Full length article

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

Zacharoula Papamitsiou*

papamits@uom.edu.gr

Anastasios A. Economides

economid@uom.gr

IPPS in Information Systems, University of Macedonia, Egnatia Street 156, 546 36 Thessaloniki, Greece

*Corresponding author.

Abstract

Personalizing computer-based testing services to examinees can be improved by considering their behavioral models. This study aims to contribute towards deeper understanding the examinee's time-spent and achievement behavior during testing according to the five personality traits by exploiting assessment analytics. Further, it aims to investigate assessment analytics appropriateness for classifying students and generating enhanced student models to guide personalization of testing services. In this study, the LAERS assessment environment and the Big Five Inventory were used to track the response times of 112 undergraduate students and to extract their personality traits respectively. Partial Least Squares was used to detect fundamental relationships between the collected data, and Supervised Learning Algorithms were used to classify students. Results indicate 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. Moreover, the ensemble RandomForest method provides accurate classification results, indicating that a time-spent driven description of students' behavior could have added value towards dynamically reshaping the respective models. Further implications of these findings are also discussed.

Keywords: Assessment analytics; BFI; Computer-based testing; Personality traits; Student behavior modelling; Supervised classification

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, Muñoz Merino, Pardo, & Kloos, 2013; Papamitsiou & Economides, 2016; Triantafillou, Georgiadou, & Economides, 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, He, & Yen, 2012; Blikstein, 2011; Shih, Koedinger, & Scheines, 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, 2006), motivation, emotions/feelings, self-regulation and self-explanation (Peña & Kayashima, 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 et al. (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, Karapistoli, & Economides, 2016; Xiong, Pardos, & Heffernan, 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) defined time management behavior as "behaviors that aim at achieving an effective use of time while performing certain goal-directed activities" (p. 36). 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).

1.1 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 (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; MacCann 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 Hoye and 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.

1.2 Objectives, research questions and suggested approach

This paper'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?

In order to answer these research questions we conducted an experimental study with the LAERS assessment environment (please, see section 2.1). One hundred and twelve (112) undergraduate students from a Greek University enrolled in a CBT procedure. Partial Least Squares (PLS) was used to explore the relationships between the included factors and evaluate the structural and measurement model, and Supervised Learning Classification algorithms were used to compare the obtained classification results based on students' level of achievement, i.e. using as class labels the students' score classes. The low misclassification rates are indicative of the accuracy of the applied method. Thus, temporal factors that imply students' behavior should be further explored regarding their added value towards modelling test-takers and dynamically reshaping the respective models to support time-management for increasing achievement during CBT.

The rest of the paper is organized as follows: next section briefly presents the LAERS assessment environment used in this study, as well as prior results from exploring the BFI with time-driven assessment analytics, that are strongly associated with the work presented in this paper. Section 3 describes the research model and develops the research hypotheses, as well as the core concepts of the student models. Section 4 explains the experiment methodology and section 5 demonstrates the results. In section 6, we elaborate on our findings, section 7 presents potential implications, and finally, section 8 focuses on the conclusions of this study, and describes our future work plans.

2 The Learning Analytics & Educational Recommender System assessment environment & Temporal Learning Analytics

2.1 The LAERS assessment environment

The Learning Analytics and Educational Recommender System (LAERS) is a CBA system developed to exploit assessment analytics to automate the provision of adaptive/personalized assessment services (Papamitsiou & Economides, 2013). The standard version of LAERS consists of two components, and integrates a testing unit and a tracker that logs the students' interaction data. The first component (i.e., the testing unit) consists of two modules: an item bank (a database) and the testing module (that operates in two states: the fixed and the adaptive –not used in this study). The testing module implements the interface that displays the multiple choice quiz tasks delivered to students separately and one-by-one. In the fixed state, the students can temporarily save their answers on the tasks, and they can change their initial choice by selecting the task to re-view from the list underneath. They submit the quiz answers only once, whenever they estimate that they are ready to do so, within the duration of the test. The second component of the system (i.e., the tracker) records the students' interaction data during testing. In a log file it tracks students' time-spent on handling the testing items, distinguishing it between the time on correctly and wrongly answered items. In the same log file, it also logs the times the students reviewed each item, the times they changed their answers, and the respective time-spent during these interactions. The overall logged features/attributes of students' activity are listed in Table 1.

