A temporal estimation of students' on-task mental effort and its effect on students' performance during computer based testing

Zacharoula Papamitsiou, Anastasios A. Economides
Interdepartmental Progr. of Postgraduate Studies in Information Systems, University of Macedonia
Thessaloniki, GREECE
papamits@uom.edu.gr, economid@uom.gr

***Abstract*—Students' on-task mental effort is an important factor of their achievement behavior. Most studies in literature adopt self-reported methods for effort estimation. In this paper we present and evaluate a method for estimating students' on-task mental effort during testing, based ontemporal, user-generated trace data. Our goal is to construct a metric that unobtrusively and seamlessly measures students' on-task mental effort. Additionally, we use this metric in order to investigate the effect of effort expenditure on students' performance during low-stakes computer based testing procedures. For that reason, we examined on-task mental effort –as defined here – with Temporal Learning Analytics. We present the results from an evaluation case study with 259 undergraduate participant students for the Computers II course. The results are encouraging, since the proposed factor significantly improves the general prediction of students' performance.**

Keywords—effort; temporal learning analytics; performance; computer-based testing.

I. Introduction

The students' performance during testing is strongly associated with their perception of task difficulty and on-task mental effort, as well as with their perception of preparation, perceived self-efficacy and expertise (e.g., [1], [2]). Most studies correlate students' perception of performance with self-confidence, task difficulty and motivation ([3], [4], [5], [6]). Theories dealing with motivation and achievement associate students' performance with task difficulty and effort. For example, according to Brehm's theory, which focuses on the intensity of motivation ([7], [8]), effort investment is directly dependent on task difficulty. Specifically, the amount of effort expended in performing a task is predicted to increase proportionally with the level of perceived task difficulty. In a sense, the higher the subjective perceived task difficulty level, the more the participant's effort expenditure on the task. The reasoning behind this claim is that high difficulty tasks evoke high effort exertion if the individual is motivated to succeed on the task. The study conducted by Fisher and Noble [9] also supports this hypothesis, as a significant positive relationship between task difficulty and effort was found. Further, Weiner's attribution theory of motivation and achievement supports that learners' current self-perceptions (e.g. perception of performance, goal-expectancy, etc.) will strongly influence the ways in which they will interpret the success or failure of their current efforts [10], [11]. According to him, the most important factors affecting attributions are ability, effort, task difficulty and luck.

It becomes apparent that there is a need to discover how students behave (effort) when dealing with assessment tasks that have different requirements. When instructors provide a set of assessment tasks to their students, they should be aware of the students' comprehensions, their ability level and an estimation of effort needed to accomplish these tasks. Going a step ahead, when an intelligent assessment system executes the same process, it should be able to automatically identify the student's ability level (e.g. predict the student's performance) and match it to the task level, taking into account the required effort for that task.

A. Brief Literature Review on Effort

Kahneman [13] was the first to define effort as an attentional control mechanism within an information processing framework. Moreover, according to Humphreys and Revelle [14], effort is "the motivational state commonly understood to mean trying hard or being involved in a task. Effort is increased when the subject tries harder, when there are incentives to perform well, or when the task is important or difficult".

There are several measures and methods that are used to assess mental workload. Two major categories of these methods have been identified during the literature review: the self-reported methods and the computational methods. The NASA-TLX [15] is one of the most widely known self-reported methods. It uses mental workload scales and consists of six subscales: (i) mental demand, (ii) physical demand, (iii) own performance, (iv) temporal demand, (v) effort, and (vi) frustration. Another well-known subjective rating scale is the SWAT (the Subjective Workload Assessment Technique) approach, consisting of three component factors: Time Load, Mental Effort Load, and Psychological Stress Load [16]. Rubio et al. [17] presented an adequate comparative study of these two methods.

As mentioned before, beyond self-reported methods, there are computational methods as well. Among these the heart rate variability (HRV) [18], the event-related potentials (ERP) [19] and the dual-task methods have been extensively studied. The HRV technique monitors the R-waves of the electrocardiogram

978-1-4799-8706-1/15/$31.00 ©2015 IEEE

curves and uses spectrum analysis of the frequencies [20]. Similarly, ERPs are electrical activity peaks recorded from the brain, averaged in the time domain and time-locked to discrete stimuli (e.g., [21]). ERPs have been used to examine cognitive workload during task performance based on the inverse relationship between cognitive workload of the primary task and performance (e.g., [21], [22]). The dual-task paradigm requires that subjects are interrupted by a tone or a visual image while they are learning, and they are asked to quickly strike a key. The speed with which they react (reaction time) to the interruption is assumed to represent the amount of mental effort they are investing in the primary task [23], [24].

