A hierarchical probabilistic framework for recognizing learners' interaction experience trends and emotions

Imène Jraidi, Maher Chaouachi, Claude Frasson

Université de Montréal, Dept. of Computer Science and Operations Research

2920 chemin de la tour, H3T-1J8 QC, Canada

{jraidiim, chaouacm, frasson}@iro.umontreal.ca

Abstract. In this paper we seek to model the users' experience while interacting with a computer-based learning environment. More precisely, we are interested in assessing the relationship between learners' emotional reactions and three extreme trends in the interaction experience, namely *flow*: the optimal interaction (a perfect immersion within the task), *stuck*: the non-optimal interaction (a difficulty to maintain focused attention), and *off-task*: the non-interaction (a drop out from the task). We propose a hierarchical probabilistic framework using a dynamic Bayesian network to model this relationship, and to simultaneously recognize the probability of experiencing each trend, as well as the emotional responses occurring subsequently. The framework combines three-modality *diagnostic variables* that sense the learner's experience including physiology, behavior and performance, *predictive variables* that represent the current context and the learner's profile, and a *dynamic structure* that tracks the temporal evolution of the learner's experience. An experimental study, with a specifically designed protocol for eliciting the targeted experiences, was conducted to validate our approach. 44 participants interacted with three computer-based learning environments involving different cognitive tasks (problem solving, memorization and reasoning), while their physiological activities (electroencephalography, skin conductance and blood volume pulse), patterns of the interaction, and performance during the task were recorded. Results revealed that multiple concurrent emotions can be associated to the experiences of flow, stuck and off-task, and that the same trend can be expressed differently from one individual to another. The evaluation of the proposed framework showed promising results in predicting learners' experience trends and emotional responses.

Keywords. Interaction experience, flow, stuck, off-task, emotional responses, biofeedback sensing, dynamic Bayesian networks

1 Introduction

Modeling and understanding the users' interaction experience is an important challenge in the design and development of adaptive intelligent systems [1]. Ongoing advances in human-computer interaction (HCI), cognitive science, psychology and neuroscience have greatly enhanced the ability of such systems to effectively diagnose users' behaviors and to provide appropriate assistance and adjustment [2-7]. In this context, a particular attention is paid to modeling users' affect and emotional reactions, as they play a critical role in users' cognitive performance, and decisively influence their

perception, concentration, decision-making, memorization and problem solving abilities [8-11]. In the field of computer-based learning and intelligent tutoring systems (ITS), a growing interest has been devoted to obtain and monitor information about learners' emotions. The combination of multimodal affect sensing technologies with artificial intelligence (AI) techniques proved its effectiveness in inferring learners' emotional states [12-18]. Physiological monitoring using wearable non-invasive biofeedback devices holds a prominent place as they provide valuable quantitative and objective information as compared to traditional evaluation methods such as questionnaires or self-report [19-21].

Nevertheless, the integration of the affective dimension within ITS has raised much debate about which emotions should be assessed. No clear consensus was reached on which emotions should be fostered or avoided within tutoring interactions [22-24]. Indeed, the relationship between emotions and learning is far more complex than a linear association that would state that positive emotions enhance learning, while negative emotions obstruct it [25]. Some emotions considered a priori as negative, are not only inevitable within technology-mediating learning [26, 27], but can also contribute positively to the learning experience. For example, stress can have two opposite effects: the 'positive' stress (or eustress) is known to stimulate cognitive abilities, while the 'negative' stress (or distress) penalizes concentration and decreases cognitive performance [28, 29]. Similarly confusion can represent a positive challenging aspect in the learning experience, or might conversely, signal a cognitive lock or impasse [23, 30, 31]. Therefore, the assessment of learners' emotions may not provide in itself, an explicit evaluation of their interaction experience. For instance, beyond which level, stress becomes harmful to the learning experience? This is obviously a challenging aspect, given the highly contextualized, person-dependent and dynamic nature of emotions.

Hence, the goal of this research is to not only assess learners' emotional responses, but also to determine how emotions impact their learning experience, whilst taking account of both contextual and individual differences, and tracking the dynamics of the learners' states over time. More precisely, we propose to model the relationship between learners' emotions and the tendency that characterizes the quality of their interaction experience (e.g. positive/favorable or negative/unfavorable). We identify three extreme trends in the interaction, namely *flow* or the optimal experience: a state in which the learner is completely focused and involved within the task, *stuck* or the unfavorable interaction: a state in which the learner has trouble to maintain focused attention, and *off-task* or the non-interaction: a state in which the learner is not involved anymore within the task. The hypothesis we establish is that these trends can be associated to multiple overlapping emotional responses, and that this relationship can be specific to each learner. We propose a hierarchical probabilistic framework using a dynamic Bayesian network to model this relationship, and to simultaneously recognize the trend that characterizes the learner's interaction experience, and the emotional responses occurring subsequently. The framework involves three different modalities to diagnose the interaction including physiology, behavior and performance, the learner's profile and context-dependent variables to account for individual differences and environmental factors, and a dynamic structure to track the evolution of the interaction experience over time.

An experimental study was conducted to test our hypothesis and validate our approach. A protocol was established to manipulate the learners' interaction experience, and elicit the three targeted trends as they used three computer-based learning environments involving different cognitive tasks namely: problem solving, memorization and reasoning. 44 participants were recruited for this experiment while monitoring their physiological activities using three biofeedback devices (electroencephalogram, skin conductance and blood volume pulse), behavioral variables tracking patterns of their interactions, and performance during the tasks. The evaluation of the proposed framework shows its capability to efficiently recognize the learners' experience. We demonstrate that our approach outperforms conventional non-dynamic modeling methods using static Bayesian networks, as well as three non-hierarchical formalisms including naive Bayes classifiers, decision trees and support vector machines.

The remainder of the paper is organized as follows. A brief literature review is outlined in Section 2. Section 3 describes the proposed hierarchical framework for assessing learners' interaction trends and emotional responses. Section 4 details our experimental setup and methodology. Finally, Section 5 discusses the experimental results, and Section 6 concludes and presents directions for future work.

2 Related work

Improving the interaction between users and computers requires both a means of measuring qualitatively the users' experience, as well as a set of adaptive mechanisms to automatically adjust the interaction. A large body of work has extensively been devoted to evaluate the users' experience by analyzing their emotions as they play a key role in mirroring the users' internal state. Approaches for measuring emotions - especially in the fields of HCI and ITS - are typically concerned with the recognition of a single emotional state. Two distinct strategies are mainly adopted: either a specific emotion is considered in isolation, or several emotions are considered, but treated as mutually exclusive. For the first case, the system is designed to identify a specific class of emotion such as frustration [17, 32, 33], stress [34-36], confusion [15, 37, 38] or fatigue [39-41]. For the second case, the system is capable of representing and recognizing several classes of emotions that vary over time, but at a given time the user is characterized by a unique emotional state (e.g. [12, 14, 18, 22, 42]). These approaches clearly restrict the evaluation of the user's experience as they provide only a limited insight into the user's actual state. Indeed, several emotions can be experienced at the same time; these emotions can have either the same or opposed valence [7, 43]. For instance at a given time, a user can be both interested and engaged within the current task, but also stressed and confused. Hence, representing and recognizing a combination of overlapping states provides a more holistic and comprehensive view of the user's experience [13].

Current approaches on affective modeling can be also categorized according to the machine learning techniques used to recognize the users' emotional states. The first category uses conventional classification algorithms including rule-based reasoning [44], support vector machines [42, 45], neural networks [46, 47], decision trees [48, 49], etc. These approaches rely mostly on a low-level mapping between manifesting features of affect and the targeted emotional states. This mapping is often inadequate to represent complex dependencies comprising contextual features or person-related characteristics, which could interfere in the experience of affect. Besides, the classification of the user's state is commonly achieved on an ad-hoc and static basis, independently of the history; that is without taking account of the past knowledge regarding the user state. Another limitation of these approaches is that they are often unable to represent and manage the uncertainty associated to both the sensory measurements and the expression of affect. To overcome these limitations the second category of approaches use hierarchical probabilistic methods such as dynamic Bayesian networks (DBN), hidden Markov models (HMM), etc. DBN are particularly used for affect recognition (e.g. [13, 41, 50, 51]) as they provide a powerful tool to model complex causal relationships at different levels of abstraction, and capture the dynamics and the temporal evolution of the user's state, while efficiently handling the uncertainty through probabilistic representation and reasoning formalisms. For instance Conati et al. [13] use a DBN to monitor learners' emotions within an educational game, using bodily expression-related features, personality traits and patterns of the interaction. Liao et al. [51] infer users' stress levels using a DBN that combines physiological measures, physical observable changes, and performance and interaction features. Ji et al. [41] use observable clues including facial expressions, gaze direction, head and eye movement, in conjunction with context-related information to assess human fatigue.

In this paper, we propose a hierarchical probabilistic framework to dynamically track the users' experience while interacting with a learning environment. Our approach differs fundamentally from previous work in that we are not only recognizing concurrent emotional states, but also measuring explicitly the tendency that characterizes the quality of their interaction experience. More precisely, our objective is to assess the relationship between emotions and the type of the interaction. That is, how emotions impact the learning experience? Or in other words, how a favorable (or an unfavorable) interaction is manifested emotionally? We propose to evaluate the learners' experience with regards to three extreme key trends, namely the states of flow, stuck and off-task, which characterize the learners' interaction along the dimensions of involvement and control (or mastery) regarding the task at hand, and which would determine whether a tutoring intervention is required. Flow is the optimal trend: a positive experience where the learner is perfectly focused and involved within the task. A feeling of being in control prevails, as an equilibrium is found between the challenge at hand and the learner's skills [52]. It is hence the moment where a tutoring intervention should be avoided to not interrupt the learner, and risk to disturb his cognitive flow. Stuck is a non-optimal trend: a negative

experience where the learner has trouble to maintain focused attention. The learner feels to be out of control, as a pronounced disequilibrium is perceived between the challenge at hand and his skills [53]. In this case, a supportive intervention should be performed to help the learner overcome the encountered difficulty, and peruse the task. The off-task trend (or the 'non-interaction') can be seen as an extremely negative experience where the learner totally loses his focus, and drops out from the task. The notion of control is no longer applicable, as the learner gave up, and 'disconnected' from the interaction. The off-task trend should therefore be carefully monitored; and if detected, a more radical intervention would be performed to motivate the learner and get him involved again in the interaction, such as changing the current task or presenting a different topic.

