17*(4), 433–448.
- Slavin, R. E. (1987). Developmental and Motivational Perspectives on Cooperative Learning: A Reconciliation. *Child Development*, *58*(5), 1161–1167.
- Sluijsmans, D., Dochy, F., & Moerkerke, G. (1998). Creating a Learning Environment by Using Self-, Peer- and Co-Assessment. *Learning Environments Research*, *1*(3), 293–319. <http://doi.org/10.1023/A:1009932704458>
- Song, D., & Lee, J. (2014). Has Web 2.0 revitalized informal learning? The relationship between Web 2.0 and informal learning. *Journal of Computer Assisted Learning*, *30*(6), 511–533. <http://doi.org/10.1111/jcal.12056>
- Srikant, R., & Agrawal, R. (1996). Mining Quantitative Association Rules in Large Relational Tables. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data* (pp. 1–12). New York, NY, USA: ACM. <http://doi.org/10.1145/233269.233311>

- Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*. Srivastava, Sanjay: Stanford U, Dept of Psychology, Jordan Hall, Building 420, Stanford, CA, US, 94305, sanjay@psych.stanford.edu: American Psychological Association. <http://doi.org/10.1037/0022-3514.84.5.1041>
- Stahl, E., & Bromme, R. (2009). Not everybody needs help to seek help: Surprising effects of metacognitive instructions to foster help-seeking in an online-learning environment. *Computers & Education*, 53(4), 1020–1028. <http://doi.org/10.1016/j.compedu.2008.10.004>
- Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), 111–147.
- Streeter, M. J. (2015). Mixture Modeling of Individual Learning Curves. In *Educational Data Mining*.
- Sun, P.-C., Tsai, R. J., Finger, G., Chen, Y.-Y., & Yeh, D. (2008). What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction. *Computers & Education*, 50(4), 1183–1202. <http://doi.org/10.1016/j.compedu.2006.11.007>
- Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, 29(1), 6–26. [http://doi.org/https://doi.org/10.1016/S0361-476X\(02\)00063-2](http://doi.org/https://doi.org/10.1016/S0361-476X(02)00063-2)
- Tabak, F., Nguyen, N., Basuray, T., & Darrow, W. (2009). Exploring the impact of personality on performance: How time-on-task moderates the mediation by self-efficacy. *Personality and Individual Differences*, 47(8), 823–828. <http://dx.doi.org/10.1016/j.paid.2009.06.027>
- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
- Tanes, Z., Arnold, K. E., King, A. S., & Remnet, M. A. (2011). Using Signals for Appropriate Feedback: Perceptions and Practices. *Comput. Educ.*, 57(4), 2414–2422. <http://doi.org/10.1016/j.compedu.2011.05.016>
- Tatar, D., Roschelle, J., Vahey, P., & Penuel, W. R. (2003). Handhelds Go to School: Lessons Learned. *Computer*, 36, 30–37. <http://doi.org/10.1109/MC.2003.1231192>
- Tempelaar, D. T., Gijsselaers, W. H., van der Loeff, S. S., & Nijhuis, J. F. H. (2007). A structural equation model analyzing the relationship of student achievement motivations and personality factors in a range of academic subject matter areas. *Contemporary Educational Psychology*, 32(1), 105–131. <http://doi.org/10.1016/j.cedpsych.2006.10.004>
- Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, 48(1), 159–205. <http://dx.doi.org/10.1016/j.csda.2004.03.005>
- Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer based assessment. *Computers & Education*, 56(4), 1032–1044. <http://dx.doi.org/10.1016/j.compedu.2010.11.017>
- Terzis, V., Moridis, C. N., & Economides, A. A. (2012). How Student's Personality Traits Affect Computer Based Assessment Acceptance: Integrating BFI with CBAAM. *Comput. Hum. Behav.*, 28(5), 1985–1996. <http://doi.org/10.1016/j.chb.2012.05.019>
- Thai-Nghe, N., Horváth, T., & Schmidt-Thieme, L. (2011). Factorization Models for forecasting student performance. In C. T. C. V. S. R. C. Pechenizkiy M. & J. Stamper (Eds.), *Proceedings of the 4th International Conference on Educational Data Mining, EDM 2011* (pp. 11–20). Eindhoven, the Netherlands.
- Thomson, D., & Mitrovic, A. (2009). Towards a negotiable student model for constraint-based

- ITs. In *17th International Conference on Computers in Education* (pp. 83–90). Hong Kong.
- Towle, B., & Halm, M. (2005). Designing Adaptive Learning Environments with Learning Design. In R. Koper & C. Tattersall (Eds.), *Learning Design: A Handbook on Modelling and Delivering Networked Education and Training* (pp. 215–226). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/3-540-27360-3_12
- Triantafyllou, E., Georgiadou, E., & Economides, A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education, 50*(4), 1319–1330. <http://doi.org/https://doi.org/10.1016/j.compedu.2006.12.005>
- Trivedi, S., Pardos, Z. A., & Heffernan, N. T. (2011). Clustering Students to Generate an Ensemble to Improve Standard Test Score Predictions. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial Intelligence in Education* (pp. 377–384). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Trueman, M., & Hartley, J. (1996). A comparison between the time-management skills and academic performance of mature and traditional-entry university students. *Higher Education, 32*(2), 199–215. <http://doi.org/10.1007/BF00138396>
- Tschofen, C., & Mackness, J. (2012). Connectivism and dimensions of individual experience. *The International Review of Research in Open and Distributed Learning, 13*(1), 124–143.
- Turner, J. R. (2013). Cross-Sectional Study. In M. D. Gellman & J. R. Turner (Eds.), *Encyclopedia of Behavioral Medicine* (p. 522). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4419-1005-9_1010
- United Nations Economic and Social Council. (2006). *Definition of basic concepts and terminologies in governance and public administration*.
- Vaessen, B. E., Prins, F. J., & Jeurig, J. (2014). University students' achievement goals and help-seeking strategies in an intelligent tutoring system. *Computers & Education, 72*, 196–208. <http://doi.org/10.1016/j.COMPEDU.2013.11.001>
- Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. F. (1992). The Academic Motivation Scale: A Measure of Intrinsic, Extrinsic, and Amotivation in Education. *Educational and Psychological Measurement, 52*(4), 1003–1017. <http://doi.org/10.1177/0013164492052004025>
- Van Der Linden, J. W. (2009). Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement, 46*(3), 247–272. <http://doi.org/10.1111/j.1745-3984.2009.00080.x>
- van der Linden, W. J. (2009). A Bivariate Lognormal Response-Time Model for the Detection of Collusion Between Test Takers. *Journal of Educational and Behavioral Statistics, 34*(3), 378–394. <http://doi.org/10.3102/1076998609332107>
- van der Linden, W. J. (2011). Test Design and Speededness. *Journal of Educational Measurement, 48*(1), 44–60. <http://doi.org/10.1111/j.1745-3984.2010.00130.x>
- van der Linden, W. J., Entink, R. H. K., & Fox, J.-P. (2010). IRT Parameter Estimation With Response Times as Collateral Information. *Applied Psychological Measurement, 34*(5), 327–347. <http://doi.org/10.1177/0146621609349800>
- Vanijdee, A. (2003). Thai Distance English Learners and Learner Autonomy. *Open Learning: The Journal of Open, Distance and e-Learning, 18*(1), 75–84. <http://doi.org/10.1080/0268051032000054130>
- Vansteenkiste, M., Sierens, E., Goossens, L., Soenens, B., Dochy, F., Mouratidis, A., ... Beyers, W. (2012). Identifying configurations of perceived teacher autonomy support and structure: Associations with self-regulated learning, motivation and problem behavior. *Learning and Instruction, 22*(6), 431–439. <http://doi.org/10.1016/j.learninstruc.2012.04.002>

- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3), 425–478.
- Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., & Duval, E. (2011). Dataset-driven Research for Improving Recommender Systems for Learning. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 44–53). New York, NY, USA: ACM. <http://doi.org/10.1145/2090116.2090122>
- Verhelst, N. D., & Glas, C. A. W. (1995). The One Parameter Logistic Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (pp. 215–237). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4612-4230-7_12
- Vermetten, Y. J., Lodewijks, H. G., & Vermunt, J. D. (2001). The Role of Personality Traits and Goal Orientations in Strategy Use. *Contemporary Educational Psychology*, 26(2), 149–170. <http://doi.org/https://doi.org/10.1006/ceps.1999.1042>
- Vesin, B., Ivanović, M., Klačnja-Milićević, A., & Budimac, Z. (2012). Protus 2.0: Ontology-based semantic recommendation in programming tutoring system. *Expert Systems with Applications*, 39(15), 12229–12246. <http://doi.org/10.1016/J.ESWA.2012.04.052>
- Wainer, H. (2000). *Computerized adaptive testing: A Primer (2nd Edition)*. Mahwah, NJ: Erlawrence Erlbaum Associates.
- Wang, C., Chang, H. H., & Boughton, K. A. (2013). Deriving Stopping Rules for Multidimensional Computerized Adaptive Testing. *Applied Psychological Measurement*, 37(2), 99–122. <http://doi.org/10.1177/0146621612463422>
- Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, 29(5), 323–339. <http://doi.org/10.1177/0146621605275984>
- Wang, X., Berger, J. O., & Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, 7(1), 126–153.
- Wang, Y.-S. (2003). Assessment of learner satisfaction with asynchronous electronic learning systems. *Information & Management*, 41(1), 75–86. [http://dx.doi.org/10.1016/S0378-7206\(03\)00028-4](http://dx.doi.org/10.1016/S0378-7206(03)00028-4)
- Watson, D., & Clark, A. L. (1997). *Extraversion and its positive emotional core*. San Diego: Academic Press.
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*. US: American Psychological Association. <http://doi.org/10.1037/0033-295X.92.4.548>
- Weiss, D. J. (2004). Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education. *Measurement and Evaluation in Counseling and Development*, 37(2), 70–84. <http://doi.org/10.1080/07481756.2004.11909751>
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to educational problems. *Journal of Educational Measurement*, 21(4), 361–375. <http://doi.org/10.1111/j.1745-3984.1984.tb01040.x>
- White, C. (1995). Autonomy and strategy use in distance foreign language learning: Research findings. *System*, 23(2), 207–221. [http://doi.org/10.1016/0346-251X\(95\)00009-9](http://doi.org/10.1016/0346-251X(95)00009-9)
- White, M. C., & Bembenuddy, H. (2013). Not All Avoidance Help Seekers Are Created Equal: Individual Differences in Adaptive and Executive Help Seeking. *SAGE Open*, 3(2), 2158244013484916. <http://doi.org/10.1177/2158244013484916>
- Whiting, S. W. (2015). *Temporal dynamics in information retrieval*. University of Glasgow.