In addition, Wise & Kong [25] introduced a method, the Response Time Effort (RTE), for measuring examinee test-taking effort based on item response time. The initial hypothesis was that unmotivated examinees will answer too quickly (i.e., before they had time to read and fully consider the item). Based on that assumption, they identified two behaviors: the rapid-guessing behavior and the solution behavior. They introduced a threshold to discriminate these two behaviors. Specifically, the threshold value is used in order to clarify whether the student spends time on solving the task or just guesses the possible answer. Given an examinee *j*'s response time, RT*ij*, to item *i*, a dichotomous index of item solution behavior, SB*ij*, is computed as

$SB_{ij} = \begin{Bmatrix} 1, & if\ RT_{ij} \geq T_i \\ 0, & otherwise \end{Bmatrix}$, where T_i is the threshold value for this item.

The index of overall response time effort for examinee j to the test is given by

$RTE_j = \frac{\sum SB_{ij}}{k}$ (1), where k = the number of items in the test.

They introduced a threshold in order to clarify whether the student spends time on solving the task or just guesses the possible answer. In their method, they aggregate the values of item solution behavior of a dichotomous index.

B. Motivation and rationale of the research

The literature review revealed that the issue of accurately estimating mental effort is a challenging research question. Different methods and measures are being adopted under different conditions, as the appropriateness and availability of each one varies from study to study.

However, self-reported measures of effort (like [15] and [16]) are potentially vulnerable to bias through motivational processes, and it is difficult to ascertain the degree to which these factors have influenced a particular set of self-report data. Further, the use of self-report data requires the assumption that examinees truthfully answered the self-report instrument. Studies have found that the reliability of self-reported measure is still in question (e.g., [26]).

Moreover, based on the above, the ERP [19] and HRV [18] techniques require the use of specialized equipment. That explains sufficiently why they cannot be extensively used during typical educational testing procedures.

Another finding from the literature review is that most of the measures and methods use the time dimension as one of the key features of the respective technique. The example of Wise and Kong [25] is the most recent and the one that unobtrusively tracks and explores the temporal dimension of on-task mental effort. However, the determination of the threshold value (which discriminates rapid-guessing from solution behavior) is still arbitrary and should be further explored.

The goal of our suggested method is to deal with these issues. In particular, we propose a method for estimating students' on-task mental effort during testing, based on temporal, user-generated trace data. Our goal is to construct a metric that a) unobtrusively measures students' on-task mental effort, b) is easy to implement and could be seamlessly applied in any Computer Based Testing (CBT) environment, and c) moves beyond subjective perceptions to objectively and directly calculating on-task mental effort, and transcends threshold values determination(i.e., is not based on students' perception of the task effort, neither on some arbitrarily set parameters).This metric could be useful for the instructors as additional information about their students' abilities on each question, in order to adjust the difficulty level accordingly and, even more, to self-assess their instructional design.

More precisely, we suggest that a student's EFF is associated to the ratio of the average total idle time the student spends on overcoming the task to the average total time to answer, and to the ratio of the average total changes of submitted answers to the average total re-views of the questions.

In order to explore the effect of on-task mental effort on student's performance, we conducted a case study with the LAERS assessment environment (see section II). 259 undergraduate students from the Department of Economics at University of Macedonia, Thessaloniki, Greece, attended the midterm examination for the Computers II course (related to databases, information systems and introduction to e-commerce). The participants were examined on 34 multiple choice questions during their midterm progress assessment. We estimated EFF for each participant and examined on-task mental effort with Temporal Learning Analytics (TLA) [12]. In this paper we present the results regarding the effect of on-task effort on student's performance. Initial results are encouraging, indicating that EFF is a good estimator of students' on-task mental effort and a significant predictor of their performance during low-stakes CBT procedures in higher education.

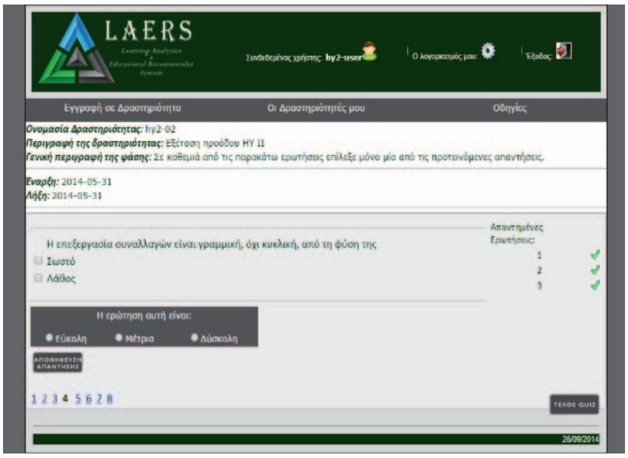
The rest of this paper is organized as follows: in section IIwe briefly present the LAERS assessment environment used in this study, as well as previous results from the TLA approach, that are strongly associated with the work presented in this paper. In section III we introduce the EFF metric.Section IV describes the followed methods for our case study, and in section V we present the findings from our approach.In section VI we discuss on the results and the conclusions from our experimentation.

II. THE LAERS ASSESSMENT ENVIRONMENT AND TEMPORAL LEARNING ANALYTICS

A. *The LAERS Assessment Environment*

The LAERS assessment environment is a Computer Based Assessment (CBA) system that we are developing in order to exploit data-driven decision making research results to automate the accurate prediction of students' performance and provision of adaptive and personalized recommendations as assessment services.