Although there have been significant attempts to model these trends, especially within ITS [16, 53-56] and video game environments [57-59], there is still a lack of a unifying framework to systematically assess, in a dynamic way both the three types of interaction (i.e. flow, stuck and off-task), and the emotional responses that occur subsequently. Indeed these states have mainly been approached in an isolated manner, and mostly associated to a single emotion within constrained interactions, such as predicting whether a learner is about to quit from a Towers of Hanoi activity as he presses a button labeled 'I'm frustrated' while resolving the task [16], or by detecting whether the user is avoiding to learn the materials by guessing or abusing hint features [54].

To summarize, the research presented in this paper extends prior work in the following ways. First, we combine the recognition of the three interaction experience trends with the emotional responses. We assume that a learner's experience can be possibly associated to several overlapping emotions, and that the same trend can be expressed differently from a learner to another. Second, we propose a hierarchical probabilistic framework based on DBN to model and train the relationship between learners' emotions and the targeted trends. The framework combines multimodal channels of affect, with the learner's profile and context-dependent variables, to automatically recognize the probability of experiencing flow, stuck and off-task, and to assess the emotional responses occurring during the interaction. Finally, we validate our approach through an experimental study where we provoke the three interaction trends as learners are performing different cognitive tasks.

3 The proposed approach

In this section we describe our framework for modeling a user's experience while interacting with a computer-based learning environment. The framework uses a dynamic Bayesian network [60] to automatically track the learner's emotional changes, where concurrent emotions are represented, and assess the probability of experiencing flow, stuck and off-task. A macro-model of the framework is given in figure 1; it includes two main portions to represent the factors (causes) and the manifesting features (effects) of a learner's state namely, a predictive component and a diagnostic component.

Predictive component. The predictive (upper) part of the network describes the factors that could cause or alter the experience of the interaction. These factors represent the current context, which includes environmental variables that can directly influence the learner's experience such as the level of difficulty of the task at hand, the relevance of the hints or help provided, the imposed time constraints, etc. The predictive portion includes also the learner's own characteristics (profile) that can directly or indirectly influence the learning experience. These include the learner's goal, preference, personality, skills, computer usage frequency, etc.

Diagnostic component. The diagnostic (lower) part denotes the evidence, i.e. the sensory observations used to infer the learner's state. Three-modality channels can be included, namely physiology, behavior and performance. (1) Physiological features can be used to track bodily changes associated to emotions. For instance, galvanic skin response (GSR) is widely known to linearly vary with the emotional arousal [61, 62]. Heart rate (HR) is extensively applied to understand the autonomic nervous system function and has shown a close correlation to the emotional valence [61, 63, 64]. Electroencephalogram (EEG) can provide neural indexes related to cognitive changes such as alertness, attention, workload, executive function, or verbal and spatial memory [2, 65-67]. Particularly, Pope and colleagues developed at NASA [68] a mental engagement index. This index

showed a great reliability in switching between manual and automated piloting modes and was used as an alertness criterion for adaptive and automated task allocation [69]. It was also used within an educational context, providing an efficient assessment of learners' mental vigilance and cognitive attention [70].

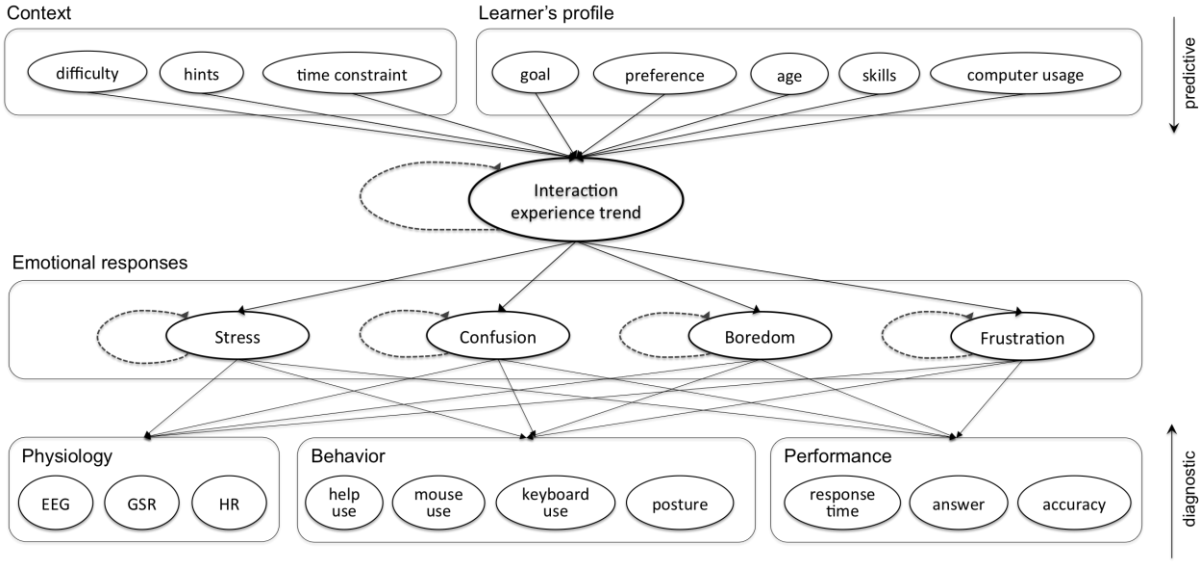

Fig. 1 The proposed framework for assessing learners' emotions and interaction experience trends using a DBN. The rectangular spaces are generic; other variables can be inserted or replaced given the available modalities. For instance, the sensory nodes can be adapted to the current environment and devices (e.g. a video camera, a posture sensitive chair, an eye-tracker, etc.). Dashed arrows denote temporal dependencies, e.g. the interaction trend at time t is affected by the experienced trend at time t-1. Note that we don't draw all the links between the emotion nodes and each diagnostic variable due to space limit.

(2) Behavioral features comprise key aspects of the interaction between the learner and the environment, which may give clues about the learners' levels of involvement (or activity/inactivity) within the task. These variables include the rate of requesting help, the hints used, mouse or keyboard pressing, click frequency, character input speed, etc. Additional devices can be used to assess learners' behaviors during the interaction such as a video camera, an eye-tracker, a posture sensitive chair, etc. (3) Performance features involve objective measures that can be influenced by changes in the learner's experience, and could provide an indication about the level of mastery of the task. These features include correctness, errors made in the task, time spend before answering, etc. More complex features can be used to track the learner's skill acquisition process such as the content that the learner knows, the practiced skills, etc.

The middle part of the model represents the learner's actual state. The first layer represents the concurrent emotional responses. Each emotion is represented by a separate random variable with different possible outcomes. In this work, we are modeling four classes of emotions pertaining to learning, and frequently observed during computer tutoring, namely stress, confusion, boredom and frustration [22, 23, 25-27, 71]. For instance, the node associated to stress can have the following outcomes: calm (no stress), low, moderate and high stress. Similarly confusion can range from confidence (no confusion) to high confusion, boredom can range from interest (no boredom) to high boredom, and frustration: from satisfaction (no frustration) to high frustration. The second layer represents the learner's interaction experience trend with the following possible outcomes: flow, stuck and off-task. The recognition is achieved through a probabilistic inference from the available diagnostic measures (bottom-up) to update the learner's emotional responses (i.e. the probability of each emotion node's outcome). This inference will be in turn, combined to a predictive (top-down) inference from the current context and personal variables, and propagate to update the learner's interaction trend (i.e. the probability of experiencing flow, stuck and off-task).

This two-layered abstraction is aimed to quantify the learner's experience trend on one hand, and to identify the emotional responses that occur subsequently on the other hand. More precisely the goal

is to determine the emotions that occur when the probability of a positive interaction (flow) tends to decrease and the probability of a negative interaction (stuck and off-task) tends to increase so that an effective intervention can be initiated, and targeted according to the predominant emotional states. In addition, the model includes a dynamic structure representing the temporal evolution of the learner's interaction trends and emotional responses. This structure is described by the dashed arcs shown in figure 1. Each random node at time t is influenced by the observable variables at time t, as well as by its corresponding random node's outcomes at time t-1. The resulting network is made up of interconnected time slices of static Bayesian networks describing each, a particular state of the learner. The relationship between two neighboring time slices is represented by a hidden Markov model (HMM). That is the inference made at time t-1 is used in conjunction with the sensory data observed at time t, to update the learner's current emotions and the probability of each trend.

4 Methodology and experimental design

An experimental protocol was established to deliberately manipulate the learners' interaction experience, while recording their physiological activities, behavioral patterns, and performance during the tasks. Data were collected from 44 participants of different ages, gender and qualifications to validate our approach. Three devices were used to record participant's physiological activities, namely electroencephalogram (EEG), skin conductance (SC), and blood volume pulse (BVP) sensors. EEG was recorded using a 6-channel headset. SC and BVP sensors were placed in the resting left hand fingers. Data were synchronized using necessary time markers, to automatically integrate the recorded signals with the rest of the instrumental setup. In addition, two video cameras were used to record the users' face, and the onscreen activity, so that to not miss any feature of their interactions.

Three environments were used for our experimentations, namely trigonometry, backward digit span (BDS), and logic. The goal was to study the learners' experience within different contexts and cognitive tasks. BDS and logic involve strict cognitive tasks with controlled laboratory conditions, namely memorizing digits, and logical exercises. The trigonometry session is a more complex learning environment, with less controlled conditions. It comprises a learning session with an introductory course covering some basic trigonometric properties and relationships, followed by a problem solving activity. Figure 2 depicts a screen shot from each environment.

One of the key points of this study was to acquire accurate data related to the learners' interaction experience trends and emotional responses. Thus the three environments were thoroughly designed in a way that would intentionally elicit the three types of interaction (i.e. flow, stuck and off-task). Each session begins with relatively simple tasks; everything was made to get the learners involved within the activity (e.g. easy problems, figures clarifying the problem statements, help/hints provided if needed, no time limit imposed, etc.). As the learner progresses within the session, the tasks become more challenging and the level of difficulty increases gradually. Different parameters were manipulated to deliberately vary the difficulty level and foster the states of stuck and off-task. These included the complexity of the task to be performed, the time limits, and the provided help. Some additional parameters (e.g. unreasonable time limits, deliberate bugs, etc.) were adjusted to systematically get the learners puzzled or even discouraged from pursuing the activity.