- Wigfield, A., & Eccles, J. S. (2000). Expectancy-Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, 25(1), 68–81.
<http://doi.org/https://doi.org/10.1006/ceps.1999.1015>
- Williams, J. D., & Takaku, S. (2011). Help Seeking and Writing Performance among College Students: A Longitudinal Study. *Journal of Writing Research*, 3(issue 1), 1–18.
<http://doi.org/10.17239/jowr-2011.03.01.1>
- Wilson, K., Boyd, C., Chen, L., & Jamal, S. (2011). Improving student performance in a first-year geography course: Examining the importance of computer-assisted formative assessment. *Computers & Education*, 57(2), 1493–1500.
<http://doi.org/10.1016/j.compedu.2011.02.011>
- Wilson, K. H., Karklin, Y., Han, B., & Ekanadham, C. (2016). Back to the Basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation. In *9th International Conference on Educational Data Mining* (pp. 539–544).
- Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, 18(2), 163–183.
http://doi.org/10.1207/s15324818ame1802_2
- Wolpers, M., Najjar, J., Verbert, K., & Duval, E. (2007). Tracking Actual Usage: The Attention Metadata Approach. *Journal of Educational Technology & Society*, 10(3), 106–121.
 Retrieved from <https://www.learntechlib.org/p/75407>
- Wolsey, T. (2008). Efficacy of Instructor Feedback on Written Work in an Online Program. *International Journal on E-Learning*, 7(2), 311–329.
- Wood, H., & Wood, D. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, 33(2–3), 153–169. [http://doi.org/10.1016/S0360-1315\(99\)00030-5](http://doi.org/10.1016/S0360-1315(99)00030-5)
- Woodside, A. G. (2013). Moving beyond multiple regression analysis to algorithms: Calling for adoption of a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory. *Journal of Business Research*, 66(4), 463–472.
- Worsley, M., & Blikstein, P. (2013). Towards the Development of Multimodal Action Based Assessment. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 94–101). New York, NY, USA: ACM.
<http://doi.org/10.1145/2460296.2460315>
- Xie, T., Zheng, Q., & Zhang, W. (2018). Mining temporal characteristics of behaviors from interval events in e-learning. *Information Sciences*, 447, 169–185.
<http://doi.org/10.1016/j.ins.2018.03.018>
- Xiong, X., Pardos, Z. A., & Heffernan, N. T. (2011). An Analysis of Response Time Data for Improving Student Performance Prediction. In *In KDD 2011 Workshop: Knowledge Discovery in Educational Data, Held as part of 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Xu, D., & Jaggars, S. S. (2014). Performance Gaps Between Online and Face-to-Face Courses: Differences Across Types of Students and Academic Subject Areas. *The Journal of Higher Education*, 85(5), 633–659. <http://doi.org/10.1353/jhe.2014.0028>
- Yager, R. R. (2003). Fuzzy Logic Methods in Recommender Systems. *Fuzzy Sets Syst.*, 136(2), 133–149. [http://doi.org/10.1016/S0165-0114\(02\)00223-3](http://doi.org/10.1016/S0165-0114(02)00223-3)
- Yang, Y., & Taylor, J. (2013). The role of achievement goals in online test anxiety and help-seeking. *Educational Research and Evaluation*, 19(8), 651–664.
<http://doi.org/10.1080/13803611.2013.811086>
- Yanhui, D., Dequan, W., Yongxin, Z., & Lin, L. (2015). A Group Recommender System for Online Course Study. In *2015 7th International Conference on Information Technology in Medicine*

- and Education (ITME) (pp. 318–320). <http://doi.org/10.1109/ITME.2015.99>
- Yao, L. (2012). Multidimensional CAT Item Selection Methods for Domain Scores and Composite Scores: Theory and Applications. *Psychometrika*, 77(3), 495–523. <http://doi.org/10.1007/s11336-012-9265-5>
- Yen, C.-J., & Liu, S. (2009). Learner Autonomy as a Predictor of Course Success and Final Grades in Community College Online Courses. *Journal of Educational Computing Research*, 41(3), 347–367. <http://doi.org/10.2190/EC.41.3.e>
- Yu, Z., Zhou, X., Hao, Y., & Gu, J. (2006). TV Program Recommendation for Multiple Viewers Based on user Profile Merging. *User Modeling and User-Adapted Interaction*, 16(1), 63–82. <http://doi.org/10.1007/s11257-006-9005-6>
- Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian Knowledge Tracing Models. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings* (pp. 171–180). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-39112-5_18
- Zapata, A., Menéndez, V. H., Prieto, M. E., & Romero, C. (2015). Evaluation and selection of group recommendation strategies for collaborative searching of learning objects. *International Journal of Human-Computer Studies*, 76(Supplement C), 22–39. <http://doi.org/https://doi.org/10.1016/j.ijhcs.2014.12.002>
- Zhou, T., Jiang, L.-L., Su, R.-Q., & Zhang, Y.-C. (2008). Effect of initial configuration on network-based recommendation. *EPL (Europhysics Letters)*, 81, 58004. <http://doi.org/10.1209/0295-5075/81/58004>
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation*. (pp. 13–39). San Diego, CA, US: Academic Press. <http://doi.org/10.1016/B978-012109890-2/50031-7>
- Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In *Self-regulated learning and academic achievement: Theoretical perspectives, 2nd ed.* (pp. 1–37). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, 41(2), 64–70.
- Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-Motivation for Academic Attainment: The Role of Self-Efficacy Beliefs and Personal Goal Setting. *American Educational Research Journal*, 29(3), 663–676.
- Zimmerman, B. J., & Pons, M. M. (1986). Development of a Structured Interview for Assessing Student Use of Self-Regulated Learning Strategies. *American Educational Research Journal*, 23(4), 614–628. <http://doi.org/10.3102/00028312023004614>
- Zlatović, M., Balaban, I., & Kermek, D. (2015). Using online assessments to stimulate learning strategies and achievement of learning goals. *Computers & Education*, 91, 32–45. <http://doi.org/http://dx.doi.org/10.1016/j.compedu.2015.09.012>
- Zweig, D., & Webster, J. (2004). What are we measuring? An examination of the relationships between the big-five personality traits, goal orientation, and performance intentions. *Personality and Individual Differences*, 36(7), 1693–1708. <http://doi.org/10.1016/j.paid.2003.07.010>

Appendix A : The Learning Analytics and Educational Recommender System

The LAERS (Learning Analytics and Educational Recommendations System) assessment environment architecture

The standard version of LAERS is a self-assessment environment, consisting of four components; (a) a testing module (presentation layer), (b) a tracker that logs the learners' interaction data (tracking layer), (c) a pre-processing engine (application logic layer) that mines the data-logs and shapes/revises the learner models, and (d) a storage module (data accumulation layer) that stores in a database information related to the learners and the self-assessment tasks, as well as analytics from the learners' past activity with the system. In the full version of the system, two additional components have been integrated in the application logic layer: (a) an assessment analytics engine that produces analytics and diagnostics in real-time during self-assessment for visualization and decision support purposes, and (b) an adaptation engine that provides personalized feedback to the students after consulting the current states of learner models and the learners' responses on the latest self-assessment tasks. An overview of the abstract architecture of the full system is illustrated in Figure A-1.