At the first phase of its implementation, we configured a testing mechanism and a tracker that logs students' temporal data. The testing unit displays the multiple choice quiz items/task delivered to students. Each item/task is displayed separately and one-by-one. The students can temporarily save their answers on the items/tasks, before finalizing their decision (by submitting the quiz). The list of items/tasks that have already been answered is displayed alongside the quiz, within the same window, with a green check-icon indicating that the student has saved an answer. The students can also change their initial choice, and save a new answer, by selecting the item/task to re-view from the list underneath, within the same window. During the test, the students can also skip an item/task (either because they are not sure about the answer, or because they think it is too difficult), and answer it (or not) later. In case the students choose not to submit an answer to an item/task, they receives zero points for this item/task. They submit the quiz answers only once, whenever they estimates that they are ready to do so, within the duration of the test.Figure 1 illustrates the student's view of the environment during testing.



Ονομασία Δραστηριότητας: hy2-02
Περιγραφή της δραστηριότητας: Εξέταση προόδου ΗΥ ΙΙ
Γενική περιγραφή της φάσης: Σε καθεμιά από τις παρακάτω ερωτήσεις επίλεξε μόνο μία από τις προτεινόμενες απαντήσεις.

Εναρξη: 2014-05-31
Λήξη: 2014-05-31

Η επεξεργασία συναλλαγών είναι γραμμική, όχι κυκλική, από τη φύση της
□ Σωστό
□ Λάθος

Απαντημένες Ερωτήσεις:
1 ✓
2 ✓
3 ✓

Η ερώτηση αυτή είναι:
● Εύκολη ● Μέτρια ● Δύσκολη

Fig. 1. The LAERS assessment environment.

The second component of the system records students' activity data during testing. In specific, the collected dataset includes the following (for each student): student ID, the item/task the student works on, the answer the student submits, the correctness of the submitted answer, how many times the student views each item/task, how many times the student changes the answer, the timestamp the student starts viewing an item/task, the timestamp the student chooses to leave the item/task (saves an answer), the idle time the student spends viewing each item/task (not saving an answer, but choosing to see another question). We also embedded into the system a pre-test questionnaire in order to measure each student's goal expectancy (GE) (a measure of student goal orientation and perception of preparation; [28]).

The whole system is developed in PHP 5.4, MySQL 5.1 and runs on Apache 2.4 for Windows. Javascript and AJAX and JQuery have also been used for implementing the system's functionalities.

B. *Temporal Learning Analytics for Computer Based Testing*

As stated in the introduction, in this paper we wanted to explore EFF with Temporal Learning Analytics(TLA). In this subsection we briefly review on TLA for computed based testing.

Temporal Learning Analytics are proposed as a complementary dimension of a more concise predictive model in order to interpret students' participation and engagement in assessment activities in terms of "time-spent". More specifically, in [12] the authors hypothesized that students who spend more time for choosing the correct answers (and consequently aggregate more time on the total time to answer correctly -TTAC variable) are more likely to have better performance. On the contrary, students who spend more time for choosing and finally submitting the wrong answers (and therefore aggregate more and more time on the total time to answer wrongly - TTAW variable) are more likely to have lower performance. Moreover, an initial hypothesis was that well prepared students are more likely to answer more questions correctly; therefore the time that they will spend for the correct answers will be higher than the spending time of poorly prepared students. Contrariwise, well prepared students will have fewer wrong answers than poorly prepared students. Consequently, well prepared student will spend less time on questions that finally will answer wrongly.

Results revealed that TTAC and TTAW have a direct positive and a direct negative effect on Actual Performance (AP) respectively. In addition, goal-expectancy (GE) – i.e. the students' self-confidence regarding their study and the assessment and their perception of preparation [28] – was found to be an indirect determinant of AP.The suggested model explains almost 62% of the variance in actual performance [12].

III. THE "EFF" METRIC

Researchers explored the use of response time information in obtaining more accurate proficiency level estimates (e.g., [27]). Additional studies have shown that the temporal interpretation of students' engagement in task-solving during CBT, could be used for predicting their progress [12].

Therefore, we believe that when a student consumes large amounts of time on viewing/reviewing a testing item/task trying to deal with the item/task, this could indicate high on-task mental effort. On the other hand, when a student spends less time/times on viewing/reviewing the testing item/task, this could be indication of less effort expenditure. Therefore, we could support that the total idle time the students spend on the item/task implies their engagement in the task and their effort to understand it and try to answer it. During dealing with the task and saving an answer, this time interval aggregates on total time to answer (TTA) the task. In the case that the students don't save an answer, but just review the item/task or choose another question, the respective time interval aggregates to total idle time (TIT) on the item/task.

Consequently, we assumed that:

If we consider the ratio of TIT to TTA for a specific student, the higher this ratio, the higher the effort.In the opposite case, if this amount is low, it implies that the effort needed is less.