Trigonometry. For this session, we used the trigonometry tutoring system developed by Chaouachi et al. [72]. The tutoring content, which formally covered six basic problem solving tasks, was enhanced with additional tasks (16 in total) structured in three series of incrementally increasing difficulty as will be described below. The session started with a trigonometry lesson explaining several fundamental trigonometric properties and relationships. Basic definitions as well as their mathematical demonstrations were given. The environment provided schemas and examples for each presented concept, and a calculator to perform the needed computations. Learners were then asked to complete a problem solving activity, which involved applying, generalizing and reasoning about the trigonometric properties. No further prerequisites were required to resolve the problems, except the concepts previously seen. However a good level of concentration was needed to successfully achieve the tasks. Three series of gradual difficulty were designed for this activity; several parameters were considered namely: the time constraints, the presence/absence of help, and the complexity of the task.

Particularly, each trigonometric problem required some intermediate steps to reach the solution and the complexity was enhanced by increasing the number of the required steps.

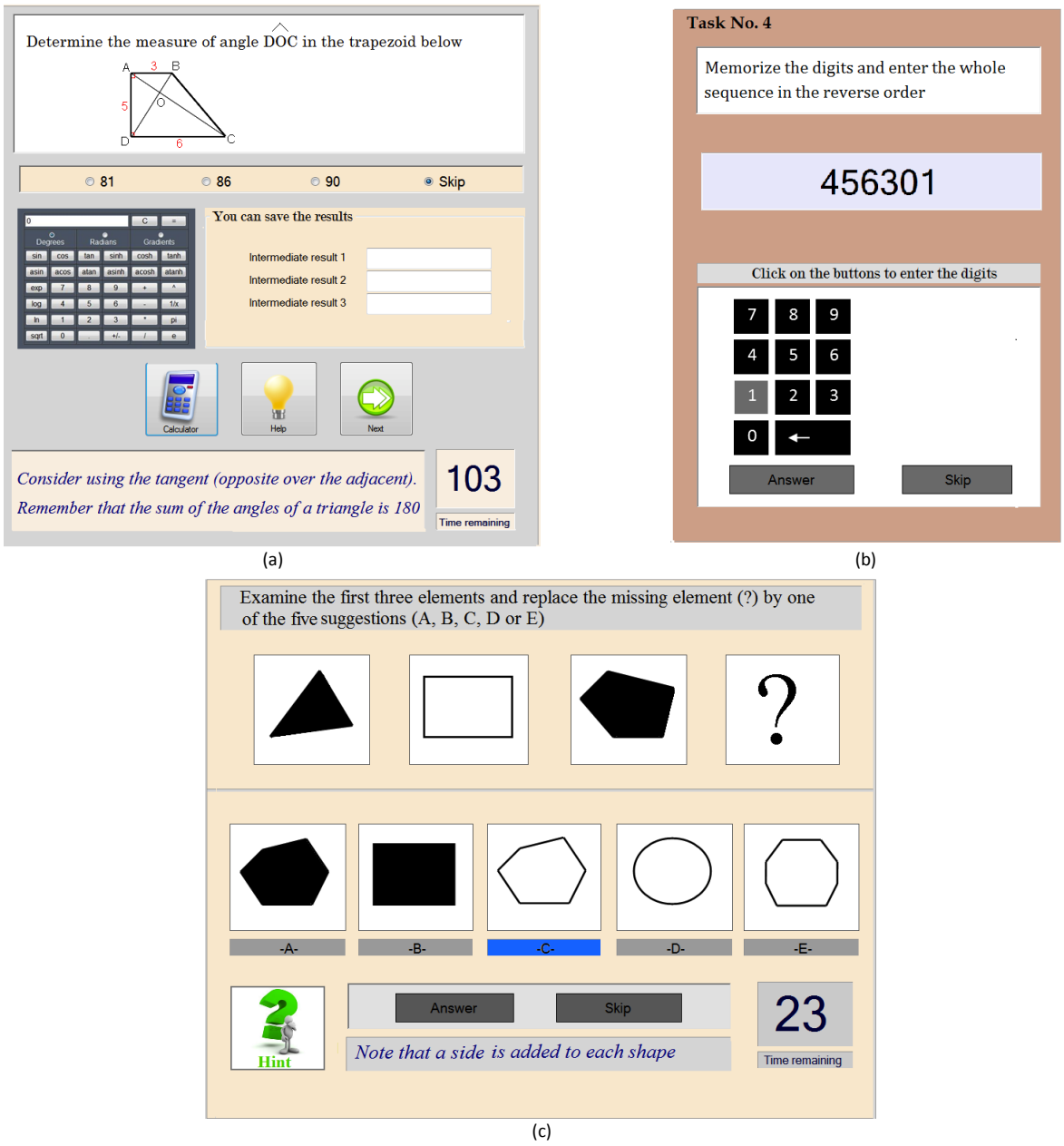

Fig. 2 Screen shots from the three environments: (a) trigonometry (b) backward digit span (BDS) and (c) logic.

Series 1 involved six rudimentary multiple-choice questions, without any time limit. The problems consisted mainly in applying simple trigonometric properties, and required few intermediate steps (e.g. calculating the measure of an acute angle within a right triangle given the length of the hypotenuse and the opposite side). The environment provided a limited number of hints for each problem. The hints (if used) provided relevant and detailed information leading to the solution (e.g. "Remember to use the sine = hypotenuse / opposite"). Schemas illustrating the problems and the necessary recalls were presented as well, to make the task easier. Series 2 consisted of five multiple-choice questions. The problems of this series were more complex and required an increased number of intermediate steps to reach the solution. For example, to compute the sine of an angle, learners had first to compute the cosine. Then, they had to square the result, and to subtract it from 1. Finally they had to compute the square root. A geometrical figure was given to illustrate the statements, and reasonable time limits, varying according to the difficulty, were fixed for each problem. Some hints

were given for the most difficult problems. However the information provided was very vague as compared to series 1 (e.g. "The sum of the angles of a triangle is equal to 180 degrees"). Series 3 involved five open response questions (i.e. without offering potential options to choose from). The problems involved more elaborated statements, and a further concentration was needed to translate the statements into a trigonometric formulation (e.g. "A 50-foot pole (height = 50 feet), perpendicular to the ground, casts a shadow of 20 feet (length = 20 feet) at a given time. Find the elevation angle (in degrees) of the sun at that moment"). Very strict gradually decreasing time limits were imposed and no hints or illustrations were given for this series.

Backward digit span (BDS). This activity involves mainly working memory and attention abilities. A series of single digits are presented successively on the screen during a short time. Learners are asked to memorize the whole sequence, and then instructed to enter the digits in the inverted order of presentation. Two levels were considered for this session, namely BDS 1 and BDS 2. Each level involved three tasks (i.e. task one to three and task four to six), of gradual difficulty by increasing the number of digits of the displayed sequence. The difficulty was further enhanced in BDS 2 by gradually decreasing the digits' display periods (from 700 ms to 600, 500 and 300 ms). Task one consisted of a series of 12 sets of 3 digits, task two: 8 sets of 4 digits, task three: 7 sets of 5 digits, task four: 5 sets of 6 digits, task five: 4 sets of 8 digits, and task six: 4 sets of 9 digits. Participants were instructed to use the mouse to enter the digits using a virtual keyboard displayed on the screen. No additional time constraints were imposed for this activity.

Logic. This activity involves inferential skills on information series and is typically used in brain training exercises or tests of reasoning. No further prerequisites are needed but a high level of concentration is required. The goal is to teach learners how to infer a logical rule from a series of data in order to find a missing element. The tutoring environment is composed of three modules. Each module is concerned with specific forms of data: the first module deals with geometrical shapes (Geo.), the second module with numbers (Num.), and the third module with letters (Lett.). The session started with a tutorial giving instructions and warm up examples, to get the learners accustomed with the user interface and types of questions, then a series of multiple-choice question tasks related to each module is given. For instance in the Geo. module, three shapes were successively presented in the interface. The first shape represented a black triangle, the second a white rectangle, and the third a black pentagon. Learners were asked to deduce the fourth missing element, which would be in this case, a white hexagon. That is, the logical rule that one should guess is to alternate between the black and white colors and to add a side in each shape. Two levels with an increasing difficulty were considered for each module namely Geo. 1 and 2, Num. 1 and 2, and Lett. 1 and 2. The difficulty was manipulated by enhancing the complexity of the logical rule between the data. In addition, for the first level, the environment provided a limited number of hints to help the learners find the logical rule that they had to infer, and no time constraint was fixed to answer. In the second level, the hints were increasingly scarcer or even omitted, and a gradually decreasing time delay was imposed to answer. Besides, some tasks were designed to systematically mislead the learner. For instance in the Num. 2 module, two perpendicular data series were presented. In the vertical series all the numbers were multiples of seven and in the horizontal series all the numbers were multiples of five. In this task, one should deduce the missing crossing element, which should be a multiple of both five and seven. But no such element was given among the possible answers. Some disturbing bugs were also intentionally provoked to get learners distracted and lose their focus (e.g. freezes, hidden statements or materials, very unreasonable time limits, etc.). A total of 20 tasks were given in this session: each sub activity consisted of 3 tasks, except Num. 2 and Lett. 2, which involved 4 tasks each.

Sensory measurements

Three-modality measures were monitored, namely behavioral variables, performance, and physiological features. Behavioral variables included the mouse movement rate (Mouse_mvt) and the frequency of requesting help/hints (Help_req). Performance measures included response time (Resp_time), answer to the current task (correct, incorrect or no-answer) and the overall accuracy rate.

Physiological features involved galvanic skin response (GSR), heart rate (HR) and mental engagement (EEG_Engag). We discuss below the methodology used to extract and pre-process the physiological data.

Physiological features. Three devices were used to record learners' physiological activities, namely skin conductance (SC), blood volume pulse (BVP), and electroencephalogram (EEG) sensors. The acquired signals were digitized using the ProComp Infinity multi-channel data acquisition system [73]. The SC device computed galvanic skin response (GSR). It measures changes in the resistance of the skin produced by the perspiration gland activity. A tiny voltage is applied through two electrodes strapped to the first and middle fingers on the palm side. This establishes an electric circuit that quantifies the skin's ability to conduct electricity, which increases as the skin is sweaty (for instance when one is experiencing stress). The SC data were recorded at a sampling rate of 1024 Hz. The BVP device is a photoplethysmograph sensor, which computes the amount of light reflected by the surface of the skin. This amount varies with the quantity of blood present in the skin, and thus with each heartbeat. The BVP signals were recorded at a sampling rate of 1024 Hz. Heart rate (HR) was calculated by measuring the inverse of the inter-beat intervals (i.e. distance between successive pulse peaks).