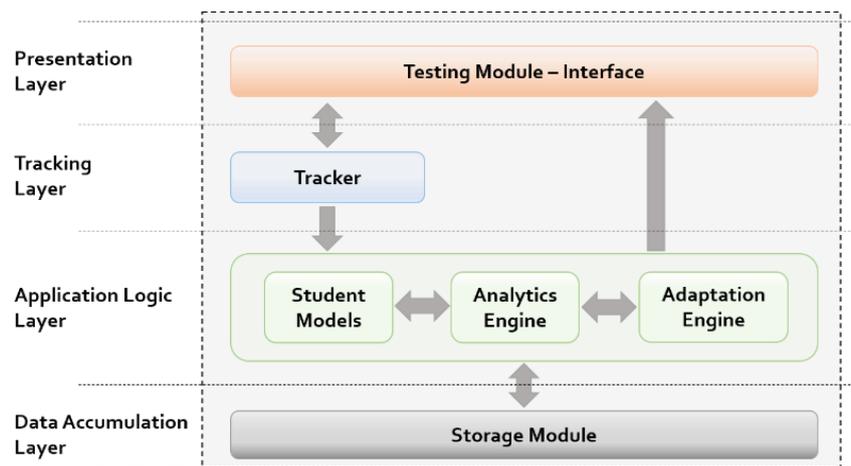


Figure A-1. The abstract architecture of the full LAERS version

The first component of the expert system (i.e., the testing module - presentation layer) implements the interface that displays the self-assessment items delivered to students separately and one-by-one. The interface delivers the items to the students in predetermined order, and it allows them to temporarily save their answers, to review them, to alter their initial answer choices, and to save new answers. Students can also skip an item (either because they are not sure about the answer, or because they think it is too difficult), and answer it (or not) later. They finalize and submit their answers only once, whenever they estimate that they are ready to do so, within the duration of the self-assessment. In a later version of LAERS, the students can also ask for hints, they can inquire about similar items, and they can ask for analytics about each item (e.g., how many students have submitted correct/wrong answers on this item, which is the average time to answer this item, which is the average perceived difficulty of the item, etc.). Figure A-2 illustrates the LAERS interface.

The second component of the system (i.e., the tracker - tracking layer) records the students' interaction data during self-assessment. In log files, it tracks and aggregates students'

time-spent on handling the self-assessment items, distinguishing it between the time-spent on correctly and wrongly answered items. It also aggregates the time-spent according to the difficulty level of the items, distinguishing it between the time-spent on easy, medium or hard items. In a log file, the tracker also logs how many times the students reviewed each item, how many times they changed the answers, and the respective time-spent during these interactions. The overall logged features of students' activity during a self-assessment procedure are listed in Table A-1. The system also calculates the learning performance (i.e., score) (LP) for each student:

$LP = \sum_{i=1}^N d_i z_i$, where $z_i \in (0,1)$ is the correctness of the student's answer on item i , and d_i is the difficulty of the item. For the score computation, only the correct answers are considered, without penalizing the incorrect answers (i.e., without negative scores). In case students choose not to submit an answer to an item, they receive zero points for this one.

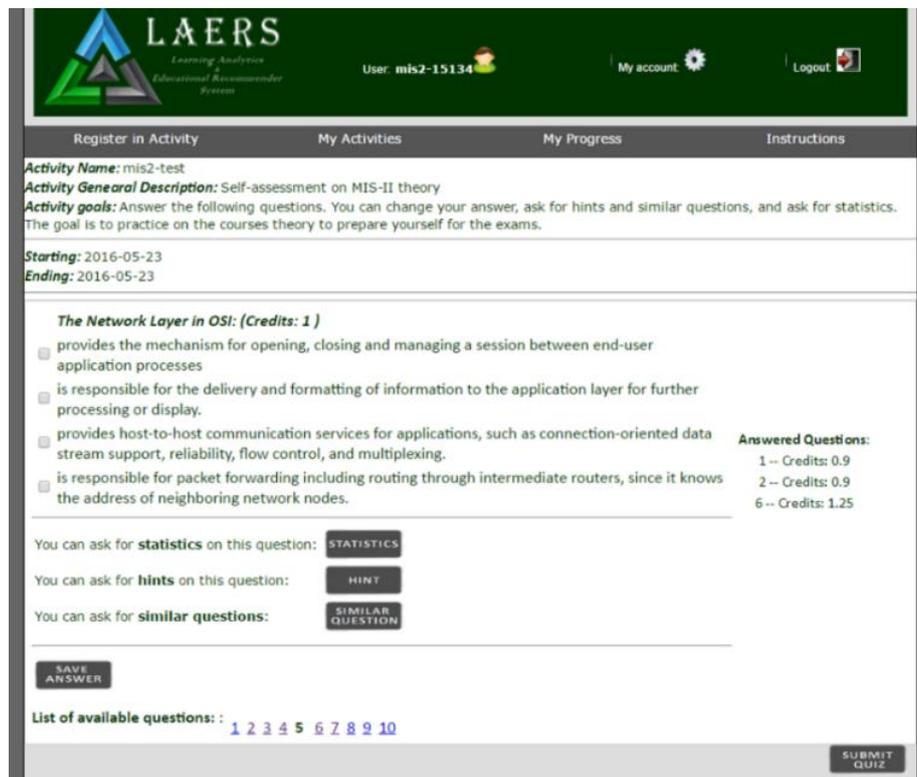


Figure A-2. The LAERS student interface

The application logic layer of LAERS includes three independent engines that continuously and repetitively exchange information (illustrated in Figure A-3); (a) a pre-processing engine that mines the data-logs and shapes/revises the student models, (b) an assessment analytics engine that produces analytics in real time during self-assessment, and (c) an adaptation engine that suggests the next most appropriate item.

More precisely, the pre-processing engine organizes the data from the data logs for each individual's observed activity (e.g., cleans, filters, classifies) and prepares them to be loaded on the student models. The exact student characteristics considered in the student models are discussed in Chapter 6. The final states of the student models are stored for later usage in the data accumulation layer, right after the completion of the self-assessment. The student models are fed to the assessment analytics engine that, on-demand, builds and administers these analytics to the students (e.g. visualizations of students' progress and system's usage to guide time-management,

self-regulation, self-monitoring, etc.). The student models are also fed to the adaptation engine; it recommends suitable self-assessment items and provides personalized feedback to the students according to the student models at that time and to their handling on the previous self-assessment items, while preserving students' autonomy on which item to choose and deal with.