Further, (un-)certainty (i.e. the students' cautiousness during testing in terms of time-spent on answering the quiz) has been explored regarding its capabilities to explain the students' actual performance [2] with satisfactory results. Therefore, we believe that when a student consumes large amounts of time on viewing/reviewing the question (TCV) and often changes his/her answer (TCA) trying to deal with the task, his/her mental effort increases. On the other hand, when a student spends less time/times on viewing/reviewing the question and does not changes his/answer, the effort expenditure is lower.

Thus, we hypothesized that:

If we consider the ratio of TCA to TCV for a specific student, the higher this ratio, the higher the effort.In the opposite case, if this amount is low, it implies that the effort needed is less.

For a student *i* that answers n questions, let us call:

- TTA_i^j the *i* student's total time to answer on question *j*,
- TIT_i^j the *i* student's total idle time on question *j*,
- TCV_i^j the *i* student's total re-views of question *j*, and
- TCA_i^j the *i* student's total number of changes of answers on question *j*.

Then, the student's *i* total time to answer on all questions is $pTTA_i = \sum_j TTA_i^j$, and the total idle time on all questions is $pTIT_i = \sum_j TIT_i^j$. Consequently, the student's *i* average time to answer on all questions is $\overline{pTTA_i} = \frac{pTTA_i}{n}$, and the average idle time on all questions is $\overline{pTIT_i} = \frac{pTIT_i}{n}$ respectively.

In the same way we define as $\overline{pTCA_i} = \frac{pTCA_i}{n}$ the average total number the student *i* changes the answers on all questions, and as $\overline{pTCV_i} = \frac{pTCV_i}{n}$ the average total check views on all questions respectively, where $pTCA_i = \sum_j TCA_i^j$ and $pTCV_i = \sum_j TCV_i^j$ are the total number of changing answers on all questions and the total re-views on all questions.

The variables used and a short definition for each of them is shown in Table I.

TABLE I. VARIABLES USED FOR EFF METRIC CALCULATION

Variable	Definition
$pTTA_i = \sum_j TTA_i^j$	student's *i* total time to answer
$\overline{pTTA_i} = \frac{pTTA_i}{n}$	student's *i* average total time to answer
$pTIT_i = \sum_j TIT_i^j$	student's *i* total idle time
$\overline{pTIT_i} = \frac{pTIT_i}{n}$	student's *i* average idle time
$pTCA_i = \sum_j TCA_i^j$	student's *i* total number of changing answers
$\overline{pTCA_i} = \frac{pTCA_i}{n}$	student's *i* average total number of changing answers
$pTCV_i = \sum_j TCV_i^j$	student's *i* total re-views
$\overline{pTCV_i} = \frac{pTCV_i}{n}$	student's *i* average total check views

Then, for a student *i*, the *EFF*$_i$is:

$$EFF_i = \alpha \frac{\overline{pTIT_i}}{\overline{pTTA_i}} + \beta \frac{\overline{pTCA_i}}{\overline{pTCV_i}} + \gamma \qquad (2).$$

where α, β are the weights for the respective factors and γ is a loading constant. Each one of the parameters entered in the equation (2) is tracked and/or calculated during testing. For simplicity reasons, let us call: $\lambda = \frac{\overline{pTIT_i}}{\overline{pTTA_i}}$, and $\mu = \frac{\overline{pTCA_i}}{\overline{pTCV_i}}$.

In this paper we attempt to interpret the tracked students' actual data during CBT into meaningful information regarding mental effort expenditure on task solving. The goal is to construct a metric that indirectly yet objectively could transform students' testing behaviour into instructor's awareness regarding their understandings of the test content and requirements.

IV. METHODS

A. Research participants and data collection

Data were collected from a total of 259 undergraduate students (108 males [41.7%] and 151 females [58.3%], aged 20-27 years old (M=22.6, SD=1.933, N=259) from the Department of Economics at University of Macedonia, Thessaloniki, Greece. 12 groups of 20 to 25 students attended the midterm exams of the Computers II course (related to databases, information systems and introduction to e-commerce), for 60 minutes each group from 26th to 31st of May 2014, at the University computer laboratory. For the purposes of the examination, we used 34 items/tasks in total, distributed in 6 equivalent tests of 8 multiple choice questions each (some of the questions were common in two tests). Each question had two to four possible answers, but only one was the correct. All questions used in the current case study correspond to the lower three levels of the cognitive domain of Bloom's taxonomy (Remembering, Understanding and Applying) [29]. In order to prevent copying phenomena between students seating next to each other, we cyclically assigned one test (out of the six) to each participant.

The participation to the midterm exams procedure was optional. As external motivation to increase their effort, we set that their score would participate at 30% of their final grade.

B. Research hypotheses, model and data analysis method

Based on previous studies, this paper goes a step further by introducing the temporal estimation of effort with TLA for prediction of actual performance during low-stakes CBT.