Table 1 Features from the raw log files.

alt-text: Table 1

Feature	
1. Student ID	8. The task the student works on
2. The answer the student submits	9. The correctness of the submitted answer
3. The timestamp the student starts viewing a task	10. The timestamp the student chooses to leave the task (saves an answer)
4. The total time the student spends on viewing the tasks and submitting the correct answers	11. The total time the student spends on viewing the tasks and submitting the wrong answers
5. The idle time the student spends viewing each task (not saving an answer)	12. How many times the student reviews each task
6. The idle time the student spends reviewing each task	13. How many times the students change the answer they submit for each task
7. The student's total idle time on task	14. The student's total active time on task

In the standard version of LAERS, a pre-test questionnaire to measure each student's goal-expectancy (GE) (a measure of student's goal orientation and perception of preparation) was embedded. The items that measure GE were proposed in Computer Based Assessment Acceptance Model (CBAAM) (Terzis & Economides, 2011) and include: a) GE1: Courses' preparation was sufficient for the CBA, b) GE2: My personal preparation for the CBA, and c) GE3: My performance expectations for the CBA. These items were measured in a seven point Likert-type scale with 1 = strongly disagree to 7 = strongly agree.

For the needs of the current study, in order to extract the students' personality traits the BFI was also embedded into LAERS in the form of a post-test questionnaire (in order not to distract students' attention before taking the

exams). 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. We selected BFI, because it has been known for its reliability, validity and clear factor structure (e.g. Srivastava, John, Gosling, & Potter, 2003).

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.

2.2 Temporal Learning Analytics (TLA)

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 (e.g., Papamitsiou & Economides, 2014a; Papamitsiou, Terzis, & Economides, 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 spent more time on correctly answering the questions, while unconfident students (i.e. with lower level of certainty) will spent more time and finally will submit the wrong answers (Papamitsiou & Economides, 2014a). 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 Fig. 1.

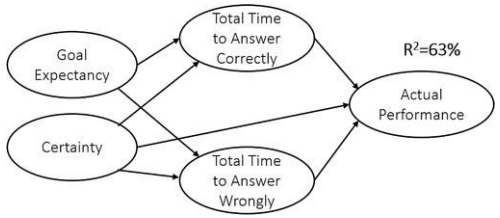

Fig. 1 TLA for predicting performance during CBT (Papamitsiou & Economides, 2014a).

alt-text: Fig. 1

Moreover, Papamitsiou and Economides (2014b) 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 modelling student behavior in learning activities (Papamitsiou et al., 2016).

3 Research model and hypothesis - concepts of student models

As stated in the previous section, goal-expectancy (GE) is a variable which measures goal orientation regarding the use of a CBA. 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. Table 2 synopsizes the variables participating in the model, while Fig. 2 illustrates the overall causal relationships between them.

Table 2 List of variables participating in the model: acronym, description and type.

alt-text: Table 2

Variable	Description	Type
TTAC	Total time to answer correctly	Simple – computed from actual data
TTAW	Total time to answer wrongly	Simple – computed from actual data

GE	Goal expectancy	Latent – measured via questionnaire
CERT	Level of certainty	Latent – composed from actual data
AP	Actual performance	Simple – computed from actual data
E	Extraversion	Latent – measured via questionnaire
A	Agreeableness	Latent – measured via questionnaire
C	Conscientiousness	Latent – measured via questionnaire
N	Neuroticism	Latent – measured via questionnaire
O	Openness to Experience	Latent – measured via questionnaire

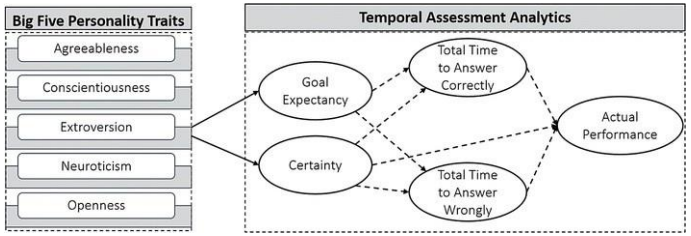

Fig. 2 Overall research model and variables relationships.

alt-text: Fig. 2

In Fig. 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.