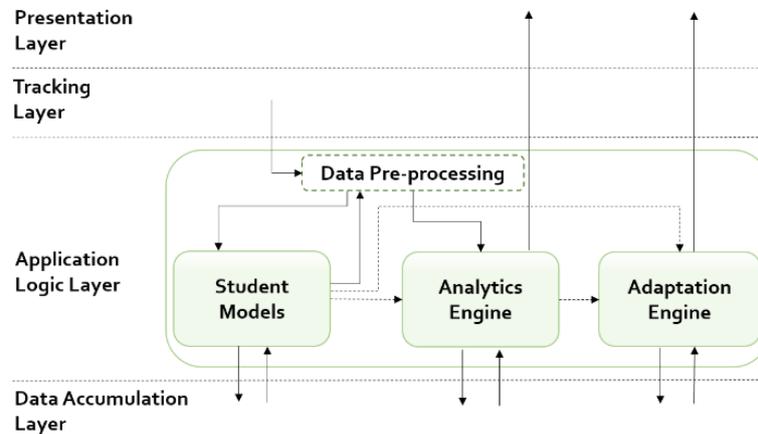


Figure A-3. Data flow in the application logic layer

The system is developed in PHP 5.4, MySQL 5.1 and runs on Apache 2.4. Javascript, AJAX and JQuery have also been used for implementing the system's functionalities.

Table A-1. Features from the raw log files

Simple Feature (Tracked)	Computed Feature
1. student ID	1. the total time the student spends on viewing the items and submitting the correct answers
2. the item the student works on	2. the total time the student spends on viewing the items and submitting the wrong answers
3. the timestamp the student starts viewing an item	3. the idle time the student spends viewing each item (not saving an answer)
4. the timestamp the student leaves an item (not saving an answer)	4. the total idle time the student spends on re-viewing the items
5. the timestamp the student saves an answer	5. the student's total active time on item
6. the timestamp the student chooses to re-view an item	6. the student's total idle time on item
7. the timestamp the student leaves an item after re-view (saving an answer)	7. how many times the student changes the answer saved for each item
8. the answer the student saves	8. how many times the student reviews each item
9. the correctness of the saved answer	9. how many times the student views the item
10. the timestamp the student requests a hint	10. the total time the student spends on viewing the easy items and submitting an answer
11. the timestamp the student closes the hint	11. the total time the student spends on viewing the medium items and submitting an answer
12. the timestamp the student opens the statistics (analytics) about an item	12. the total time the student spends on viewing the hard items and submitting an answer
13. the timestamp the student closes the statistics	13. The total time the student spends on viewing the hint
14. the timestamp the student chooses to see the remaining time (shows the timer)	14. The total time the student spends on viewing the statistics (analytics)
15. the timestamp the student hides the timer	15. The total time until making the decision what type of item to ask for
16. the timestamp the student asks for the next item	16. The frequency the student requests hints
	17. The frequency the student views analytics
	18. The frequency the student shows the timer
	19. The total time the student has the timer "ON"
	20. How many times the student asks for harder items
	21. How many times the student asks for easier items
	22. How many times the student asks for items of the same difficulty
	23. How many times the student asks for random items
	24. How many times the student takes the items that the system suggests
	25. How many times the student asks to replace the items that the system suggested

Brief description of the variables used in this study.

Table A-2. Variables used in this study and short description

Variable	Description
Perceived clarity of the content (CONT)	Clarity of content was proposed as a determinant of student satisfaction. This subjective perception of how well a learning item meets the student's expectations for learning and supports success (Lo, 2010) may help refine our insight on the test result. Students' perceptions about the clarity of the assessment content (CONT) may affect their overall beliefs regarding the assessment to be considered as difficult or easy, interesting or boring. Thus, CONT stores information related to whether the students considered the items to be clear, understandable and relative to the course's content. It is a measure of learning satisfaction.
Effort-regulation (ER)	Effort regulation is the control of "one's effort expenditure". It incorporates the learners' ability to exert effort and to persist in their engagement with the learning items.
Goal expectancy (GE)	Goal-expectancy (GE) reflects the students' dispositions regarding their achievement expectations in the assessment, and was proposed in the Computer Based Assessment Acceptance Model (CBAAM). Goal-expectancy has two dimensions: (a) students' perception of preparation for the assessment – how satisfied they are with their preparation – and (b) their desirable level of success. Before taking the assessment, the students set a goal regarding the percentage of correct answers that they perceive as a satisfying performance.
Help Seeking (HS)	Help seeking (HS) is a behavior associated to students' need to ask for multiple levels of hints during dealing with assessment items. Students' help-seeking behavior reflects their metacognitive and domain-specific skills and knowledge; knowing when and how to seek help is a key self-regulatory skill.
Prior Skill Mastery (PSM)	Student's prior knowledge to successfully complete tasks that correspond to the specific skill.
Perceived usefulness (PU)	Perceived usefulness (PU) is one of the two basic constructs of the original version of Technology Acceptance Model (TAM) – the other one is perceived ease-of-use – and is defined as "the degree to which a person believes that using a particular system will enhance his/her job performance". This feature corresponds to and influences the students' behavioral intention to use again the assessment system. Thus, perceived usefulness reflects the students' satisfaction regarding the offered assessment services.
Perceived Usefulness of Visualizations (PUV)	Particularization of the Perceived Usefulness variable, to measure how useful and helpful does the student perceive the analytics visualizations.
Self-efficacy (SE)	Self-efficacy (SE) has been defined as one's belief in one's ability to succeed in specific situations or accomplish a task. Perceived self-efficacy plays a major role in how the subject approaches goals, tasks, and challenges. Students who score high in self-efficacy—that is, those who believe they can perform well—are more likely to view difficult tasks as something to be mastered rather than something to be avoided, and are more likely to make efforts to complete these tasks, and to persist longer in their efforts.
Time Management (TM)	Time management (TM) is a meta-skill aiming at maximize the overall benefit of a set of other activities within the boundary condition of a limited amount of time, as time itself cannot be managed because it is fixed. It is associated with students' exercising conscious control over the amount of time-spent on learning items during the learning process.

