Going Beyond Performance Scores: Understanding Cognitive-Affective States in Kindergarteners

Priyashri K. Sridhar [1,2]
priya@ahlab.org

Samantha W.T. Chan [1]
samantha@ahlab.org

Suranga Nanayakkara [1]
suranga@ahlab.org

[1]Augmented Human Lab, Auckland Bioengineering Institute, The University of Auckland
[2]Singapore University of Technology and Design

ABSTRACT

Cognitive-affective states during learning or interactions with technologies is dependent on the mental effort of the learner and / or the cognitive load imposed by the system. Despite the growing research on the importance of understanding cognitive-affective state and their relationship to learning, measurement of such states during the learning process in Kindergartners is still unclear. While most assessments of learning and usability evaluations with Kindergartners focus on performance, self-reports and inferring from observable behaviours, they provide limited insights into the cognitive load and emotional state during the learning or interaction that are essential for a holistic picture of learning. Through a study with 18 kindergartners, we explore the feasibility of understanding cognitive-affective states associated with mental effort by triangulating the data obtained from observations, physiological markers, self-reports and performance as they performed tasks of varying mental effort. We present findings on the reliable markers within these sources across tasks. Results reveal that such a triangulation offers deeper insights into the cognitive-affective state of the learner. We believe this work would be a step towards better understanding of the learning process thereby facilitating instruction that is more aligned with the learner's cognitive-affective architecture as well as establishing guidelines for comprehensive usability / evaluation processes based on well-defined associations between child behaviour and child action.

CCS Concepts

•**Human-centered computing** → **Interaction design process and methods;**

Author Keywords

Emotion / Affect; Children; Learning; Empirical Study

INTRODUCTION

Educational psychologists have increasingly emphasized the importance of kindergarten education in a child's overall development. A significant way of creating enriching experiences comes from a thorough understanding of the cognitive-affective state of the child and following the learning behaviour. Extensive research on measuring cognitive load through self-reports [41] provide limited insight to the quantity of knowledge and no information on the learner's cognitive load or emotions during the learning process [42]. The major pitfall of these methods when used alone is that these measures are static (measured at a single point in time), thereby making them inappropriate for measuring variations in cognitive load over a continuous time frame. Furthermore, there are mixed views on the accuracy and reliability of self-reports especially with children [46].

In order to determine how to respond to the temporal and subtle changes of cognitive-affective states as well as improve the reliability of subjective responses, it is necessary to objectively measure the cognitive load of kindergartners in real-time and in-situ. While physiological and neurological measures such as skin conductance, heart rate variability and Functional Near-Infrared Spectroscopy (fNIRS) have been explored in the context of cognitive load, such research is primarily focused on adults. To our knowledge there has been very limited exploration in understanding the physiological changes that correspond to cognitive-affective states during learning in kindergartners. By determining the objective and subjective markers that correspond to increased cognitive load, we can help understand the learner's cognitive-affective state during learning.

As a first step in this direction, we investigated the feasibility of obtaining physiological measurements from kindergartners. We then conducted a controlled study with standardized cognitive tasks of varying difficulty to identify the suitable physiological markers. We found that specific markers within skin conductance and heart rate are linked to increasing cognitive load. However, due to the specificity problem related to physiological measures, where one physiological measure maps to different psychological phenomena, it was challenging to infer the cognitive load experienced by the learner only with physiological data. Hence, we triangulated the physiological measures with observational data to understand the events and emotions that accompanied or triggered the physiological change. This triangulation helped us tease apart the pattern of physiological measures and revealed better insights into the cognitive-affective state.

IDC '18, June 19–22, 2018, Trondheim, Norway

 ISBN 978-1-4503-5152-2/18/06…15.00

DOI: https://doi.org/10.1145/3202185.3202739

The contribution of this paper are three-fold:

1. Exploration of the feasibility of obtaining physiological measurements of cognitive load from kindergartners as they engage in cognitive tasks
2. Conducting a user study to obtain physiological, observational and performance measures and triangulating the results from the three sources to better understand cognitive-affective states during performance of cognitive tasks.
3. Discussing insights on formulating and implementing triangulated study designs with kindergartners.

RELATED WORK

Assessment of Cognitive States in Learning

One of the popular approaches of assessment of cognitive states has focused on cognitive load imposed by the learning or the mental effort exerted by the student and the demonstration of proficiency in a subject matter. York et al. [58] attempt to evaluate programs that work towards attaining such focus such as "Knowledge of Individual Students' Skills" (KISS) [7] that are based on how well teachers' ratings align with students' actual proficiencies. Teachers may also rely on observable behaviours in the classroom to infer about a child's learning status. While these offer insights into a learner's knowledge acquisition, they usually are conducted after lesson delivery and therefore do not offer much on what really happens during the learning process itself. As a result, there is not much information on how much mental effort is being exerted by a student on a learning task, whether the task imposes extreme cognitive load, the nature of the cognitive load as well as what are the range of emotions the student goes through. According to Kirschner [28], performance is not a measure of cognitive load but rather determined by the cognitive abilities of the learner, mental effort, task complexity and environment. Access to such information enables timely intervention of the teacher to redesign instruction in a way that makes learning enjoyable to the child.