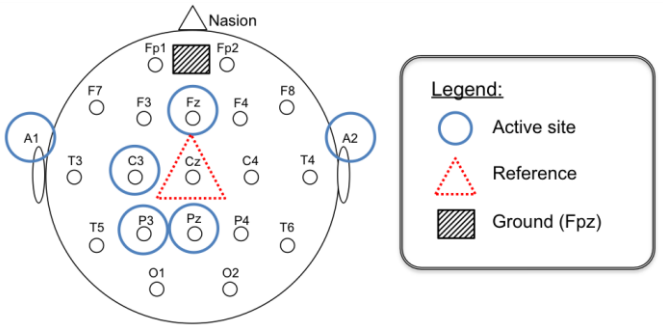

Fig. 3 EEG channel electrode placement

EEG was recorded using an electro-cap that measures the electrical brain activity produced by the synaptic excitations of neurons. Signals were received from sites P3, C3, Pz, and Fz as defined by the international 10-20 electrode placement system [74]. Each site was referenced to Cz and grounded at Fpz. Two more active sites were used namely A1 and A2 (i.e. the left and right earlobes respectively). This setup is known as the "referential linked ear montage", and is illustrated in figure 3. In this montage, roughly speaking, the EEG signal is equally amplified throughout both hemispheres. Moreover, the "linked-ear" setup calibrates each scalp signal to the average of the left and right earlobe sites, which yields a cleaner and a more precise signal. For example, the calibrated C3 signal is given by (C3 - (A1 + A2) / 2). Each scalp site was filled with a non-sticky proprietary gel from Electro-Cap and impedance was maintained below 5 Kilo Ohms. Any impedance problems were corrected by rotating a blunted needle gently inside the electrode until an adequate signal was obtained. The recorded sampling rate was at 256 Hz.

Due to its weakness (at the order of a few microvolts), the EEG signal needs to be amplified and filtered. Besides, the brain electrical signal is usually contaminated by external noise such as environmental interferences caused by surrounding devices. Such artifacts alter clearly the quality of the signal. Thus a 60-Hz notch filter was applied during data acquisition to remove these artifacts. In addition, the acquired EEG signal easily suffers from noise caused by user body movements or frequent eye blinks. Thus a 48-Hz high pass and 1-Hz low pass de-noising filters were applied. The engagement index was derived using three EEG frequency bands, namely Theta (4-8 Hz), Alpha (8-13 Hz) and Beta (13-22 Hz). A fast Fourier transform (FFT) was applied to transform the EEG signal from each active site into a power spectrum. The transformed signal was divided to extract the estimated power with respect to each band. A combined power was then summed from the measured scalp sites in order to compute the EEG band ratio given by: Beta / (Alpha + Theta) [68]. The EEG engagement index was then smoothed using a sliding moving average window: at each instant T, the

engagement index is computed by averaging each ratio within a 40s-sliding window preceding T. This procedure is repeated every 2s and a new 40s-sliding window is used to update the index.

Participants and protocol

44 participants (31 males) aged between 19 and 52 (M = 28.61, ± 8.40) were recruited for this research. Participation was compensated with 20 dollars. Upon arrival at the laboratory, participants were briefed about the experimental objectives and procedure and asked to sign a consent form. They were then outfitted with the biofeedback devices and familiarized with the materials and environments. Next, participants filled in demographic information (age, gender, qualification, frequency of computer usage per day, etc.). They were also asked about their preferences regarding the three activities (i.e. whether they like or not trigonometry, digit recall and logical reasoning, respectively), and their perceived skill levels (low, moderate or high) in each of the three activities. Then, the Big Five Inventory (BFI) was administrated to assess learners' personality traits, namely openness, conscientiousness, extraversion, agreeableness, and neuroticism [75]. After that, participants completed a 5-minute eyes open baseline followed by another 5-minute eyes closed baseline to establish a neutral reference for the physiological variables.

Participants were then instructed to complete the trigonometry session, followed by BDS, and logic. To make the tasks more stimulating, participants were informed that a correct answer is rewarded 4 points, -1 point is given for a bad answer, 0 point is given for a no-answer, and that they could, if they choose to, get their score and ranking as compared to other participants, at the end of the three sessions. All participants completed the levels of the three activities in the same order, namely: series 1 to 3 for trigonometry, next BDS 1-2, and then Geo. 1-2, Num. 1-2 and Lett. 1-2 for logic. They were allowed to self-pace with respect to the time required to complete each task and were given breaks and rest periods between the three sessions and levels. Before starting each level, participants were asked what were their goals regarding the next tasks by choosing between the following: "realizing the highest score/fewest incorrect answers possible", "learning or discovering new concepts", or just "finishing the task". The experiment ended with a debriefing interview.

Subjective measurement collection. After completing each task, participants reported how they have been experiencing the last trial. Participants were instructed to select the trend that would characterize their overall state during the last task (i.e. flow, stuck or off-task), and rate their experienced levels of stress, confusion, frustration and boredom. A definition of each trend was given to the participants, to help them choosing the descriptions that best match to their experiences. Flow was defined as: "I felt like I was immersed in the activity. I was totally involved, and I was focused and attentive. I was totally controlling the task, and I felt that I had the necessary skills to fulfill it". Stuck was defined as: "I felt that I was blocked. I had trouble to maintain focused attention. I was not totally controlling the task, and felt like I could not make it". Off-task was defined as: "I was likely to drop out. I could not (or did not want to) concentrate and I was no more involved in the task. I felt like I gave up, or that I did not want to pursue".[1]

A definition of each of the four emotions was provided as well. Stress was defined as: a reaction from a state of calm (relaxed) to an excited state, a feeling of tension or worry due to environmental pressure or constraint. Confusion was defined as: having doubts or uncertainty; may be due to a lack of knowledge or understanding. Frustration was defined as annoyance, irritation or dissatisfaction. Boredom was defined as being wearied or listless due to a lack of interest. Four graduated scroll bars ranging from 0 to 100 were used to rate the intensity of each emotion. The bars included the following subdivisions 0 = no negative emotion (i.e. calm, confident, satisfied or interested, respectively for stress, confusion, frustration or boredom),]0; 35] = a low level,]35; 65] = moderate, and]65; 100] = high. For instance, if a participant rated 17 for stress, 52 for confusion, 0 for frustration and 0 for boredom, we get the following overlapping states: low stress, moderate confusion, satisfied and interested.

[1] If participants reported an 'off-task' trend, they were given a little break before resuming the session.

5 Results and discussion

A total of 1848 samples (42 * 44 participants), were collected from the experiment. Results are organized as follows: first we describe the statistical analysis conducted to validate our experimental design. Then, we study the relationship between the reported emotions and the experienced trends. Finally, we evaluate our framework for recognizing learners' interaction experience trends and emotional responses.

Analysis of the reported experiences

A preliminary statistical analysis was performed to analyze the experienced trends with regards to the task design. More precisely, the goal was to investigate how participants perceived their interactions throughout the sessions: What was the distribution of the targeted trends (i.e. flow, stuck and off-task) across the different activities? Did the reported experiences vary in line with the established experimental process?

A two-way repeated measure ANOVA was conducted to evaluate the incidence (occurrence) and the variation (increase or decrease) of flow, stuck and off-task [2] across the levels of difficulty of the three sessions (i.e. trigonometry, BDS and logic). The within-subject dependent variable was the proportions of the interaction trends, and the independent variables were: (i) the type of the trend (flow, stuck or off-task) and (ii) the testing time (i.e. series 1-3 for trigonometry, levels 1-2 for BDS, and Geo. 1-2, Num. 1-2 and Lett. 1-2 for logic). Results revealed a statistically significant main effect of the trend: $F(1.80, 77.71) = 61.85$, $p < 0.001$; degrees of freedom were corrected using Huynh-Feldt estimates of sphericity (epsilon = 0.89), as the assumption of sphericity has been violated (chi-square = 6.79, $p < 0.05$). Post-hoc tests with a Bonferroni adjustment indicated that the state of flow was in overall (i.e. across the three environments), the most prominent trend (M = 0.59 (0.028)), the state of stuck was less frequent (M = 0.27 (0.022)), and off-task was the least prevalent state (M = 0.14 (0.021)).

The interaction effect (trend * testing time) showed that the rates of occurrence of flow, stuck and off-task differed significantly across the 11 sub-activities: $F(10.24, 440.40 = 20.07)$, $p < 0.001$; degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (epsilon = 0.59), as the assumption of sphericity has been violated (chi-square = 441.62, $p < 0.001$). Bonferroni corrected posthoc tests yielded the following patterns at the 0.05 significance level: flow > stuck > off-task for the beginning of the trigonometry session (series 1) and flow > (stuck = off-task) for the beginning of BDS (level 1) and logic (Geo. 1). Hence in the first tasks of each of the three environments, experiences of flow were the most common, and experiences of stuck were either on par with or higher than off-task. For instance an average occurrence of 79% for flow, 17% for stuck and 4% for off-task was found in series 1 of trigonometry. These patterns were reversed towards the end of each activity: (flow = off-task) < stuck for trigonometry (series 3), and flow = off-task = stuck for BDS (level 2) and logic (Lett. 2). That is experiences of stuck were either as or more likely than experiences of off-task and flow. For instance for the last sub-activity in logic (Lett. 2), flow occupied 38% of the interaction time, stuck 30% and off-task 32% of the time.

Figure 4 shows the estimated marginal means of each trend over the 11 sub-activities of the three learning environments. The proportions of flow were significantly lower at the end of the trigonometry session (series 3) compared to the beginning of the session (series 1) ($d = -0.497$, $p < 0.001$), and the proportions of stuck and off-task were significantly higher ($d = 0.239$ and 0.258 respectively, $p < 0.05$). The same pattern was observed from level 1 to level 2 in BDS ($d = -0.523$ for flow, 0.258 for stuck, and 0.265 for off-task, $p < 0.001$), as well as from Geo. 1 to Lett. 2 in logic ($d = -0.534$ for flow, 0.218 for stuck, and 0.316 for off-task, $p < 0.05$). Hence for each of the three environments, the further the learners got within the tasks, the more the occurrences of flow decreased and the occurrences of suck and off-task increased. This pattern was inverted between the end of the trigonometry session (series 3) and the beginning of BDS (level 1): the incidence of flow increased

[2] During debriefing, participants were asked whether there were other trends that would characterize their overall state throughout their interactions. All responses were negative.

(d = 0.611, p < 0.001), and the incidence of stuck and off-task decreased (d = -0.338 and -0.273 respectively, p < 0.001). The same variations were observed from BDS (level 2) to logic (Geo. 1): d = 0.530 for flow, -0.250 for stuck and -0.280 for off-task (p < 0.05). Besides, within the logic session, the occurrences of flow increased from Geo. 2 to Num. 1 (d = 0.220, p <0.05), and from Num. 2 to Lett. 1 (d = 0.237, p < 0.05), that is as a different type of materials (numbers or letters) was presented, with a lower difficulty level. The difference was not significant for stuck and off-task.