Self-reported constructs measured with questionnaires – See Appendix C

Analytics parameters Extracted from the raw interactions data and Developed/Evaluated during this study	Time to answer correctly (TTAC)	Time to answer correctly (TTAC) and Time to answer wrongly (TTAW) are defined as the total time that students spend on viewing the assessment items and submitting the correct and wrong answers respectively. By definition, they indicate the respective response-time the students constantly aggregate on answering the assessment questions.
	Time to answer wrongly (TTAW)	
	Time to answer easy items (TTAE)	Item's difficulty is a critical factor that reflects the item's quality, and has been strongly associated with students' achievement behavior, motivation and performance. In a sense, the total time the students aggregate to answer easy (TTAE), medium (TTAM) and hard (TTAH) assessment items, represent and quantify the student's current level of knowledge, skills and abilities, similar to stereotypes for students' mastery of the subject domain.
	Time to answer fair items (TTAM)	
	Time to answer hard items (TTAH)	
	Time-spent on analytics visualizations viewing (TAVV)	The total average time the student spends on viewing the analytics visualizations and engage on sense-making, and the frequency (i.e., how many times) that the student asks for analytics visualizations, i.e., a counter that increases every time that the students make the respective request.
	Frequency of analytics visualizations request (FAVR)	
	Time-spent on decision making (TTDM)	The average total time that the students spend from the moment that they answer to an item until they make a decision on what item they want to answer next, and they ask for it.
	Frequency of choosing easier (FEAS)	How many times the student asks for easier questions (compared to the current one), How many times the student asks for harder questions (compared to the current one), How many times the student asks for question of the same difficulty with the current one, How many times the student asks for a random question, How many times the student takes the question that the system delivers, How many times the student asks for another question to replace the one that the system delivers. Simple counters for the respective choices, and they increase every time that the students make the respective selection of next item.
	Frequency of choosing harder (FHAR)	
	Frequency of choosing same (FSAM)	
	Frequency of choosing random (FRAN)	
	Frequency of acceptance (FACC)	
	Frequency of replacement (FREP)	
Frequency of showing/hiding the timer (FSHT)	How many times the student asks to show/hide the timer.	
Duration of showing the timer (TTST)	For how long the student shows the timer (it has it in the "ON" mode).	
Level of certainty (CERT)	Certainty is used to describe a person's strength of belief about the accuracy of a choice. In assessment procedures, the level of certainty (CERT) reflects how certain the students want to be before answering a question; the more certain the students want to be before finalizing their answers, the more the idle time they spend on re-viewing the items and the more the times they re-view the items. Overall, level of certainty is included in the student models as a core feature representing students' cautiousness.	
Calculated Analytics Factor	Response Time Effort (RTE)	Effort is "the motivational state commonly understood to mean trying hard or being involved in a task. Effort is increased when the subject tries harder, when there are incentives to perform well, or when the task is important or difficult". Thus, effort is about how much engaged the learners are in completing the tasks. Response Time Effort (RTE) measures the proportion of items which the students try to solve (solution behaviour – SB) instead of guessing the answers. Less engaged students will answer too quickly, before they had time to fully consider the items (see Appendix B).

Appendix B : Algorithms and Formulas

1. Algorithm 1: The self-assessment test adaptation algorithm

For the adaptive phase of the self-assessment tests, we utilized Measurement Decision Theory (MDT – Rudner, 2003) to classify the students in three mastery classes based on their item responses, a priori item information, and a priori population classification proportions. The core of the methodology in use is the estimation of the students' mastery class every time they submit an answer. This estimation is reached by knowing prior probabilities and Bayes Theorem: $P(m_k|z) = c \cdot P(z|m_k) \cdot P(m_k)$, with $z = (z_1, z_2, \dots, z_n)$ being a student's response vector with $z_i \in (0,1)$, and: (a) $P(m_k|z)$ is the probability that the student belongs to mastery class m_k given z , (b) $P(z|m_k)$ is the probability of responses z given the student's mastery class, (c) $P(m_k)$ is the probability of a randomly selected student belonging to mastery class m_k , and (d) c is a standardization constant so that $P(m_1|z) + P(m_2|z) + P(m_3|z) = 1$.

At each step, the posterior classification probabilities $P(m_k|z)$ are treated as updated prior probabilities $P(m_k)$ and are used to help identify the next item to be assigned. The selection of the next item is based on entropy, a maximum information gain strategy from Information Theory. The goal is to have a peaked distribution of $P(m_k)$, and to next select the item that has the greatest expected reduction in entropy. This process continues until either a fixed number of items is reached, or a degree of decision accuracy is attained. For the termination of the test, we used the Sequential Probability Ratio Test (SPRT) criterion and a maximum number of items assigned. In our study, if the SPRT criterion was not met after assigning 12 items, the test ended and the student was classified into the mastery class with the largest probability $P(m_k|z)$ to this point.

For the calibration of the item-bank, prior testing (involving students who have already been classified) can be used. The priors $P(z|m_k)$ and $P(m_k)$ were computed by previously administering our bank of items to previous students. We identified three mastery classes of students: Class A (advanced): final grade ≥ 7 , Class B (basic): final grade ≥ 4 , and Class C (below basic): final grade < 4 . $P(m_k)$ is an approximation of the portion of students below basic, basic, and advanced. For each student and each item we logged the correctness of each answer (right (1) - wrong (0), $P(z_1=0|m_k) = 1 - P(z_1=1|m_k)$). After the estimation of $P(m_k)$, and for each of the items, we estimated three probabilities, according to how likely is a student of the given class m_k to answer correctly to this item. The probability $P(z|m_k)$ of the response vector z is the conditional probability of students in each mastery class responding correctly to each item, and equals to the product of the conditional probabilities of the item responses.

We adopted this methodology due to the low computational resources demanded, the little pre-testing required, and, the satisfactory level of the accuracy of classification using moderately small samples for item calibration. In fact, MDT can produce results that are comparable to that of more complicated IRT classification procedures. Another reason that justifies the choice of MDT is that one doesn't need to be concerned with the fit of the data to a theoretical model as in IRT or in most latent class models.