In particular, and in agreement with previous results, we believe that the higher the effort exertion during testing, the higher the score a student gets. That is because, the more engaged the student remain during examination, the more possible it is to achieve a higher performance. Thus, we hypothesized that:

H1: *Effort (EFF) will have a positive effect on Actual Performance (AP).*

Moreover, and since goal expectancy is a measure of self-preparation regarding the assessment, we believe that a student who is well prepared will spend higher amount of effort on achieving a higher score compared to a less prepared student. On the contrary, we expect a low prepared student to spend less effort during overcoming the testing items/tasks. Thus, we hypothesized that:

H2: *Goal expectancy (GE) will have a positive effect on effort (EFF).*

To summarize, this paper extends the previously suggested causal model (i.e., introduced in [12]) as follows (Fig. 2):

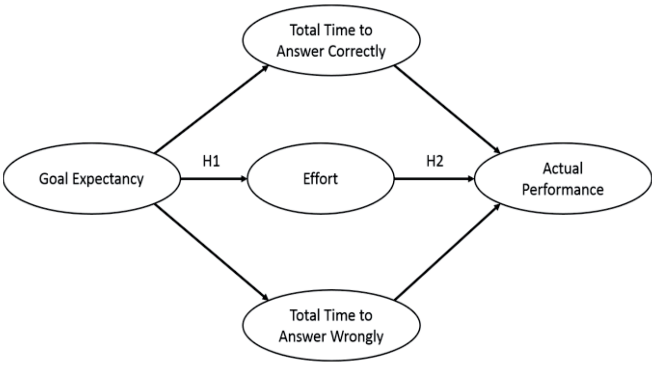


Fig. 2. Research model.

C. Estimation of the EFF metric&Data Analysis techniques

Initially we looked up for correlations in our data and searched for significant outliers. The previously identified predictors werethen entered into a multiple linear regression model, ordered accordingto their importance in determining the outcome variable. There are four principal assumptions which justify the use of multiple linearregression modelsfor purposes of inference or prediction [30] and confirm validity of the regression results. Thus, we checked for a) significant outlier, b) autocorrelations by using the Durbin-Watson statistic, c) homoscedacity with scatterplots, and d) approximately normally distributed residuals (errors) with Normal P-P Plot. Next, and since none of the above assumptions was violated, we performed multiple linear regression in order to determine the coefficients of each one of the parameters of Equation 2 (i.e. α, β, γ) and generate a best-fit predictive model that explains satisfactorily the variance in effort. The next step was to compute the EFF metric for each one of the questions. We analysed our data with SPSS 19.0. In section V.A we present the results from the regression analysis.

In this study we also used Partial least-squares (PLS) analysis for the construction of a path diagram that contains the structural and measurement model showing the causal dependencies between latent variables and the relations to their indicators [31], [32]. 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.

In our model, the most complex construct variable is GE with three items. Further, the largest number of independent variables impacting a dependent variable is also three (TTAC, TTAW, EFF to AP). Thus, the sample for our group (259) is large enough, since it surpasses the recommended value of 30 [31]. Reliability and validity of the measurement model are proved by measuring the internal consistency (Cronbach's a), convergent validity and discriminant validity [33], [34]. In particular, these values should satisfy the following:

- Items' factor loadings on the corresponded constructs have to be higher than 0.7 [31],

- Average Variance Extracted (AVE) have 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 [31], [35],

- Cronbach'sa and composite reliability have to be greater than 0.7 [36].

The structural model is evaluated by examining the variance measured (R^2) by the antecedent constructs. Values of the variance equal to 0.02, 0.13 and 0.26 are considered as small, medium and large respectively [37]. Moreover, a bootstrapping procedure is used in order to evaluate the significance of the path coefficients and total effects, by calculating t-values. In addition, Goodness of Fit (GoF) provides an overall prediction capability of the research model by taking into consideration the measurement and the structural models. GoF values of 0.10, 0.25 and 0.36 are defined as small, medium and large respectively [38].

V. RESULTS

A. The EFF metric calculation - "EFF" vs. "RTE" (Response Time Effort)

The normal P-P plot for the residuals shown that the residuals are normally distributed and consistent to the standardized predicted value (i.e. the *RTE*[25]) as well. Further, the scatterplot for the homoscedasticity check shown that there are no significant outliers in *RTE* values. Pearson correlation coefficients were computed to assess the relationship between RTE and the factors introduced in section III (i.e. the factors that constitute the EFF metric). As seen from Table II, there was a strong positive correlation between RTE and λ, and between RTE and μ.

TABLE II. COEFFICIENTS CORRELATION

		RTE	λ	μ
RTE	Pearson Correlation	1	.646**	.165**
	Sig. (2-tailed)		.001	.008
	n	259	259	259
λ	Pearson Correlation		1	.104
	Sig. (2-tailed)			.096
	n		259	259
μ	Pearson Correlation			1
	Sig. (2-tailed)			
	n			259

**. Correlation is significant at the 0.01 level (2-tailed).

In particular, r = 0.646, n = 259, p = 0.01 (2-tailed), between RTE and λ. Similarly, there was a positive correlationbetween *RTE* and μ, r = 0.165, n = 259, p = 0.08 (2-tailed) respectively. Similarly it was found that $\lambda$and μ are correlated with each other.