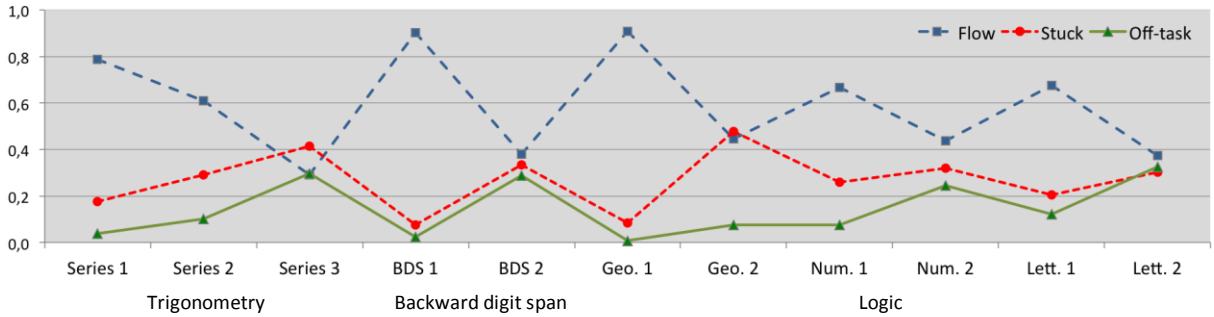

Fig. 4 Estimated marginal means of the proportions of the experienced trends over the three sessions.

To sum up, the experienced trends accurately tracked the intended experimental design. At the beginning of their interactions, learners were more likely to experience flow. The occurrences of negative interactions (stuck and off-task) were also probable, but with very low proportions. As the level of difficulty of the task increased (i.e. more complex tasks, imposed time constraints, scarcer hints, provoked bugs, etc.), the incidence rate of flow decreased significantly and both stuck and off-task behaviors were more likely experienced. Particularly towards the end of the sessions, stuck and off-task became more common; a negative interaction trend was as or more likely than a positive interaction experience. Switching the learning environment (i.e. starting a new activity with a lower level of difficulty) reversed this pattern. That is the incidence of stuck and off-task decreased and the state of flow became dominant again.

Emotional expressions of the experienced trends

Our next investigation was to analyze the learners' emotional responses with regards to the states of flow, stuck and off-task. More precisely, we were interested in answering the following questions: (1) Are there any significant differences in terms of stress, confusion, boredom and frustration, as the learners' interaction was optimal, problematic or completely inhibited? (2) If so, is there a particular emotional pattern associated to each trend? That is which emotion(s) could potentially characterize or contribute to each state, and how? (3) Did all the learners share the same pattern?

Three MANOVAs were conducted to test the relationship between the interaction experience trends and the emotional responses reported in each of the three environments. The dependent variable was the combined intensities of the four emotions (i.e. stress, confusion, boredom and frustration), and the independent variable was the interaction trends (i.e. flow, stuck, and off-task). We found that each of three MANOVAs was statistically significant, showing that there is a significant interplay between the combined expressed emotions and the interaction feedback. $F(8, 1398) = 73.81$, $p < 0.001$; Pillai's Trace = 0.59, partial $\varepsilon^2 = 0.29$ for trigonometry, $F(8, 518) = 27.32$, $p < 0.001$; Pillai's Trace = 0.59, partial $\varepsilon^2 = 0.29$ for BDS, and $F(8, 1750) = 101.66$, $p < 0.001$; Pillai's Trace = 0.63, partial $\varepsilon^2 = 0.31$ for logic. Hence the emotional responses do seem to significantly characterize the type of the interaction. An analysis of each emotion aside was performed using distinct ANOVAs (4 * 3). The results were statistically significant for all the ANOVAs ($p < 0.001$); a summary is given in Table 1. Bonferroni posthoc tests showed that the three trends were significantly different in terms of the four emotions ($p < 0.001$). The state of flow was characterized by a low level (]0; 35]) of stress (around 17 and 25) and confusion (around 16 and 19), and a very low level of boredom (around 6 and 9) and frustration (around 10 and 15). The state of stuck was marked by a moderate level (]35; 65]) of stress (44 to 46) , confusion (about 57) and frustration (36 to 50), and a

low level of boredom (20 to 25). The off-task trend concurred with the highest level of stress (still moderate: 50 to 57, but more intense as compared to stuck), a high level (]65; 100]) of confusion (70 to 77), a moderate level of boredom (49 to 64), and a moderate to high level of frustration (about 58 to 68).

Tab. 1 Descriptive statistics on intensities of emotions for each interaction trend in the three learning environments. Standard errors given in parentheses. ANOVAs reported in italics ($p < 0.001$).

Environment	Experience trend	Stress	Confusion	Boredom	Frustration
Trigonometry	Flow	22.72 (1.37)	18.46 (1.36)	7.64 (1.20)	15.28 (1.40)
	Stuck	46.37 (1.94)	57.96 (1.94)	25.03 (1.71)	50.17 (1.99)
	Off-Task	49.69 (2.80)	74.30 (2.80)	58.66 (2.46)	67.58 (2.87)
	F(2, 701) =	*69.47*	*238.65*	*180.45*	*189.163*
BDS	Flow	24.91 (1.98)	18.88 (2.07)	9.47 (1.87)	10.62 (1.78)
	Stuck	45.85 (3.53)	57.26 (3.70)	20.47 (3.34)	36.04 (3.17)
	Off-Task	50.36 (3.97)	70.36 (4.16)	49.41 (3.75)	58.45 (3.56)
	F(2, 261) =	*24.51*	*84.53*	*45.77*	*82.36*
Logic	Flow	17.13 (1.20)	16.36 (1.23)	6.71 (1.11)	11.25 (1.16)
	Stuck	44.01 (1.72)	57.32 (1.76)	24.73 (1.58)	45.47 (1.66)
	Off-Task	56.79 (2.30)	77.19 (2.35)	64.45 (2.11)	66.39 (2.22)
	F(2, 877) =	*157.47*	*353.30*	*297.16*	*307.07*

From these analyses, it can be said that there was not a unique emotion behind the nature of the interaction, but the four concurrent emotions (stress, confusion, boredom and frustration) seemed to contribute significantly in the expression of flow, stuck and-off-task. In overall (i.e. across all the participants), low stress and confusion seemed to be more likely associated to a positive trend of interaction. Frustration was also experienced, but with a very smaller degree, and boredom was practically absent with flow. The state of stuck was characterized with significantly higher levels of stress, confusion, frustration and boredom. The off-task behavior was likely associated to the worst emotional responses (i.e. the highest levels of stress, confusion, boredom and frustration). However, the case-by-case analysis showed that this pattern was not shared by all the study subjects.

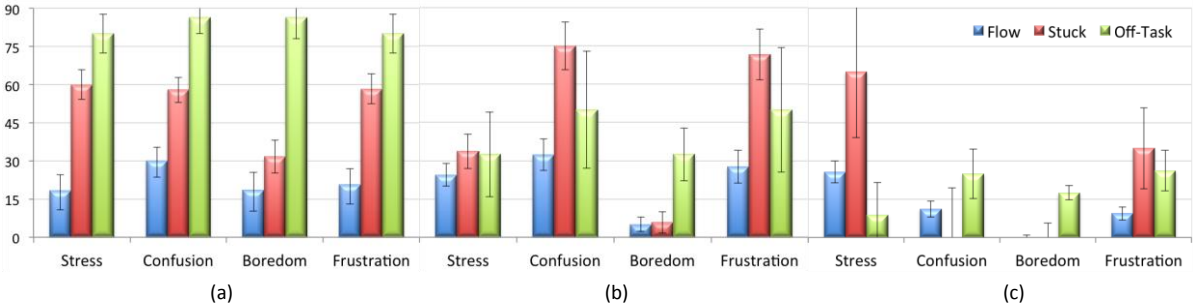

Fig. 5 Three different patterns of learners' emotional responses. MANOVA tests revealed a significant effect of the interaction trends for the three cases: (a) $F(8, 74) = 7.75$, $p < 0.001$; Pillai's Trace = 0.91, partial $\varepsilon^2 = 0.46$ (b) $F(8, 74) = 3.43$, $p < 0.05$; Pillai's Trace = 0.54, partial $\varepsilon^2 = 0.27$ (c) $F(8, 74) = 4.42$, $p < 0.001$; Pillai's Trace = 0.65, partial $\varepsilon^2 = 0.33$.

Separate correlational analyses were run for each participant. MANOVAs results revealed a statistically significant effect of the experienced trends for all the participants ($p < 0.05$), but with different emotional reactions. Figure 5 depicts an example of three distinct patterns: for the first participant (a), a significant effect was found for the four emotions ($F(2, 39) = 22.43$ for stress, 23.18 for confusion, 20.56 for boredom, and 19.73 for frustration, $p < 0.001$ for the four ANOVAs), showing that all four emotions do significantly contribute in the expression of flow, stuck and-off-task. Bonferroni post-hoc tests showed a statistically significant increase of the intensity of the four emotions from the state of flow to stuck, and from the state of stuck to off-task; that is as the typical

case discussed above. For the second subject (b), there were no significant differences of stress between the three types of interaction ($F(2, 39) = 0.71$, p = n.s.), and as a matter of fact, this subject did not seem to experience much stress during the experiment (Max = 33.75). A significant contribution of boredom was found ($F(2, 39) = 3.32$, $p < 0.05$), with the highest values (a low level), for the off-task trend (M = 32.5 (10.35))), but there were no reliable differences between flow and stuck. Significant contributions of confusion ($F(2, 39) = 7.30$, $p < 0.05$), and frustration ($F(2, 39) = 6.85$, $p < 0.05$), were also found. But unlike the overall pattern, the highest values were associated to the state of stuck rather than off-task (M = 75 (9.36), and M = 71.67 (9.99) respectively for confusion and frustration). Besides, there were no significant differences between the states of flow and off-task, in terms of confusion and frustration. For the third subject (c), a totally different pattern was found: there was no significant contribution of stress ($F(2, 39) = 2.00$ p = n.s.) or confusion (($F(2, 39) = 1.11$, p = n.s.), and a tendency towards significance for frustration ($F(2, 39) = 3.12$, $p = 0.056$). The unique significant effect was found for boredom ($F(2, 39) = 17.64$, $p < 0.001$), with a low level in the off-task trend (M = 17.5 (2.80), values were close to zero for flow and stuck).