Rudner, L. M. (2003). The classification accuracy of Measurement Decision Theory, *Annual meeting of the National Council on Measurement in Education*, Chicago.

2. The formula for Response Time Effort calculation

The Response Time Effort (RTE) measures the proportion of items which the students try to solve (solution behaviour – SB) instead of guessing the answers (Wise & Kong, 2005). The RTE for a

student j is: $RTE_j = \frac{\sum SB_{ij}}{k}$, where k is the number of items, and $SB_{ij} = \begin{cases} 1, & \text{if } RT_{ij} \geq T_i \\ 0, & \text{otherwise} \end{cases}$, where

RT_{ij} is student's j response time to item i , and T_i discriminates solution from guessing (threshold).

Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, 18(2), 163–183.

3. Algorithm 2: Hoeffding Adaptive Tree (HAT)

HAT algorithm as it is implemented in MOA – (see Bifet, A. and Gavaldà, R. (2009))

HOEFFDING WINDOW TREE(Stream, δ)

- 1 Let HT be a tree with a single leaf(root)
- 2 Init estimators A_{ijk} at root
- 3 **for** each example (x, y) in Stream
- 4 **do** HWTREEGROW((x, y) , HT, δ)

HWTREEGROW((x, y) , HT, δ)

- 1 Sort (x, y) to leaf l using HT
- 2 Update estimators A_{ijk}
- 3 at leaf l and nodes traversed in the sort
- 4 **if** current node l has an alternate tree T_{alt}
- 5 HWTREEGROW((x, y) , T_{alt} , δ)
- 6 Compute G for each attribute
- //Evaluate condition for splitting leaf l
- 7 **if** $G(\text{Best Attr.}) - G(\text{2nd best}) > \epsilon(\delta', \dots)$
- 8 **then** Split leaf on best attribute
- 9 **for** each branch of the split
- 10 **do** Start new leaf
- 11 and initialize estimators
- 12 **if** one change detector has detected change
- 13 **then** Create an alternate subtree T_{alt} at leaf l if there is none
- 14 **if** existing alternate tree T_{alt} is more accurate
- 15 **then** replace current node l with alternate tree T_{alt}

Here δ' should be the Bonferroni correction of δ to account for the fact that many tests are performed and we want all of them to be simultaneously correct with probability $1 - \delta$. It is enough e.g. to divide δ by the number of tests performed so far.

Bifet, A., & Gavaldà, R. (2009). Adaptive Learning from Evolving Data Streams. In N. M. Adams, C. Robardet, A. Siebes, & J.-F. Boulicaut (Eds.), *Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, August 31 - September 2, 2009. Proceedings* (pp. 249–260). Berlin, Heidelberg: Springer Berlin Heidelberg.

Appendix C : Questionnaires & Instruments

1. Table C-1. Chapter 3: Measuring Goal-Expectancy in Study 1

Construct	Items	Description
Goal Expectancy (GE) - Terzis & Economides (2011)	GE1	Courses' preparation was sufficient for the test
	GE2	My personal preparation for the test was sufficient
	GE3	My performance expectations for the test

2. Table C-2. Chapter 3: Measuring self-regulation and satisfaction from content in Study 2 – Constructs and items from the questionnaires

Construct	Items	Description
Time Management (TM) - Macan (1990)	TM1	I spend more time than I want trying to find things.
	TM2	I use goal setting to determine my most important activities.
	TM3	I put off tasks that are difficult or I don't like.
Goal Expectancy (GE) - Terzis & Economides (2011)	GE1	Courses' preparation was sufficient for the test
	GE2	My personal preparation for the test was sufficient
	GE3	My performance expectations for the test
Comprehensibility of Content (CONT) - Terzis & Economides (2011)	CONT1	Questions were clear and understandable
	CONT2	Questions were relative with the syllabus
	CONT3	Questions were suitable for measuring my understanding of the course's concepts

3. Table C-3. Chapter 4: The testing analytics paradigm: Measuring non-cognitive/motivational factors – Constructs and items from the questionnaires

Construct	Items	Description
Self-Efficacy (SE) - Bandura (2006)	SE1	I remember well information presented in class and textbooks
	SE2	I get myself to study when there are other interesting things to do
	SE3	I finish my homework assignments by deadlines
Goal Expectancy (GE) - Terzis & Economides (2011)	GE1	Courses' preparation was sufficient for the test
	GE2	My personal preparation for the test was sufficient
	GE3	My performance expectations for the test
Perceived Usefulness (PU) - Davis (1989)	PU1	Using the test will improve my learning
	PU2	Using the test will enhance my effectiveness
	PU3	Using the test will increase my productivity
Comprehensibility of Content (CONT) - Terzis & Economides (2011)	CONT1	Questions were clear and understandable
	CONT2	Questions were relative with the syllabus
	CONT3	Questions were suitable for measuring my understanding of the course's concepts

4. Table C-4. Chapter 5: The Big Five Inventory (BFI)

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

Disagree strongly 1	Disagree a little 2	Neither agree nor disagree 3	Agree a little 4	Agree strongly 5
------------------------	------------------------	------------------------------------	---------------------	---------------------

I see Myself as Someone Who...

- | | |
|--|---|
| __1. Is talkative | __23. Tends to be lazy |
| __2. Tends to find fault with others | __24. Is emotionally stable, not easily upset |
| __3. Does a thorough job | __25. Is inventive |
| __4. Is depressed, blue | __26. Has an assertive personality |
| __5. Is original, comes up with new ideas | __27. Can be cold and aloof |
| __6. Is reserved | __28. Perseveres until the task is finished |
| __7. Is helpful and unselfish with others | __29. Can be moody |
| __8. Can be somewhat careless | __30. Values artistic, aesthetic experiences |
| __9. Is relaxed, handles stress well | __31. Is sometimes shy, inhibited |
| __10. Is curious about many different things | __32. Is considerate and kind to almost everyone |
| __11. Is full of energy | __33. Does things efficiently |
| __12. Starts quarrels with others | __34. Remains calm in tense situations |
| __13. Is a reliable worker | __35. Prefers work that is routine |
| __14. Can be tense | __36. Is outgoing, sociable |
| __15. Is ingenious, a deep thinker | __37. Is sometimes rude to others |
| __16. Generates a lot of enthusiasm | __38. Makes plans and follows through with them |
| __17. Has a forgiving nature | __39. Gets nervous easily |
| __18. Tends to be disorganized | __40. Likes to reflect, play with ideas |
| __19. Worries a lot | __41. Has few artistic interests |
| __20. Has an active imagination | __42. Likes to cooperate with others |
| __21. Tends to be quiet | __43. Is easily distracted |
| __22. Is generally trusting | __44. Is sophisticated in art, music, or literature |

Please check: Did you write a number in front of each statement?