Furthermore, the Durbin-Watson statistic was found to be 1.720. This means that the residuals are uncorrelated since this measure is approximately 2.

Tables III and IV illustrate the model summary and the corresponding F test. As seen from these tables, the process generated a best predictive model of effort (F=95.189, p=.00) as a linear combination of the proposed factors.

TABLE III. MODEL SUMMARY^{B}

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.653^{a}	.426	.422	.17416	1.720

a. Predictors: (Constant), λ, μ

b. Dependent Variable: RTE

TABLE IV. ANOVA^{B}

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.715	3	.572	66.083	.000^{a}
	Residual	.259	30	.010		
	Total	1.974	33			

a. Predictors: (Constant), λ, μ

b. Dependent Variable: RTE

The model explains almost 65.3% of the variance in effort.

Finally, Table V provides the necessary information (coefficients) to predict task difficulty from its loading factors.

TABLE V. COEFFICIENTS^{A}

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Std. Coefficients Beta	t	Sig.
1	(Constant)	.181	.060		3.012	.003
	λ	.246	.018	.635	13.347	.000
	μ	.141	.068	.100	2.094	.037

a. Dependent Variable: RTE

Thus, Equation 2 becomes:

$$\mathbf{EFF=0.246\lambda+0.141\mu+0.181}$$

B. Measurement Model and Hypothesis Results

Table VI confirms the adequate values for the measurement model. This table displays the items' reliabilities (Cronbach's a, Composite Reliability), Average Variance Extracted (AVE) and factor loadings and confirms convergent validity.

TABLE VI. RESULTS FOR MEASUREMENT MODEL

Construct Items	Factor Loading (>0.7)^{a}	Cronb. a (>0.7) ^{a}	C.R. (>0.7) ^{a}	AVE (>0.5) ^{a}
GE		0.81	0.89	0.72
GE1	0.75			
GE2	0.88			
GE3	0.91			
EFF	1.00	1.00	1.00	1.00
TTAC	1.00	1.00	1.00	1.00
TTAW	1.00	1.00	1.00	1.00
AP	1.00	1.00	1.00	1.00

a. Indicates an acceptable level of reliability and validity

In addition, Table VII presents the correlation matrix for the measurement model. The diagonal elements are the square root of the AVE of a construct. According to the Fornell-Larcker criterion [35], the AVE of each latent construct should be higher than the construct's highest squared correlation with any other latent construct. Consequently, discriminant validity is confirmed since the diagonal elements are higher than any correlation with another variable.

TABLE VII. DISCRIMINANT VALIDITY FOR THE MEASUREMENT MODEL

Construct	GE	EFF	TTAC	TTAW	AP
GE	**0.85**				
EFF	0.26	**1**			
TTAC	0.36	0.10	**1**		
TTAW	-0.32	-0.35	-0.10	**1**	
AP	0.50	0.40	0.46	-0.72	**1**

A bootstrap procedure with 1000 resamples was used to test the statistical significance of the path coefficients in the model. The results for the hypotheses are summarized in Table VIII. EFF has significant direct positive effect on AP. Moreover GE is a determinant of EFF as well. Thus all the hypotheses were confirmed.

TABLE VIII. HYPOTHESIS TESTING RESULT

Hypothesis	Path	Path coeff.	t value	Results
H1	GE-> EFF	0.26*	5.2	support
H2	EFF->AP	0.16*	4.3	support
	TTAC -> AP	0.40*	10.0	
	TTAW ->AP	-0.64*	18.8	
	GE ->TTAC	0.36*	7.6	
	GE ->TTAW	-0.32*	6.2	

Additional to the direct effects, the structural model includes also indirect effects (Table IX).

TABLE IX. R^2 AND DIRECT, INDIRECT AND TOTAL EFFECTS

Dependent Variable	R^2	Independent Variables	Direct effect	Indirect effect	Total effect
AP	0.71	TTAC	0.40	0.00	0.40*
		TTAW	-0.64	0.00	-0.64*
		GE	0.00	0.38	0.38*
		EFF	0.16	0.00	0.16*

According to the GoF measure [38], the model explains almost the 71% of the variance in AP(Table IX). These results are summarized in Figure 3. This figure illustrates the path coefficients for all initial hypotheses of the research model. It also depicts the overall variance (R^2) explained by the proposed model for actual performance during fixed testing (AP).

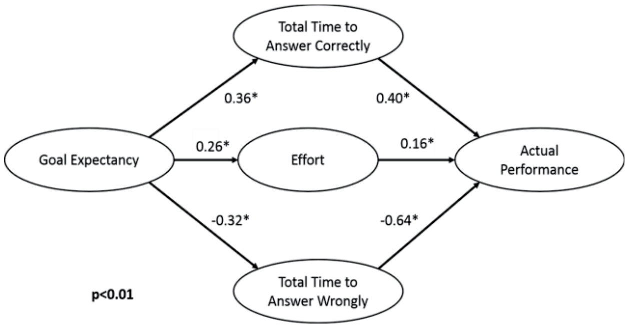


Fig. 3. Path coefficients of the research model.