3.1 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 (A) will have a positive effect on goal-expectancy (GE)

H2 Agreeableness (A) will have a positive effect on certainty (CERT)

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 (E) will have a positive effect on goal-expectancy (GE)

H4 Extroversion (E) will have a negative effect on certainty (CERT)

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 (C) will have a positive effect on goal-expectancy (GE)

H6 Conscientiousness (C) will have a positive effect on certainty (CERT)

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 Heggestad'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 (N) will have a negative effect on goal-expectancy (GE)

H8 Neuroticism (N) will have a negative effect on certainty (CERT)

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, Gijselaers, 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 (O) will have a positive effect on goal-expectancy (GE)

H10 Openness to experience (O) will have a negative effect on certainty (CERT)

The research model and hypotheses are illustrated in Fig. 3.

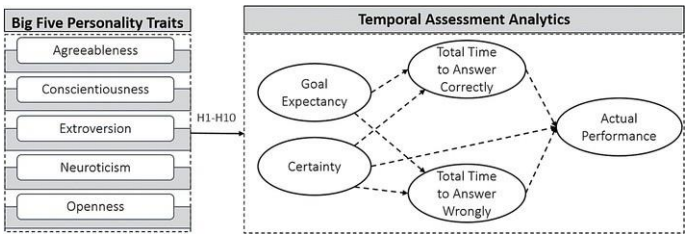

Fig. 3 Research model and hypothesis.

alt-text: Fig. 3

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.

3.2 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.

According to these hypotheses and assumptions, the descriptions of the five classes of achievers during CBT are synopsized in Table 3. 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.

Table 3 Description of achievers' classes and their characteristics.

alt-text: Table 3

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 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.

4 Methodology

4.1 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 20–25 students attended the midterm exams of the Management Information Systems II course (related to databases, telecommunications and e-commerce), for 50 min each group, on May 18th, 2016, at the University computer laboratory.

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 & Krathwohl, 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. 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.

4.2 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; Sellin, 1989; Tenenhaus, Vinzi, Chatelin, & Lauro, 2005). PLS allows comparisons between multiple response variables and multiple explanatory variables (Tenenhaus, 1998) and is a statistical technique for estimating and testing causal dependencies between latent variables. Our decision to use PLS instead of an ordinary least-squares regression method (like Hierarchical Linear Modelling, used in Bergstrom, Gershon, and Lunz (1994) for example) was based on our aim to reduce the predictors (complex constructs) to a smaller set of uncorrelated components and perform least-squares regression on these components, instead of on the original data. 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 our model, the most complex predictor is O with ten items (see section 3.1), 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, Higgins, & Thompson, 1995; Fornell & Larcker, 1981; Henseler, Ringle, & Sinkovics, 2009). Cronbach's a and composite reliability (CR) are used to confirm reliability of the measurement model, and they both have to be higher than 0.7 (Tenenhaus et al., 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 (Fornell & Cha, 1994). For the measurement and the structural model we used SmartPLS 3.0 (Ringle, Wende, & Becker, 2015).

4.3 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; Cristianini & Shawe-Taylor, 2000). 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, Steinbach, & Kumar, 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, 1996, 2001; Tan et al., 2005).
- Classification rule mining aims to discover a small set of rules in the dataset to form an accurate classifier (e.g., Breiman, Friedman, Olshen, & Stone, 1984). 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; Tan et al., 2005). 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 (Mitchell, 1997). 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 (Hall et al., 2009).

5 Results

5.1 Structural and measurement model - hypothesis testing

The results support the measurement model. Table 4 displays the items' reliabilities (Cronbach's alpha, C.R), AVE and factor loadings and confirms convergent validity for the latent constructs.

Table 4 Results for the latent constructs of the measurement model.

alt-text: Table 4

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
AP	1.000	1.00	1.00	1.00

[a] Indicates an acceptable level of reliability and validity.