There has been a host of methods beyond performance scores that have been used to infer about cognitive load / mental effort as well as how a student feels about a task. The subjective or self-reporting method [41] has been the most commonly used method with adults due to its convenience. However, for children below age 11, self-reports have low validity due to their limited language ability, reading age, motor skills, temperamental effects such as confidence, self-belief and desire to please [46]. Another important aspect of these questionnaires is the time of administration [9]. Most of the studies present the questionnaires after the learning has occurred [44, 20]. As a result, there is a high possibility that the participant may provide an average estimate for the whole task that is affected by memory effects. This loses its purpose of capturing the dynamic and fluctuating nature of load that is imposed during learning [59].

The second common method of measuring cognitive load is using dual or secondary tasks [53, 5] that draw from psychology where a secondary task is introduced along with the primary task of learning. However, as Yuksel et al. [59] point out, a major disadvantage of these tasks is their interference with the primary task especially when the primary task itself is complex and draws much of the learner's cognitive capacity [40]. In addition, the sensory modality of the secondary task may interfere with that of the primary task.

With the inclusion of technology into the learning environments, there have been some encouraging explorations on human thinking and information processing abilities [42]. Much of the studies of cognitive load and learning outcome measures administer each of these measures either before or after test performance [31]. Even though they are static and considered unreliable [43], they continue to be popular in real-world contexts partly because single measures are easy to administer whereas other objective measures of cognitive load may require expensive and hard-to-use instrumentation. In contrast, Mayer et al. [34] have highlighted the need for direct measure of cognitive overload. Physiological measures such as skin conductance and heart-rate variability offer a direct measure of cognitive load [42, 38].

Galvanic Skin Response (GSR)

Research on skin conductance looks at the skin conductance response (SCR) that is triggered by the action of sweat glands to an external stimulus. Researchers [49, 50] have used GSR to differentiate between stress state and cognitive load state, and found correlations between the GSR signal and cognitive load. It has been shown that GSR reduces with increase in cognitive load [23] while others find a weak relationship between skin conductance and cognitive load [16]. Ferreira et al. [18] used perceptual speed and visuo-spatial cognitive processing tasks and collected psycho-physiological data in young and old adults that included GSR. With pre-schoolers, researchers have explored GSR as objective indicators of anxiety [37] or aggression [45]. Such work shows GSR as a potential physiological marker for different behaviours.

Heart Rate Variability

Cognitive load has been shown to have an effect on various components of Heart Rate Variability (HRV) such as mean heart rate (HR), breathing rate, low frequency (LF) and high frequency (HF) components of HRV [3, 22, 36]. People under high mental workload have reduced HF components [22]. The HF component of HRV is indicative of the parasympathetic influence on the heart and is high during rest. During high-attention tasks, absolute measures of LF and HF HRV power have been observed to decrease when compared to a baseline [36]. Mc Duff et al. [35] used remote HRV measures to monitor effect of cognitive workload on HRV and identified the LF and HF components of HRV to be the most indicative of cognitive stress.

Although physiological measures provide a direct measure of cognitive load, there are limitations when they are used by their own. Changes in GSR and HRV can also be mapped to other phenomenon such as changes in emotional states [17]. Bearing this in mind, we use the physiological measures in conjunction with other observational data and the performance across trials and over time.

Role of Affect in Identifying Cognitive Load

Even though affect has been proposed to play a great role in learning [51, 39], it has not been explored as an extension of cognitive theory [42]. There has been consensus on the role of intrinsic vs. extrinsic influences, the influence of past pleasurable experiences and motivation on learning [56, 25, 26, 55]. Some researchers have integrated both affective and cognitive components into motivation theories [12, 1, 13]. While these have provided insight into emotions and reaffirmed their role in learning, there is not much consensus on the kind of emotions involved in learning. Csikszentmihalyi [8] has emphasized the tendency for a pleasurable state of "flow" to accompany problem solving. Kort [29] attempted to list the emotions involved in learning and proposed a four quadrant model relating phases of learning in emotions. There have also been some scattered attempts by other researchers [32, 30] to identify emotions in learning.

One of the reasons why affect is not explored much is that it is hard to measure. While it is easy to get performance scores and test the ability to transfer learning, it is harder to measure how the learner feels during the entire process [42]. Typically affect has been measured through questionnaires that ask how much pleasure, frustration or interest one felt [33, 10, 57]. Moreover, there are specialized instruments for evaluating the motivational characteristics of an instructor's classroom delivery [27] but as shared earlier, self-reports for children are not reliable and valid. Much of earlier work has focused on emotions from exaggerated expressions, making it hard to generalize them to typical learning situations. Kapoor et al. [24] attempted to build a system that detects surface level affect behaviour such as posture, eye-gaze, facial expressions and head movements using pressure sensors and gaze-tracking.