VI. DISCUSSION AND CONCLUSIONS

When low-stakes assessments are administered to examinees, the degree to which examinees give their best effort is often unclear, complicating the validity and interpretation of the resulting test scores.Without adequate effort, performance is likely to suffer, resulting in the examinees' test score under-representing their true level of proficiency.

The principal idea of the work presented in this paper is to exploit the temporal user-generated trace data regarding the estimation of students' on-task mental effort during testing. Our goal was to construct a metric that a) unobtrusively formulates students' on-task mental effort, b) is easy to implement in any CBT environment, and c) moves beyond subjective perceptions to objectively and directly calculating on-task mental effort, and transcends threshold values determination.

We introduced a data-driven method for effort (EFF) estimation and suggested that a student's EFF is associated to the ratio of average total idle time that the student spends on overcoming the task to the average total time to answer. Moreover, effort is associated to the ratio of the average total changes of the submitted answers to the average total re-views of the questions.

Initially, we assumed that the Response Time Effort (RTE) measure suggested in literature by Wise & Kong [25] is an accurate estimator, despite the fact that it takes under consideration a threshold value that is weakly justified. Next, we performed multiple linear regression analysis in order to check how good predictor of effort is our proposed method compared to the RTE metric. It was found that the EFF metric proposed in this paper explains almost 65.3% of the variance in effort. This indicates that EFF is a good measure of effort, if one accepts that RTE is also a credible measure of effort.

We should mention that in this study we compared the EFF metric only to RTE [25]. It would be interesting to examine the interrelationship between EFF and some of the self-reported measures of on-task mental effort suggested in literature (see section I.A).

Next, we hypothesized that effort would have a positive effect on actual performance and that students' goal expectancy – i.e. their perception of preparation – would also have a positive effect on effort expenditure during testing.

In order to explore our hypotheses, we conducted a case study with the LAERS assessment environment. 259 undergraduate students from the Department of Economics at University of Macedonia, Thessaloniki, Greece, participated in the midterm examination for the Computers II course (related to databases, information systems and introduction to e-commerce). We estimated EFF for each participant. Next, we used the PLS technique to examined on-task mental effort with Temporal Learning Analytics (TLA) and evaluate the measurement and the structural model.

Initial results verified both hypotheses, indicating that EFF is a good estimator of students' on-task mental effort and a significant predictor of their performance during low-stakes CBT procedures in higher education. In particular, the model explains almost the 71% of the variance in actual performance. The total effects of TTAC (0.40), of TTAW (-0.64) and GE (0.38) on AP are strong, while the effect of EFF (0.16) on AP is medium.

Our model also demonstrates that GE is a direct strong determinant of all the temporal variables inserted into the model, including the EFF variable. In this study we confirmed that self-perceptions of goal expectancy have a strong direct positive effect on the effort exertion during testing. This finding is in agreement with Weiner's attribution theory of motivation and achievement [10], which supports that learners' current self-perceptions (e.g. perception of performance, goal-expectancy, etc.) will strongly influence the ways in which they will interpret the success or failure of their current efforts.

It is also important to investigate underlying relations between task difficulty and effort. Task difficulty is suggested to precede high effort. Specifically, the higher the subjective perceived task difficulty level, the more the participant's effort expenditure on the task. The reasoning behind this claim is that high difficulty tasks evoke high effort exertion if the individual is motivated to succeed on the task. The study conducted by Fisher and Noble [9] also supports this hypothesis, as a significant positive relationship between task difficulty and effort was found. This metric could be taken under consideration for the construction of a measure for actual task-difficulty.

Concluding, these findings are encouraging towards the development of an intelligent assessment system that should be able to automatically identify the student's ability level (e.g. predict the students' performance) and match it to the task level, taking into account the required effort for that task.

By embedding this type of metrics in CBT systems, the intelligent assessment system would be able to automatically identify the student's effort, match it to the task level of difficulty and proceed to further adjustments and personalization of the next assessment item, according to the detected student's level.In a sense, that could be useful to the teacher in two ways: a) by exempting the teacher from trying to diagnose the student's effort needed and b) at the same time, by allowing the teacher to be aware of that effort (e.g. through a visualization graph that provides this information to the teacher), and monitor how this effort evolves over time and assessment items. In that way, the added value of this measure, beyond the testing results themselves, is a deeper understanding of students' behavior during testing that could lead to prior critical decisions during the assessment design by the teacher (e.g. exclude items that require a lot of effort for the time-limited testing procedure or identify items that are treated as trivial or least challenging). Moreover, predicting performance through variables that are obtained during the testing process would allow for real-time adaptation of the test to better detect and identify the examinee's level of knowledge or ability.

REFERENCES

[1] Capa, R.L., Audiffren, M., &Ragot, S. (2008). The effects of achievement motivation, task difficulty, and goal difficulty on physiological, behavioral, and subjective effort, *Psychophysiology*,45(5), 859–868.