Table 5 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 criterion (Fornell & Larcker, 1981), 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 Discriminant validity for the measurement model.

alt-text: Table 5

Construct	GE	CERT	TTAC	TTAW	E	A	C	N	O	AP
GE	0.857									

CERT	0.252	0.901								
TTAC	0.390	0.240	1.000							
TTAW	−0.415	−0.177	−0.302	1.000						
E	0.512	0.155	0.227	−0.428	0.735					
A	0.364	0.161	0.042	−0.097	0.355	0.714				
C	0.407	0.342	0.385	−0.364	0.345	0.151	0.714			
N	−0.134	−0.216	−0.144	0.065	0.018	−0.008	−0.115	0.721		
O	0.245	−0.069	0.217	−0.236	0.553	0.237	0.275	−0.050	0.728	
AP	0.645	0.340	0.773	−0.561	0.394	0.152	0.552	−0.114	0.257	1.000

A bootstrap procedure with 3000 resamples was used to test the statistical significance (t-value) of the path coefficients (β) in the model. Table 6 summarizes the results for the hypotheses.

Table 6 Hypothesis testing results.

alt-text: Table 6

Hypothesis	Path	β	t	P	Result
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

*$p < 0.05$.

As seen from Table 6, 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 (AP) are depicted in Table 7. According to these results, the suggested model explains almost the 73% of the variance in AP.

Table 7 R^2, Q^2 and Direct, Indirect and Total effects.

alt-text: Table 7

Dep. Variable	R^2	Q^2	Indep. Variables	Dir. effect	Indir. effect	Total effect	t-value	P-value
AP	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
			O		−0.110	−0.110	1.850	0.064
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		_0.396*	4.622	0.000
			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

*$p < 0.05$.

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 7 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 Fig. 4. This figure illustrates the path coefficients for the initial hypotheses of the research model.

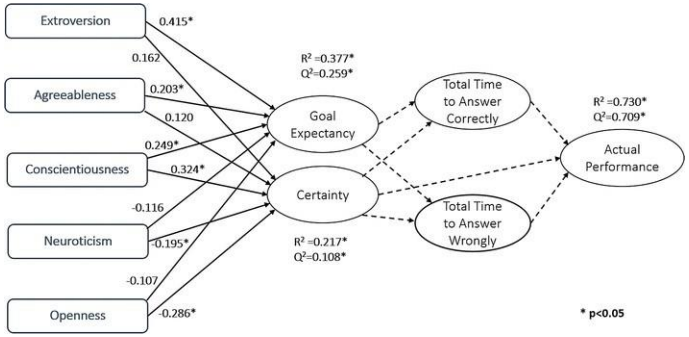

Fig. 4 Path coefficients of the research model and overall variance (R^2).

alt-text: Fig. 4

5.2 Classification results

Table 8 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 8 A summary of the classification approach.

alt-text: Table 8

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 (Hall & Holmes, 2003).

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

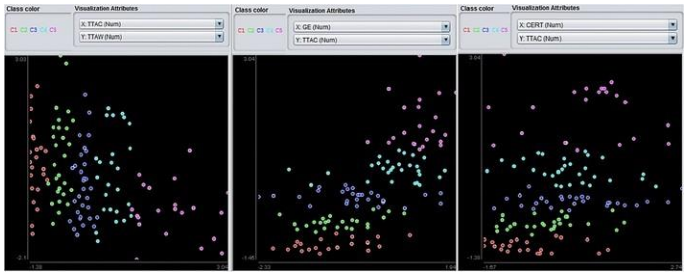

Fig. 5 Graphical exploratory analysis on classes' characteristics: (a) the five classes according to their time-spent, (b) the five classes according to goal-expectancy, and (c) the five classes according to their level of certainty.

alt-text: Fig. 5

Table 9 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 9 Performance metrics for cross-validation 10% with seven features.

alt-text: Table 9

Test Set Size	|cvpartition| = 10% (k-fold = 10)			
Classifier	SVM	NB	RF	JCBA
Resub Error	0.29	0.30	0.22	0.34
True Test Error[a]	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

[a] 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.

6 Discussion

6.1 RQ1: 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 8, 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 et al. (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, Nguyen, Basuray, & Darrow, 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 & Heggestad, 1997). Furthermore, according to Van Hoye and 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 and Heggestad (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 et al. (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 and 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 ($\beta = -0.110$, $t = 1.850$, $p = 0.064$).