In summary, emotions do seem to be key indicators of a user's learning experience. This relationship showed to include several emotions that may differ from a person to another, which confirmed our expectations about the person-specific nature in expressing emotions. Indeed the case-by-case analysis showed that the emotional responses associated to the states of flow, stuck and off-task, can be specific to each learner. Some learners can experience the same stress, confusion or frustration when they are immersed within a task or get stuck, and fewer reactions when they drop out. Besides, some subjects seemed to have calmer temper and showed little emotional activations, that is no considerable emotional changes between a positive and a negative interaction. Different factors such as the learner's goals, personality or skills, could intervene and make that learners do not all react the same way as they are fully involved within a task, get stuck, or are about to give up; hence the importance of accounting for these individual differences in the assessment of learners' experience.

Learners' interaction experience modeling

Our last objective was to implement and validate our framework for recognizing a learner's interaction experience trend and emotions, based on the observable diagnostic features, the current context and the learner's characteristics. More precisely, given the macro-model described in figure 1, the diagnostic component involved the following modalities: (1) physiological features including EEG_Engag, GSR and HR (2) behavioral variables: Help_req and Mouse_mvt, and (3) performance measures: Resp_time, answer and accuracy. The context involved three variables namely: the difficulty of the task being executed (Task_diff), the presence/absence of hints or help (Help_given) and time constraints (Time_const). The profile involved: the learner's goal regarding the importance of performing the task (having the best score, learning new concepts or just finishing the task), preference (e.g. whether the learner likes trigonometry or not), skill level, frequency of computer usage (Computer_use), conscientiousness personality trait (Perso_consc), and age.[3]

Once the structure of the DBN has been defined, the next step was to train the model parameters that quantify the relationships between the connected nodes. These parameters are given by the a priori probabilities of each predictive node (e.g. p(Skill) over the values 'low', 'moderate' and 'high'), the conditional probability distribution of each node given the outcomes of its parents (e.g. p(Experience trend | Goal) over the values 'flow', 'stuck', 'off-task' given each of the values 'having the best score', 'learning new concepts' and 'finishing the task'), and the transition probabilities between the time slices (e.g. $p(Boredom_t \mid Bordedom_{t-1}, Experience\ trend_t)$ over the values 'interest', 'low', 'moderate' and 'high' boredom given the corresponding values at time t-1 and the current parent's outcomes 'flow', 'stuck' and 'off-task'). We used an iterative approach to automatically train the model parameters from the collected data, namely the EM algorithm [76]. Starting with a random parameter initialization, EM alternates between two steps. The E-step (Expectation) computes the

[3] Although beyond the scope of this paper, it should be mentioned that these particular variables were selected as they showed statistically significant associations with the experienced trends. For instance, no significant correlation was found with regards to the gender variable, which was not included in the model.

likelihood of the completed data given the current parameter estimate and the observed data; unobserved data are filled in with their expected probability distributions.[4] The M-step (Maximization) updates the current parameters by maximizing the data likelihood; i.e. the model parameters that best fit the data. The two steps are iterated until parameter convergence where a local optimal solution is reached. A 10-fold cross validation technique was used to train the parameters and evaluate the model inference for categorizing both the interaction trends and the four concurrent emotions (i.e. stress, confusion, boredom and frustration). The data set was divided into 10 subsets, where 9 subsets were used for the training and the remaining subset was used for the evaluation. The process was repeated 10 times, the accuracy estimates were averaged to yield the overall model inference accuracy reported in table 2 (DBN). The accuracy results were compared to a static approach (i.e. without the temporal dependencies) using static Bayesian networks (SBN), as well as to three non-hierarchical static formalisms namely: naive Bayes (NB) classifiers [77], decision trees (DT) [78] and support vector machines (SVM) [79].

Tab. 2 Model inference accuracy. Outright classification is done by assigning each instance to the class with the highest probability (maximum a posteriori procedure). Participants' matching self-reports are used as a ground truth. For the non-hierarchical approaches (NB, DT and SVM), the inference is achieved only for the experienced trends.

Target	Classes	DBN	SBN	NB	DT	SVM
Interaction experience trend	Flow, Stuck, Off-task	75.63	69.31	63.71	64.07	69.09
Stress	No (calm), Low, Moderate, High	61.09	47.01	N/A	N/A	N/A
Confusion	No (confidence), Low, Moderate, High	60.02	53.71	N/A	N/A	N/A
Boredom	No (interest), Low, Moderate, High	79.95	63.45	N/A	N/A	N/A
Frustration	No (satisfaction), Low, Moderate, High	67.46	55.36	N/A	N/A	N/A
Interaction experience trend	Positive, Negative	82.25	73.12	68.78	69.26	72.23
Stress	Calm to low stress, Moderate to high	82.18	68.95	N/A	N/A	N/A
Confusion	Confidence to low confusion, Moderate to high	81.88	67.41	N/A	N/A	N/A
Boredom	Interest to low boredom, Moderate to high	90.97	71.04	N/A	N/A	N/A
Frustration	Satisfaction to low frustration, Moderate to high	85.38	69.02	N/A	N/A	N/A

As shown in table 2, two test cases were considered. The first case (top) categorizes three outcomes for the interaction experience trend namely: flow, stuck and off-task, and four outcomes for each emotion (e.g. the target variable stress has the following possible outcomes: calm (no stress), low, moderate or high stress). The second case (down) shows the accuracy of a binary categorization, where two outcomes are considered for both the experience trend and the emotion labels. Although there is a loss of information, this last setting is intended to focus on two reverse behaviors in a learner's experience and emotional responses. Thereby the experience trend is either positive/ favorable (i.e. flow) or negative/unfavorable (i.e. stuck or off-task). In the same way, each emotion can be either positive to low, or moderate to highly negative (e.g. calm to low stress, or moderate to high stress). In both cases, DBN yielded the highest accuracy rates for the experience trend and emotion recognition as compared to SBN, NB, DT and SVM. For the first test case, an accuracy rate of 75.63% was achieved for assessing the experience trend, and an accuracy ranging from 60.02% (for confusion) to 79.95% (for boredom), to discriminate between four levels of emotions. For the second (binary) case, an accuracy of 82.25% was reached for categorizing between a positive and a negative interaction, and an accuracy ranging from 81.88% to categorize between a state of confidence-to-low-confusion and moderate-to-high-confusion, to an accuracy of 90.97% to discriminate between the states of interest-to-low-boredom and moderate-to-high-boredom.

These results suggest that the inference of a learner's interaction experience can be accurately achieved through probabilistic inference using three modality measures (physiology, behavior and

[4] Unobserved data included missing information such as corrupted readings due to sensor failure.

performance), in conjunction with context and person-dependent (profile) variables. The dynamic approach using a DBN outperformed the static approaches (SBN, NB, DT and SVM) that do not track the temporal evolution of the learners' states over time. Besides, with non-hierarchical formalisms (i.e. NB, DT and SVM), no distinction can be made between the input variables on the basis of their causal relationships to the learners' states (i.e. the predictive variables of the interaction experience on one side, and the diagnostic variables on the other side): all the features are equally entered as input variables for the three classifiers. Moreover with the three latter techniques, the recognition is done only for the interaction experience trends. Indeed unlike Bayesian networks (SBN and DBN), where a simultaneous inference of several target nodes is made possible through the two-layered hierarchical structure, these approaches do not allow a straight representation of several unknown classes simultaneously.

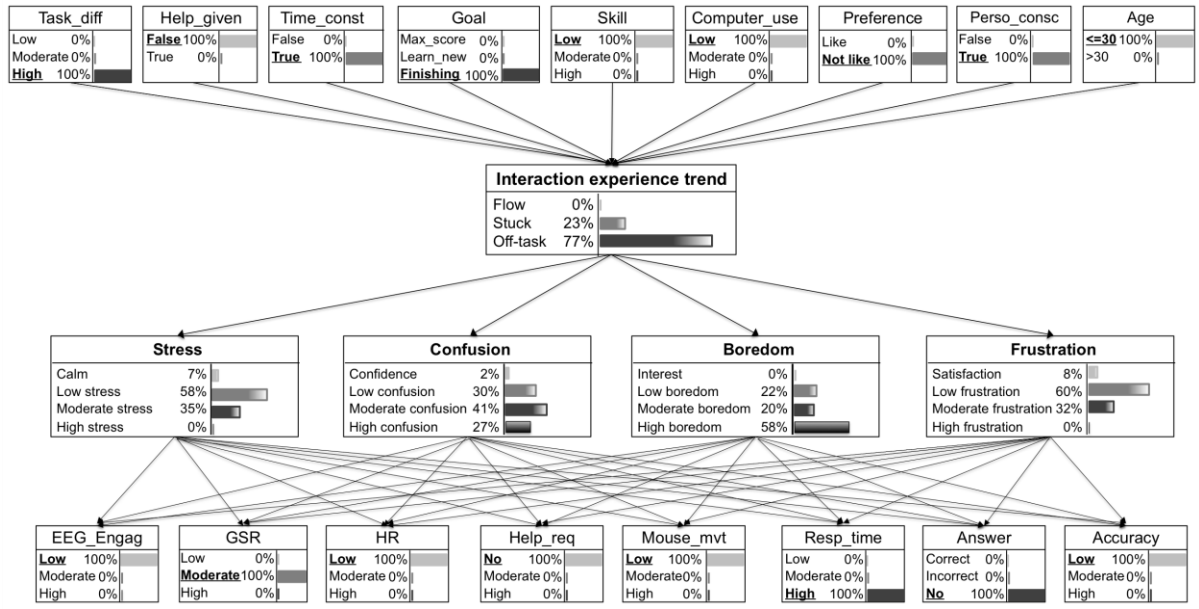

Fig. 6 Inference of a learner's interaction experience from three modality measures (physiology, behavior and performance) and personal and contextual information. The observed evidence are given by the probability values of 100%. Posterior probability distributions are updated for the nodes associated to the interaction experience trend and emotional responses.

The underlying inference of a learner's level of stress, confusion, boredom and frustration through the DBN, can be used as a dashboard for real time adaptation by continuously monitoring the learner's state and assessing the potential cause of a favorable vs. unfavorable interaction, so that an effective intervention can be undertaken. For instance in case of a favorable interaction (i.e. a high probability of flow), the tutoring system would let the learner free to go through the materials without interruption. Implicit interventions such as affective or cognitive primings, can be made to enhance the interaction experience without interrupting the learner's immersion (see [80] for more details). If the learner is about to get stuck (i.e. a high probability for stuck), an explicit intervention would be initiated, while taking into account the learner's emotional changes. For instance in case of high boredom, a more challenging task could be proposed. If frustration, hints could be made available for the learner. In case of high stress, the time constraints can be alleviated, and in case of confusion, a piece of advice or help can be proposed to guide the learner. Similarly, if the learner is about to give up (i.e. a high probability for the off-task trend), a different activity can be proposed with a varying level of challenge, constraints or help, depending on the predominant emotional states.