[2] Papamitsiou, Z. & Economides, A. A. (2014). Students' perception of performance vs. actual performance during computer-based testing: a temporal approach, *In Proc. 8th International Technology, Education and Development Conference (INTED2014)*, 401-411.

[3] Adams, T.M.&Ewen, G.W. 2009. The importance of confidence in improving educational outcomes,*25th annual conference on Distance Learning and Teaching*, Madison.

[4] Bol, L.&Hacker,D.J. 2012. Calibration*, In N. Seel (ed.), Encyclopedia of the Sciences of Learning*, pp. 495-498.

[5] Chevalier, A., Gibbons, S., Hoskins, S., Snell, M.& Thorpe, A. 2008. Students' academic self-perception, *CEE Discussion Papers 0090*, Centre for the Economics of Education, LSE.

[6] Sit, C.H.P, Braman, O. R., Kerr, J.H. and Lindner, K.J. 2013. Motivational style and actual and perceived academic performance of secondary school students in Hong Kong, *School Psychology International*, 34(1), pp. 17-32.

[7] Brehm, J.W. & Self, E.(1989). The intensity of motivation. *Annual Review of Psychology*. 40. 109–131.

[8] Wright, R. A., & Kirby, L. D. (2001). Effort determination of cardiovascular response: An integrative analysis with applications in social psychology. *Advances in Experimental Social Psychology*, 33, 255– 307.

[9] Fisher, C. D. & Noble, C. S. (2004). A within-person examination of correlates of performance and emotions while working. *Human Performance*, 17(2), 145-168.

[10] Weiner, B. (1974). *Achievement motivation and attribution theory*. Morristown, N.J.: General Learning Press

[11] Weiner, B. 1985. An attributional theory of achievement motivation andemotion. *Psych. Review*, 97, pp. 548-573.

[12] Papamitsiou, Z., Terzis, V. & Economides, A.A. (2014). Temporal Learning Analytics during computer based testing, *In proc. of the 4th International Conference on Learning Analytics and Knowledge (LAK'14)*, 31-35.

[13] Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

[14] Humphreys, M.S. &Revelle, W. (1984). Personality, motivation, and performance: a theory of the relationship between individual differences and information processing. *Psychol. Rev*. 91, 153–184.

[15] Hart, S., &Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam: North Holland.

[16] Reid, G. B. &Nygren, T. E. (1988). The subjective workload assessment technique: a scaling procedure for measuring mental workload. In P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (Amsterdam: North-Holand), 185 - 218.

[17] Rubio, S., Diaz, E., Martin, J., & Puente, J. M. (2004). Evaluation of subjective metnal workload: A comparison of SWAT, NASA-TLX, and workload profile methods, *Applied Psychology: An International Review*, 53(1), 61- 72.

[18] Kalsbeek, J. W. H. &Ettema, J. H. (1963): Scored Regularity of the Heart Rate Pattern and the Measurement of Perceptual or Mental Load. *Ergonomics*, 6, 306

[19] Luck, Steven J. (2005). *An Introduction to the Event-Related Potential Technique*. The MIT Press. ISBN 0-262-12277-4.

[20] Luckzak, H., &Laurig, W. (1973). An analysis of heart rate variability, *Ergonomics*, 16, 85-97.

[21] Kramer, A.F., Trejo, L.J. & Humphrey, D. (1995) Assessment of mental workload with task-irrelevant auditory probes. *Biological Psychology*. 40, 83–100.

[22] Sirevaag, E., Kramer, A., Wickens, C., Reisweber, M., Strayer, D., Grenell, J., 1993. Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics* 36, 1121–1140.

[23] Meyer, D. E., &Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory period phenomena. *Psychological Review*, 104, 749–791.

[24] Navon, D., & Miller, J. (2002). Queuing or sharing? A critical evaluation of the single-bottleneck notion. *Cognitive Psychology*, 44, 193-251

[25] Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer -based tests. *Applied Measurement in Education*, 16, 163-183.

[26] Fan, X., Miller, B. C., Park, K., Winward, B. W., Christensen, M., Grotevant, H. D., et al. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, 18, 223–244.

[27] Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press

[28] Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer based assessment. *Comput. Educ*., 56(4), 1032–1044.

[29] Bloom, B. S., Taxonomy of educational objectives: The classification of educational goals. New York: Longmans, Green, 1956.

[30] Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8(2).

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

[32] Wold, H. (1982). *Soft Modeling: The basic design and some extensions*. In Karl G. Jöreskog& Herman Wold (Eds.), iSystems under indirect observation: Causality, structure prediction II (1–54). Amsterdam, Netherlands: North Holland.

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

[34] Wixon, B. H., & Watson, H. J. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly*, 25(1), 17–41.

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

[36] Cortina, J. M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology,* 78(1), 98-107.

[37] Cohen, J. (1988). *Statistical power analysis for the behavioural sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.

[38] Wetzels, M., Odekerken-Schröder, G., & Van Oppen, C. (2009). Using PLS path modelling for assessing hierarchical construct models: Guidelines and empirical illustration. *MIS Quarterly*, 33(1), pp. 177–195.