6.2 RQ2: How accurately can we classify the students during testing according to their personality traits and behavior expresse 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 (e.g. Papamitsiou et al., 2016; Papamitsiou et al., 2014; Shih et al., 2008; Xiong et al., 2011).

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 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 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 "straggling 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 3 (please see section 3.2), are confirmed. However, the assumptions that were not confirmed are reconsidered and synopsized as follows (Table 10).

Table 10 Achievers' classes and their characteristics (reconsideration).

alt-text: Table 10

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 3) 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.

7 Implications

The findings presented in this paper, 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.

7.1 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 finding 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 spent more time to view the questions again and again before saving an answer, trying to assure that they will submit the correct answer. This mean 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 straggling 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.

7.2 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).

8 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 paper 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 modelling with a notion of dynamics, allowing for adaptive question sequencing.

References

Abdous M., He W. and Yen C.-J., Using data mining for predicting relationships between online question theme and final grade, Educational Technology & Society **15** (3), 2012, 77–88.

Alpaydin E., Introduction to machine learning, 2010, MIT Press.

Anderson L.W. and Krathwohl D.R., A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, 2001, Longman; New York.

Barclay D., Higgins C. and Thompson R., The partial least squares approach to causal modelling: Personal computer adoption and use as an illustration, Technology Studies **2** (1), 1995, 285–309.

Barrick M.R., Mount M.K. and Judge T.A., 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**, 2001, 9–30.

Belk M., Germanakos P., Fidas C. and Samaras G., A personalization method based on human factors for improving usability of user authentication tasks, In: 22nd int. conf. on user modeling, adaptation and personalization, 2014, 13–24.

Bennett R.E., Reinventing assessment: Speculations on the future of large scale educational testing, 1998, Educational Testing Service, Policy Information Center; Princeton, NJ.

Bergstrom B., Gershon R.C. and Lunz M.E., Computerized Adaptive Testing exploring examinee response time using hierarchical linear modeling, In: Annual meeting of the national council on measurement in education, 1994, New Orleans.

Biderman M.D., Nguyen N.T. and Sebren J., Time-on-task mediates the conscientiousness–performance relationship, Personality and Individual Differences **44** (4), 2008, 887–897.

Bidjerano T. and Dai D., The relationship between the big-five model of personality and self-regulated learning strategies, Learning and Individual Differences **17**, 2007, 69–81.

Bipp T., Steinmayr R. and Spinath B., Personality and achievement motivation: Relationship among Big Five domain and facet scales, achievement goals, and intelligence, Personality and Individual Differences **44** (7), 2008, 1454–1464.

Blikstein P., Using learning analytics to assess students' behavior in open-ended programming tasks, In: Long P., Siemens G., Conole G. and Gasevic D., (Eds.), Proceedings of the 1st international conference on learning analytic and knowledge, 2011, ACM; New York, NY, 110–116.

Boroujeni A., Roohani A. and Hasanimanesh A., The impact of extroversion and introversion personality types on EFL learners' writing ability, Theory & Practice In Language Studies **5** (1), 2015, 212–218.

Breiman L., Bagging predictors, Machine Learning **24**, 1996, 123–140.

Breiman L., Random forests, Machine Learning **45** (1), 2001, 5–32.

Breiman L., Friedman J., Olshen R. and Stone C., Classification and regression trees, 1984, Wadsworth; Belmont.

Casal G.B., Caballo V.E., Cueto E.G. and Cubos P.F., Attention and reaction time differences in introversion-extraversion, Personality and Individual Differences **11**, 1990, 195–197.

Chamorro-Premuzic T. and Furnham A., Personality and intellectual competence, 2005, Lawrence Erlbaum; London.

Chamorro-Premuzic T., Furnham A. and Lewis M., Personality and approaches to learning predict preference for different teaching methods, Learning and Individual Differences **17**, 2007, 241–250.

Chatzopoulou D.I. and Economides A.A., Adaptive assessment of student's knowledge in programming courses, Journal of Computer Assisted Learning **26** (4), 2010, 258–269.

Chin W.W., The partial least squares approach to structural equation Modeling, In: Marcoulides G.A., (Ed), Modern business research methods, 1998, Lawrence Erlbaum Associates; Mahwah, NJ, 295–336.