BFI scale scoring (“R” denotes reverse-scored items):

- Extraversion: 1, 6R, 11, 16, 21R, 26, 31R, 36
 Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42
 Conscientiousness: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R
 Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39
 Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R,

5. Table C-5. Chapters 6, 7: Measuring non-cognitive/motivational factors – Constructs and items from the questionnaires

Construct	Items	Description
Self-Efficacy (SE) - Bandura (2006)	SE1	I remember well information presented in class and textbooks
	SE2	I get myself to study when there are other interesting things to do
	SE3	I finish my homework assignments by deadlines
Goal Expectancy (GE) - Terzis & Economides (2011)	GE1	Courses’ preparation was sufficient for the test
	GE2	My personal preparation for the test was sufficient
	GE3	My performance expectations for the test

6. Table C-6. Chapter 7: Multiple Comparisons (Bonferroni test) for the motivational factors

Dependent Variable			Mean Difference (I-J)	Std. Error	Sig.	95% CI	
						Lower Bound	Upper Bound
Goal-expectancy	1	2	-2.38738*	.13805	.000	-2.7245	-2.0502
		3	-1.27374*	.13412	.000	-1.6013	-.9462
	2	1	2.38738*	.13805	.000	2.0502	2.7245
		3	1.11364*	.11968	.000	.8213	1.4059
	3	1	1.27374*	.13412	.000	.9462	1.6013
Self-efficacy	1	2	-1.11364*	.11968	.000	-1.4059	-.8213
		3	-2.57711*	.11681	.000	-2.8624	-2.2918
	2	1	2.57711*	.11681	.000	2.2918	2.8624
		3	1.18352*	.10127	.000	.9362	1.4309
	3	1	1.39359*	.11349	.000	1.1164	1.6708
		2	-1.18352*	.10127	.000	-1.4309	-.9362

*. The mean difference is significant at the 0.05 level.

7. Table C-7. Chapter 7: Multiple Comparisons (Bonferroni test) for the help-seeking factors

Dependent Variable			Mean Difference (I-J)	Std. Error	Sig.	95% CI	
						Lower Bound	Upper Bound
Time-spent on Analytics Visualizations Viewing	1	2	-.33997*	.02216	.000	-.3941	-.2859
		3	-.69655*	.02667	.000	-.7617	-.6314
	2	1	.33997*	.02216	.000	.2859	.3941
3		-.35658*	.02572	.000	-.4194	-.2938	
3		1	.69655*	.02667	.000	.6314	.7617
Frequency of Analytics Visualizations Requests	1	2	.35658*	.02572	.000	.2938	.4194
		3	-.22446*	.02680	.000	-.2899	-.1590
	2	1	-.33261*	.03226	.000	-.4114	-.2538
3		.22446*	.02680	.000	.1590	.2899	
3	1	-.10816*	.03111	.002	-.1841	-.0322	
	2	.33261*	.03226	.000	.2538	.4114	
			.10816*	.03111	.002	.0322	.1841

*. The mean difference is significant at the 0.05 level.

8. Table C-8. Chapter 8: Measuring perceived usefulness of the visualizations- Construct and items from the questionnaire

Construct	Items	Description
Perceived Usefulness of the visualizations (PUV) - Venkatesh, Morris, Davis, & Davis (2003)	PUV1	The visualizations provided meaningful information about the tasks
	PUV2	The visualizations helped me better allocate time and effort on the tasks
	PUV3	The visualizations confused me
	PUV4	I found the visualizations helpful for my performance

9. Table C-9. Chapter 10: Measuring self-regulation in online self-assessment – Constructs and items from the questionnaires

Construct	Items	Description
Effort-Regulation (ER) - MSLQ; Pintrich et al. (1993)	ER1	I work hard to do well in this class even if I don't like what we are doing
	ER2	Even when course materials are dull and uninteresting, I manage to keep working until I finish
	ER3	I often feel so lazy or bored when I study for this class that I quit before I finish what I planned to do.
Goal Expectancy (GE) – CBAAM; Terzis & Economides (2011)	GE1	Courses' preparation was sufficient for the test
	GE2	My personal preparation for the test was sufficient
	GE3	My performance expectations for the test
Help-Seeking (HS) - OSLQ; Barnard et al. (2009)	HS1	I am persistent in getting help from the instructor through e-mail.
	HS2	I find someone who is knowledgeable in course content so that I can consult with him or her when I need help.
	HS3	I share my problems with my classmates online so we know what we are struggling with and how to solve our problems
Time-Management (TM) - OSLQ; Barnard et al. (2009)	TM1	I allocate extra studying time for my online courses because I know it is time-demanding
	TM2	I try to schedule the same time every day or every week to study for my online courses, and I observe the schedule
	TM3	Although we don't have to attend daily classes, I still try to distribute my studying time evenly across days.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (Vol. 5). Greenwich, CT: Information Age Publishing.

Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S.-L. (2009). Measuring self-regulation in online and blended learning environments. *The Internet and Higher Education*, 12(1), 1–6. <http://dx.doi.org/10.1016/j.iheduc.2008.10.005>

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research*, 2nd ed. (pp. 102–138). New York, NY, US: Guilford Press

Macan, T. H., Shahani, C., Dipboye, R. L., & Phillips, A. P. (1990). College students' time management: Correlations with academic performance and stress. *Journal of Educational Psychology*, 82(4), 760-768.

Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer based assessment. *Computers & Education*, 56(4), 1032–1044. <http://dx.doi.org/10.1016/j.compedu.2010.11.017>

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3), 425–478.