Figure 6 depicts such an example where a learner's emotional responses and interaction experience trend are inferred using the trained DBN, as new evidence are introduced into the model (predictive and diagnostic nodes). The task at hand is the last trigonometric problem in series 3. The predictive variables are given by the current context: a high level of difficulty, no help provided and a time constraint imposed, the learner's current goal: finishing the activity, and characteristics: less than 30 years, conscientious, low computer usage, low skills and does not like trigonometry. The diagnostic evidence are given by the learner's cerebral activity: low EEG engagement, dermal

response: moderate GSR, and cardiac activity: low HR; behavioral variables: no help request, and low mouse movement rate; and performance: high response time, no-answer to the given problem, and low accuracy rate. The inference yields the following outcomes: a low level of stress (with a probability P = 58%), a moderate confusion (P = 41%), a high level of boredom (P = 58%) and a low frustration (P = 60%). The predominant inferred experienced trend is an off-task behavior (P = 77%). In this case, the system would for instance interrupt the learner to propose a break and change the type of the activity with a more challenging task, as a state of high boredom is detected with a high probably of giving up.

6 Conclusion

In this paper we described a hierarchical probabilistic framework to model the user's experience while interacting with a computer-based learning environment. The framework uses a dynamic Bayesian network to recognize three trends of the interaction experience, namely: flow or the optimal interaction (a total involvement within the task), stuck or the non-optimal interaction (a difficulty to maintain focused attention) and off-task or the non-interaction (a drop out from the task), as well as the emotional responses occurring subsequently. The network integrates three-modality measurements to diagnose the learner's experience namely: physiology, behavior and performance, predictive variables including contextual features and the learner's personal characteristics (profile), and a dynamic structure to track the temporal changes of the learner's state. An experimental protocol was conducted, while 44 participants performed different cognitive tasks (trigonometry, backward digit span and logic) with a gradual difficulty level to provoke the three-targeted trends, and analyze their relationship with the reported emotional responses. Three biofeedback devices were used to record participants' physiological activities including skin conductance, heart rate and EEG engagement. Behavioral variables included the help use and mouse movement rate, and performance measures included response time, answer and accuracy.

The statistical analysis supported our hypothesis about the complexity of the relationship between emotions and learners' experiences. Results showed that concurrent emotional responses can be associated to the experiences of flow, stuck and off-task, and that the same trend could be expressed with different emotional patterns for different participants; which confirmed the importance of accounting for overlapping emotional changes and individual differences in the assessment of the learners' interaction experience. The evaluation of the proposed framework showed its capability to efficiently assess the probability of experiencing flow, stuck and off-task, as well as the emotional responses associated to each trend. The experimental results showed that our framework outperformed conventional non-dynamic modeling approaches using static Bayesian networks, as well as three non-hierarchical formalisms including naive Bayes classifiers, decision trees and support vector machines. An accuracy rate of 82% was reached to characterize a positive vs. a negative experience, and an accuracy ranging from 81% to 90% was achieved to assess four emotions related to the interaction namely stress, confusion, frustration and boredom.

Our findings have implications for intelligent tutoring systems in particular, and for human-computer applications more generally, seeking to acquire a precise monitoring of the user state, by simultaneously identifying concurrent emotional responses occurring during the interaction, and the tendency that characterizes their experiences within the task. As our next steps, we plan to enhance the proposed framework with a decision theoretic formalism, and incorporate it within a real time interaction based tutoring system, so that timely interventions can be formulated on the basis of the user's inferred state. Further diagnostic variables will be included within the model to track additional features of the user's experience including keyboard interaction patterns, facial expressions, etc., as well as a cognitive component to monitor the learner's skill acquisition process including the history of the presented concepts, the practiced skills, etc., to optimally adapt the pedagogical content and strategies according to the learner's state.

Acknowledgments. This research was funded by the National Science and Engineering Research Council (NSERC). The models described in this paper were implemented using SMILE, a reasoning

engine for graphical probabilistic models and the GeNIe modeling environment, both developed at the Decision Systems Laboratory, University of Pittsburgh (http://genie.sis.pitt.edu/).

References

[1] L. Alben, 1996, Quality of experience: Defining the criteria for effective interaction design, *interactions* 3(3), 11-15. doi: 10.1145/235008.235010

[2] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven, 2007, EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks, *Aviation, Space, and Environmental Medicine* 78(5), B231-B244.

[3] S. Brave, and C. Nass, 2002, Emotion in human-computer interaction., in J. Jacko, and A. Sears, eds.: *Handbook of human-computer interaction* (Elsevier Science Pub Co., pp. 81-96).

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, 2001, Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine* 18(1), 32-80. doi: 10.1109/79.911197

[5] E. Hudlicka, 2003, To feel or not to feel: The role of affect in human-computer interaction, *International Journal of Human-Computer Studies* 59(1-2), 1-32. doi: 10.1016/s1071-5819(03)00047-8

[6] S.-H. Matthias, M. Uwe, and K. Thomas, 1993. *Adaptive user interfaces: Principles and practice* (Elsevier Science Inc., New York, NY, USA).

[7] R. Picard, 1997. *Affective computing* (MIT Press).

[8] A. Bechara, H. Damasio, and A. R. Damasio, 2000, Emotion, decision making and the orbitofrontal cortex, *Cerebral Cortex* 10 (3), 295-307. doi: 10.1093/cercor/10.3.295

[9] A. Damasio, 1994. Descartes' error: Emotion, reason and the human brain (Grosset and Putnam, New York).

[10] D. Goleman, 1995. *Emotional intelligence* (Bantam Books, New York).

[11] A. M. Isen, 2000, Positive affect and decision making, in M. Lewis, and J. M. Haviland-Jones, eds.: *Handbook of emotions* (Guilford Press, New York, pp. 417-435).

[12] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, 2009, Emotion sensors go to school, *Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (IOS Press, pp. 17-24).

[13] C. Conati, and H. Maclaren, 2009, Empirically building and evaluating a probabilistic model of user affect, *User Modeling and User-Adapted Interaction* 19(3), 267-303. doi: 10.1007/s11257-009-9062-8

[14] S. D'Mello, S. Craig, A. Witherspoon, B. McDaniel, and A. Graesser, 2008, Automatic detection of learner's affect from conversational cues, *User Modeling and User-Adapted Interaction* 18(1), 45-80. doi: 10.1007/s11257-007-9037-6

[15] K. Forbes-Riley, and D. Litman, 2010, Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system, *Comp. Speech Lang.* 25(1), 105-126. doi: 10.1016/j.csl.2009.12.002

[16] A. Kapoor, W. Burleson, and R. W. Picard, 2007, Automatic prediction of frustration, *International Journal of Human-Computer Studies* 65(8), 724-736. doi: 10.1016/j.ijhcs.2007.02.003

[17] S. W. McQuiggan, S. Lee, and J. C. Lester, 2007, Early prediction of student frustration, *2nd international conference on Affective Computing and Intelligent Interaction* (Springer-Verlag, Lisbon, Portugal).

[18] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard, 2009, Affect aware tutors : Recognising and responding to student affect, *International Journal of Learning Technology* 4,(3/4) 129-164. doi: http://dx.doi.org/10.1504/IJLT.2009.028804

[19] J. Allanson, and S. H. Fairclough, 2004, A research agenda for physiological computing, *Interacting with Computers* 16(5), 857-878.

[20] J. T. Cacioppo, G. C. Berntson, K. M. Poehlmann, and T. A. Ito, 2000, The psychophysiology of emotions., in M. Lewis, and J. M. Haviland-Jones, eds.: *Handbook of emotions, 2nd edition* (The Guilford Press).

[21] S. H. Fairclough, 2009, Fundamentals of physiological computing, *Interacting with Computers* 21(1-2), 133-145. doi: http://dx.doi.org/10.1016/j.intcom.2008.10.011

[22] R. S. J. d. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser, 2010, Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments, *International Journal of Human-Computer Studies* 68(4), 223-241. doi: 10.1016/j.ijhcs.2009.12.003

[23] S. Craig, A. Graesser, J. Sullins, and B. Gholson, 2004, Affect and learning: An exploratory look into the role of affect in learning with autotutor, *J. Edu. Med.* 29(3), 241-250.