Claessens B.J.C., van Eerde W., Rutte C.G. and Roe R.A., A review of the time management literature, Personnel Review **36** (2), 2007, 255–276.

Cohen J., Statistical power analysis for the behavioral sciences, 2nd ed., 1988, Erlbaum; Hillsdale, NJ.

Conard M.A., Aptitude is not enough: How personality and behavior predict academic performance, Journal of Research in Personality **40**, 2006, 339–346.

Cortes C. and Vapnik V., Support-vector networks, Machine Learning **20** (3), 1995, 273–297.

Costa P.T., Jr. and McCrae R.R., NEO-PI-R: Professional manual, 1992, Psychological Assessment Resources; Odessa, FL.

Cristianini N. and Shawe-Taylor J., An introduction to Support Vector Machines and other kernel-based learning methods, 2000, Cambridge University Press; London, UK.

De Raad B. and Schouwenburg H.C., Personality traits in learning and education, European Journal of Personality **10**, 1996, 185–200.

Devaraj S., Easley R. and Crant J., How does personality matter? Relating the five-factor model to technology acceptance and use, Information Systems Research **19** (1), 2008, 93–105.

Dickman S.J. and Meyer D.E., Impulsivity and speed accuracy trade-offs in information processing, Journal of Personality and Social Psychology **54**, 1988, 274–290.

Digman J.M., Personality structure: Emergence of the five-factor model, Annual Review of Psychology **41**, 1990, 417–440.

Digman J.M., Higher-order factors of the big five, Journal of Personality and Social Psychology **73**, 1997, 1246–1256.

Duda R.O., Hart P.E. and Stork D.G., Pattern classification, 2nd ed., 2000, Wiley-Interscience.

Economides A.A., Adaptive orientation methods in computer adaptive testing, In: Proc. E-Learn 2005 world conference on E-learning in corporate, government, healthcare, and higher education, 2005, 1290–1295.

Ellis C., Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics, British Journal of Educational Technology **44** (4), 2013, 662–664.

Fornell C. and Cha J., Partial least squares, In: Bagozzi R.P., (Ed), Advanced methods of marketing research, 1994, Blackwell Publishers; Cambridge, MA, 52–78.

Fornell C. and Larcker D.F., Evaluating structural equations models with unobservable variables and measurement error, Journal of Marketing Research **18** (1), 1981, 39–50.

Furnham A., Christopher A., Garwood J. and Martin N.G., Ability, demography, learning style, and personality trait correlates of student preference for assessment method, Educational Psychology: An International Journal of Experimental Educational Psychology **28** (1), 2008, 15–27.

Geisser S., A predictive approach to the random effects model, Biometrika **61** (1), 1974, 101–107.

Gellatly I.R., Conscientiousness and task performance: Test of cognitive process model, Journal of Applied Psychology **81**, 1996, 474–482.

Graziano W.G. and Eisenberg N.H., Agreeableness: A dimension of personality, In: Hogan R., Johnston J. and Briggs S., (Eds.), Handbook of personality psychology, 1997, Academic Press; San Diego, CA, 795–824.

Guyon I. and Elisseeff A., An introduction to variable and feature selection, Journal of Machine Learning Research **3**, 2003, 1157–1182.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I.H., The WEKA data mining software: An update, SIGKDD Explorations **11** (1), 2009.

Hall M.A. and Holmes G., Benchmarking attribute selection techniques for discrete class data mining, IEEE Transactions on Knowledge and Data Engineering **15** (6), 2003, 1437–1447.

Hamdan A., Nasir R., Rozainee W. and Sulaiman W.S., Time management does not matter for academic achievement unless you can cope, International Proceedings of Economics Development and Research **78**, 2013, 22–26.

Henseler J., Ringle C.M. and Sinkovics R.R., The use of partial least squares path modelling in international marketing, Advances in International Marketing **20**, 2009, 277–319.

John O.P. and Srivastava S., The Big Five trait taxonomy: History, measurement, and theoretical perspectives, In: Pervin Lawrence A. and John Oliver P., (Eds.), Handbook of Personality: Theory and research, 1999, 102–138.

Jo I.H., Kim D. and Yoon M., Constructing proxy variables to measure adult learners' time management strategies in LMS, Educational Technology & Society **18** (3), 2015, 214–225.

Judge T.A. and Ilies R., Relationship of personality to performance motivation: A meta-analytic review, Journal of Applied Psychology **87**, 2002, 797–807.

Kanfer R. and Heggestad E.D., Motivational traits and skills: A person-centered approach to work motivation, Research in Organizational Behavior **19**, 1997, 1–56.

Kelly W.E. and Johnson J.L., Time use efficiency and the five-factor model of personality, Education **125** (3), 2005, 511–515.

Kotsiantis S. and Kanellopoulos D., Discretization techniques: A recent survey, GESTS International Transactions on Computer Science and Engineering **32** (1), 2006, 47–58.

Leony D., Muñoz Merino P.J., Pardo A. and Kloos C.D., Provision of awareness of learners' emotions through visualizations in a computer interaction-based environment, Expert Systems with Applications **40** (13), 2013, 5093–5100.

Liu B., Hsu W. and Ma Y., Integrating classification and association rule mining, In: Proceedings of the 4th int. Conf. on knowledge discovery and data mining (KDD-98), 1998, AAAI Press, 80–86.

MacCann C., Fogarty G.J. and Roberts R.D., Strategies for success in education: Time management is more important for part-time than full-time community college students, Learning and Individual Differences **22**, 2012, 618–623.

McCalla G., The central importance of student modeling to intelligent tutoring, In: Costa E., (Ed), New directions for intelligent tutoring systems, 1992, Springer Verlag; Berlin.

McCrae R.R., Social consequences of experiential openness, Psychological Bulletin **120**, 1996, 323–337.

McCrae R.R. and John O.P., An introduction to the five-factor model and its applications, Journal of Personality **60**, 1992, 175–215.

Mitchell T., Machine learning, 1997, Mcgraw-Hill; New York.

Mitrovic A. and Martin B., Evaluating the effects of open student models on learning, In: 2nd international conference on adaptive hypermedia and adaptive web-based systems, 2006, 296–305.

Murphy K. and Cleveland J., Understanding performance appraisal: Social, organizational and goal-oriented perspectives, 1995, Sage; Newbury Park, CA.

Nonis S.A. and Hudson G.I., Academic performance of college students: Influence of time spent studying and working, Journal of Education for Business **81** (3), 2006.

Papamitsiou Z. and Economides A.A., Towards the alignment of computer-based assessment outcome with learning goals: The LAERS architecture, In: Proc. of the IEEE conf. on e-Learning, e-Management and e-Services, 2013, 13–17.

Papamitsiou Z. and Economides A.A., Students' perception of performance vs. actual performance during computer-based testing: A temporal approach, In: 8th int. technology, education and development conference, 2014a, 401–411.

Papamitsiou Z. and Economides A.A., The effect of personality traits on students' performance during computer-based testing: A study of the big five inventory with temporal learning analytics, In: 14th IEEE int. conf. on advanced learning technologies, 2014b, 378–382.

Papamitsiou Z. and Economides A.A., An Assessment Analytics Framework (AAF) for enhancing students' progress, In: Caballé S. and Clarisó R., (Eds.), Intelligent data-centric systems: Formative assessment, learning data analytics and gamification, 2016, Academic Press; Boston, 117–133.

Papamitsiou Z., Karapistoli E. and Economides A.A., 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), 2016, 299–303.

Papamitsiou Z., Terzis V. and Economides A.A., Temporal Learning Analytics during computer based testing, In: 4th int. conference on learning analytics and knowledge, 2014, 31–35.

Peña-Ayala A., Educational data mining: A survey and a data mining-based analysis of recent works, Expert Systems With Applications **41** (4), 2014, 1432–1462.

Peña A. and Kayashima M., Improving students' meta-cognitive skills within intelligent educational systems: A review, In: 6th international conference on foundations of augmented cognition, 2011, 442–451, Orlando, Florida, USA.

Pervin L.A. and John O.P., Personality theory and research, 8th ed., 2001, John Wiley & Sons Inc; New York.

Ringle C.M., Wende S. and Becker J.-M., SmartPLS 3, 2015, SmartPLS; Hamburg, Retrieved from http://www.smartpls.com.