[24] R. W. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, and C. Strohecker, 2004, Affective learning - a manifesto, *BT Tech. J.* 22(4), 253-269. doi: http://dx.doi.org/10.1023/B:BTTJ.0000047603.37042.33
[25] A. C. Graesser, S. K. D'mello, P. Chipman, B. King, and B. McDaniel, 2007, Exploring relationships between affect and learning with autotutor, *13th International Conference on Artificial Intelligence in Education* (pp. 16-23).
[26] N. Hara, 2000, Student distress in a web-based distance education course, *Information, Communication & Society* 3(4), 557-579. doi: 10.1080/13691180010002297
[27] K. O'Regan, 2003, Emotion and e-learning, *Journal of Asynchronous Learning Networks, JALN* 7(3), 78-92.
[28] L. A. Muse, S. G. Harris, and H. S. Feild, 2003, Has the inverted-u theory of stress and job performance had a fair test?, *Human Performance* 16(4), 349-364. doi: 10.1207/S15327043HUP1604_2
[29] S. E. Sullivan, and R. S. Bhagat, 1992, Organizational stress, job satisfaction and job performance: Where do we go from here?, *Journal of Management* 18(2) 353-374. doi: 10.1177/014920639201800207
[30] A. C. Graesser, and B. A. Olde, 2003, How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down, *Journal of Educational Psychology* 95, 524-536.
[31] K. VanLehn, S. Siler, C. Murray, T. Yamauchi, and W. B. Baggett, 2003, Why do only some events cause learning during human tutoring?, *Cognition and Instruction* 21(3), 209-249.
[32] K. Hone, 2006, Empathic agents to reduce user frustration: The effects of varying agent characteristics, *Interact. Comput.* 18(2), 227-245. doi: 10.1016/j.intcom.2005.05.003
[33] J. Klein, Y. Moon, and R. Picard, 2002, This computer responds to user frustration—theory, design, and results, *Interacting with Computers* 14(2), 119-140.
[34] J. Healey, and R. Picard, 2000, Smartcar: Detecting driver stress, *15th International Conference on Pattern Recognition* (Vol. 4, pp. 218-221).
[35] R. W. Picard, and K. K. Liu, 2007, Relative subjective count and assessment of interruptive technologies applied to mobile monitoring of stress, *International Journal of Human-Computer Studies* 65(4), 361-375. doi: http://dx.doi.org/10.1016/j.ijhcs.2006.11.019
[36] J. Zhai, and A. Barreto, 2006, Stress detection in computer users based on digital signal processing of noninvasive physiological variables, *EMBS '06. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 1355-1358).
[37] I. Jraidi, and C. Frasson, 2013, Student's uncertainty modeling through a multimodal sensor-based approach. , *Educational Technology & Society* 16(1), 219-230.
[38] H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark, and S. Peters, 2006, Responding to student uncertainty in spoken tutorial dialogue systems, *International Journal of Artificial Intelligence in Education (IJAIED)* 16(2), 171-194.
[39] J. A. Caldwell, K. K. Hall, and B. S. Erickson, 2002, EEG data collected from helicopter pilots in flight are sufficiently sensitive to detect increased fatigue from sleep deprivation, *The International Journal of Aviation Psychology* 12(1), 19 - 32.
[40] A. Heitmann, R. Guttkuhn, U. Trutschel, and M. Moore-Ede, 2001, Technologies for the monitoring and prevention of driver fatigue, *First International Driving Symposium on Human Factors in Driving Assessment, Training, and Vehicle Design* (pp. 81-86).
[41] Q. Ji, P. Lan, and C. Looney, 2006, A probabilistic framework for modeling and real-time monitoring human fatigue, *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans,* 36(5), 862-875. doi: 10.1109/tsmca.2005.855922
[42] K. H. Kim, S. W. Bang, and S. R. Kim, 2004, Emotion recognition system using short-term monitoring of physiological signals, *Medical and Biological Engineering and Computing* 42(3), 419-427. doi: 10.1007/bf02344719
[43] K. R. Scherer, T. Wranik, J. Sangsue, V. Tran, and U. Scherer, 2004, Emotions in everyday life: Probability of occurrence, risk factors, appraisal and reaction patterns, *Social Science Information* 43(3), 499-570.
[44] M. Pantic, and L. J. M. Rothkrantz, 2004, Facial action recognition for facial expression analysis from static face images, *IEEE Transactions on Systems, Man, and Cybernetic, Part B,* 34(3), 1449-1461. doi: 10.1109/tsmcb.2004.825931
[45] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, 2009, The painful face – pain expression recognition using active appearance models, *Image and Vision Computing* 27(12), 1788-1796. doi: http://dx.doi.org/10.1016/j.imavis.2009.05.007
[46] L. Ma, and K. Khorasani, 2004, Facial expression recognition using constructive feedforward neural networks, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* 34(3), 1588-1595. doi: 10.1109/tsmcb.2004.825930

[47] V. Petrushin, 1999, Emotion in speech: Recognition and application to call centers, *Artificial Neural Networks in Engineering* (St. Louis, MO).
[48] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, 2002, Prosody-based automatic detection of annoyance and frustration in human-computer dialog, *7th International Conference on Spoken Language Processing* (Denver, Colorado, pp. 2037–2040).
[49] D. J. Litman, and K. Forbes-Riley, 2004, Predicting student emotions in computer-human tutoring dialogues, *42nd Annual Meeting on Association for Computational Linguistics* (Association for Computational Linguistics, Barcelona, Spain).
[50] R. E. Kaliouby, and P. Robinson, 2004, Real-time inference of complex mental states from facial expressions and head gestures, *Computer Vision and Pattern Recognition Workshop* (IEEE Computer Society, Vol. 10, pp. 154).
[51] W. Liao, W. Zhang, Z. Zhu, Q. Ji, and W. D. Gray, 2006, Toward a decision-theoretic framework for affect recognition and user assistance, *International Journal of Human-Computer Studies - Human-computer interaction research in the managemant information systems discipline* 64(9), 847-873. doi: 10.1016/j.ijhcs.2006.04.001
[52] M. Csikszentmihalyi, 1990. *The psychology of optimal experience* (Harper & Row, New York).
[53] W. Burleson, and R. W. Picard, 2004, Affective agents: Sustaining motivation to learn through failure and a state of stuck, *Social and Emotional Intelligence in Learning Environments, Workshop In conjunction with the 7th International Conference on Intelligent Tutoring Systems* (Maceio - Alagoas, Brasil).
[54] R. S. Baker, A. T. Corbett, I. Roll, and K. R. Koedinger, 2008, Developing a generalizable detector of when students game the system, *User Modeling and User-Adapted Interaction* 18(3), 287-314. doi: 10.1007/s11257-007-9045-6
[55] J. E. Beck, 2005, Engagement tracing: Using response times to model student disengagement, *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* (IOS Press).
[56] M. M. T. Rodrigo, R. S. J. d. Baker, M. C. V. Lagud, S. A. L. Lim, A. F. Macapanpan, S. A. M. S. Pascua, J. Q. Santillano, L. R. S. Sevilla, J. O. Sugay, S. Tep, and N. J. B. Viehland, 2007, Affect and usage choices in simulation problem-solving environments, *Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work* (IOS Press, pp. 145-152).
[57] B. Cowley, D. Charles, M. Black, and R. Hickey, 2008, Toward an understanding of flow in video games, *Comput. Entertain.* 6(2), 1-27. doi: 10.1145/1371216.1371223
[58] K. Jegers, 2007, Pervasive game flow: Understanding player enjoyment in pervasive gaming, *Comput. Entertain.* 5(1), 9. doi: 10.1145/1236224.1236238
[59] P. Sweetser, and P. Wyeth, 2005, Gameflow: A model for evaluating player enjoyment in games, *Comput. Entertain.* 3(3), 3-3. doi: 10.1145/1077246.1077253
[60] K. P. Murphy, 2002, *Dynamic bayesian networks: Representation, inference and learning*, (PhD thesis. Computer science division, University of California, Berkeley, CA. USA.).
[61] P. J. Lang, 1995, The emotion probe: Studies of motivation and attention, *American Psychologist* 50(5), 372-385. doi: 10.1037/0003-066x.50.5.372
[62] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, 1993, Looking at pictures: Affective, facial, visceral, and behavioral reactions, *Psychophysiology* 30(3), 261-273. doi: 10.1111/j.1469-8986.1993.tb03352.x
[63] J. Altimiras, 1999, Understanding autonomic sympathovagal balance from short-term heart rate variations. Are we analyzing noise?, *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* 124(4), 447-460.
[64] R. W. Levenson, 1992, Autonomic nervous system differences among emotions, *Psychological Science* 3(1), 23-27. doi: 10.2307/40062748
[65] C. Berka, D. J. Levendowski, M. M. Cvetinovic, M. M. Petrovic, G. Davis, M. N. Lumicao, V. T. Zivkovic, M. V. Popovic, and R. Olmstead, 2004, Real-time analysis of eeg indexes of alertness, cognition, and memory acquired with a wireless eeg headset, *International Journal of Human-Computer Interaction* 17(2), 151 - 170.
[66] M. B. Sterman, and C. A. Mann, 1995, Concepts and applications of eeg analysis in aviation performance evaluation, *Biological Psychology* 40(1-2), 115-130. doi: http://dx.doi.org/10.1016/0301-0511(95)05101-5
[67] R. Stevens, T. Galloway, and C. Berka, 2007, EEG-related changes in cognitive workload, engagement and distraction as students acquire problem solving skills, in C. Conati, K. McCoy, and G. Paliouras, eds.: *User modeling 2007* (Springer Berlin / Heidelberg, Vol. 4511, pp. 187-196).
[68] A. T. Pope, E. H. Bogart, and D. S. Bartolome, 1995, Biocybernetic system evaluates indices of operator engagement in automated task, *Biological Psychology* 40(1-2), 187-195.
[69] F. G. Freeman, P. J. Mikulka, L. J. Prinzel, and M. W. Scerbo, 1999, Evaluation of an adaptive automation system using three eeg indices with a visual tracking task, *Biological Psychology* 50(1), 61-76.

[70] M. Chaouachi, P. Chalfoun, I. Jraidi, and C. Frasson, 2010, Affect and mental engagement: Towards adaptability for intelligent systems, *23rd International FLAIRS Conference*, (AAAI Press, Daytona Beach, Florida, USA).
[71] R. Pekrun, A. J. Elliot, and M. A. Maier, 2006, Achievement goals and discrete achievement emotions: A theoretical model and prospective test, *Journal of Educational Psychology* 98(3), 583-597.
[72] M. Chaouachi, I. Jraidi, and C. Frasson, 2011, Modeling mental workload using eeg features for intelligent systems, in J. Konstan, R. Conejo, J. L. Marzo & N. Oliver, eds.: *User modeling, adaption and personalization* (Springer Berlin Heidelberg, Vol. 6787, pp. 50-61).
[73] Thought_Technology_Ltd., 2007, Http://www.Thoughttechnology.Com/bioinf.Htm.
[74] H. H. Jasper, 1958, The ten-twenty electrode system of the international federation, *Electroencephalography and Clinical Neurophysiology* (10) 371-375.
[75] O. P. John, L. P. Naumann, and C. J. Soto, 2008, Paradigm shift to the integrative big-five trait taxonomy: History, measurement, and conceptual issues, in O. P. John, R. W. Robins, and L. A. Pervin, eds.: *Handbook of personality: Theory and research* (NY, pp. 114-158).
[76] S. L. Lauritzen, 1995, The EM algorithm for graphical association models with missing data, *Computational Statistics & Data Analysis* 19(2), 191-201. doi: http://dx.doi.org/10.1016/0167-9473(93)E0056-A
[77] P. Domingos, and M. Pazzani, 1997, On the optimality of the simple bayesian classifier under zero-one loss, *Machine Learning* 29(2-3), 103-130. doi: 10.1023/a:1007413511361
[78] J. R. Quinlan, 1986, Induction of decision trees, *Machine Learning* 1(1), 81-106. doi: 10.1023/a:1022643204877
[79] J. C. Platt, 1999, Fast training of support vector machines using sequential minimal optimization, *Advances in kernel methods* (MIT Press, pp. 185-208).
[80] I. Jraidi, P. Chalfoun, and C. Frasson, 2012, Implicit Strategies for Intelligent Tutoring Systems, In S. Cerri, W. Clancey, G. Papadourakis & K. Panourgia, eds.: *Intelligent Tutoring Systems* (Springer Berlin Heidelberg, Vol. 7315, pp. 1-10).