Robinson T.N. and Zahn T.P., Preparatory interval effects on the reaction time performance of introverts and extraverts, Personality and Individual Differences **9**, 1988, 749–761.

Romero C., López M.-I., Luna J.-M. and Ventura S., Predicting students' final performance from participation in on-line discussion forums, Computers & Education **68**, 2013, 458–472.

Self J.A., Bypassing the intractable problem of student modeling, In: Frasson C. and Gauthier G., (Eds.), Intelligent-tutoring systems: At the crossroads of AI and education, 1990, Ablex; Norwood, NJ, 107–123.

Sellin N., PLS PATH - Version 3.01. Application manual, 1989, Universitat Hamburg; Hamburg.

Shih B., Koedinger K.R. and Scheines R., A response time model for bottom-out hints as worked examples, In: de Baker R., Barnes T. and Beck J., (Eds.), Proc. 1st international conference on educational data mining, 2008, 117–126.

Srikant R. and Agrawal R., Mining quantitative association rules in large relational tables, In: SIGMOD-96, 1996, 1–12.

Srivastava S., John O.P., Gosling S.D. and Potter J., Development of personality in early and middle adulthood: Set like plaster or persistent change?, Journal of Personality and Social Psychology **84**, 2003, 1041–1053.

Stone M., Cross-Validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society **36** (2), 1974, 111–147.

Tabak F., Nguyen N., Basuray T. and Darrow W., Exploring the impact of personality on performance: How time-on-task moderates the mediation by self-efficacy, Personality and Individual Differences **47** (8), 2009, 823–828.

Tan P.-N., Steinbach M. and Kumar V., Introduction to data mining, 1st ed., 2005, Addison-Wesley Longman Publishing Co., Inc.; Boston, MA, USA.

Tempelaar D.T., Gijselaers W.H., van der Loeff S.S. and Nijhuis J.F.H., 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), 2007, 105–131.

Tenenhaus M., La régression PLS, 1998, Technip; Paris.

Tenenhaus M., Vinzi V.E., Chatelin Y. and Lauro C., PLS path modeling, Computational Statistics & Data Analysis **48**, 2005, 159–205.

Terzis V. and Economides A.A., The acceptance and use of computer based assessment, Computers & Education **56** (4), 2011, 1032–1044.

Terzis V., Moridis C.N. and Economides A.A., How student's personality traits affect computer based assessment acceptance: Integrating BFI with CBAAM, Computers in Human Behavior **28** (5), 2012, 1985–1996.

Thomson D. and Mitrovic A., Towards a negotiable student model for constraint-based ITSs, In: 17th international conference on computers in education, 2009, 83–90, Hong Kong.

Triantafillou E., Georgiadou E. and Economides A.A., The design and evaluation of a computerized adaptive test on mobile devices, Computers & Education **50** (4), 2008, 1319–1330.

Trueman M. and Hartley J., A comparison between the time-management skills and academic performance of mature and traditional-entry university students, Higher Education **32** (2), 1996, 199–215.

Van Hoye G. and Lootens H., Coping with unemployment: Personality, role demands, and time structure, Journal of Vocational Behavior **82** (2), 2013, 85–95.

Vermetten Y.J., Lodewijks H.G. and Vermunt J.D., The role of personality traits and goal orientations in strategy use, Contemporary Educational Psychology **26**, 2001, 149–170.

Watson D. and Clark L.A., Extraversion and its positive emotional core, 1997, Academic Press; San Diego.

Xiong X., Pardos Z. and Heffernan N., An analysis of response time data for improving student performance prediction, In: Proc. KDD 2011 workshop: Knowledge discovery in educational data, held as part of 17th ACM SIGKDD conference on knowledge discovery and data mining in San Diego, 2011.

Zweig D. and Webster J., 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**, 2004, 1693–1708.

Highlights

- We exploit assessment analytics for exploring time-spent and achievement behavior.
- We associate examinees' personality traits with time-spent and achievement behavior.
- Examinees' extraversion and agreeableness positively affect goal-expectancy.
- Conscientiousness positively affects examinee goal-expectancy and level of certainty.
- Examinees' neuroticism and openness have a negative effect on level of certainty.