There is a strong need to understand affective states with respect to cognitive load in order to understand the learning state as well as suitably intervene or design interfaces that intervene appropriately. For example, a learner who makes mistakes while appearing engaged and curious is different from a learner who makes mistakes while fidgeting, frowning or displaying other anxious behaviours. While the former is important as it encourages exploration, the latter case demands some intervention, re-instructions and maybe feedback to facilitate learning. While there seems to be a general consensus among educators that interest and engagement are important factors in the learning process, much of this consensus is based on intuition and there is a need for studies that look at affect in learning and in relation to cognitive load.

Cognitive-Affective States in Usability Evaluations in Children

There has been much interest in measuring the usability and engagement of a novel interface that was being designed for children. For a long time now, including children in the design of new technologies, either as informants or design partners has been highlighted as being beneficial to understand users, gather design ideas and to test out new concepts [47, 11]. Since recruiting children in testing of technology designed for them involves ethical concerns as well as methodological considerations, choosing the appropriate usability evaluation method (UEM) is important. UEMs that elicit only verbalization maybe too strict for Kindergartners thereby necessitating a need for some flexibility in the approach to allow the child to express emotions, thoughts and opinions in activities. In addition, survey methods that adopt a question-answer process often are impacted by developmental effects; language, reading age, and motor abilities, as well as temperamental effects including confidence, self-belief and desire to please [46]. Also, when assessing children between 2-6 years for appeal or engagement, testers will need to closely observe behaviors such as sighing, smiling, or sliding under the table. Given the challenges with eliciting the exact response of how children truly feel, an objective method to measure cognitive-affective states as children go through the interaction with a novel interface or UEM maybe potentially useful. Therefore, in this study, we also explore the feasibility of obtaining physiological, behavioural and observational measures with kindergartners keeping in mind its potential in conducting UEMs with children in the future.

METHOD

Pilot Study: Feasibility of Collecting Physiological Parameters from Kindergartners

The pilot was conducted to check the feasibility of the study design and collecting physiological parameters from kindergartners. We selected two established sensors to obtain physiological measurements: Empatica E4 wristband [19] and Consensys Shimmer3 GSR Kit[1].

With our initial exploration with some kindergartners, we noticed that since Shimmer3 GSR had to be secured to the fingers using the strap, it limited the range of movement on the hand where the skin conductance was measured. In addition, we noticed that the participants tend to get distracted with the wires and the electrode placement especially when the tasks demanded key press. Hence, we resorted to use only the Empatica E4 for its ease of use, adjustability to children's wrist, and possibility of obtaining multiple measures simultaneously. In addition, we custom built a mobile app that was synced to the E4 band to visualize data in real-time and save them into a local database.

Participants

Three English-Mandarin bilingual kindergarteners (2 female, 1 male) participated in this study (M_{age} = 5.58, SD = 0.49, age range: 4 to 7 years). The children were recruited from a Kindergarten in a middle-class neighborhood. The average level of parental highest education was a university degree.

Stimuli

We used standardized tasks that map onto retrieval of stored knowledge from long-term memory as well as executive functions namely, inhibition, flexibility and working memory. These skills have been shown to impact learning and elicit mental effort on part of the participant. These tasks elicited different levels of mental effort.

[1] http://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor

Johnson Woodcock IV Test: Three sections from the Test of Cognitive Abilities that constitute the Brief Intellectual Ability were used. The sub-tests included: verbal ability (antonyms and synonyms), verbal attention and number series. Verbal ability requires recall / retrieval, verbal attention is a test of working memory while the number series tests working memory as well as inhibition. The items in each sub-test are arranged in the increasing order of difficulty. In the sub-test on antonyms and synonyms, the child was asked to say the antonym and synonym of the stimulus item respectively. In the verbal attention task, the child was asked to repeat the animal and number combinations in the same order as presented by the experimenter. In the number series task, the child was asked to identify the missing number by understanding the pattern / sequence of the stimulus item. For every stimulus item, the child was given a maximum of 1 minute to respond, failing which, the next item was presented. Each response was scored and each of the sub-test was terminated when the participant responded inaccurately or did not respond for six consecutive test items.

Executive Function (EF) Tasks: A computerized version of the Simon Task [52] and the Dimensional Change Card Sort (DCCS) test [60] were presented using E-Prime software [48]. The Simon tasks involves executive function skills of inhibition and to a small extent, working memory. In Simon Task, the subjects were presented with a red or a blue square on the screen. They were instructed to press a button on the corresponding side of the stimulus. There were three types of trials: congruent, incongruent and mixed. In the congruent trial (lowest load on inhibition), the presentation of the stimulus matched the side of the response key. In the incongruent trial (higher load on inhibition in the first half but reduces over time), the stimulus presentation was located on the side opposite to the response key. In the mixed trial (highest load on inhibition because the trials are all randomly mixed making prediction impossible), there were congruent and incongruent blocks in a random order. The participant was instructed to press the right key as quickly as possible. The DCCS task is a test of inhibition and flexibility. In DCCS, the children were required to sort through a series of bivalent test pictures first according to one dimension (colour) or another dimension (shape). There were two blocks (congruent and mixed) of 20 trials each. In the congruent block (lower load on inhibition and flexibility as the participant sorts according to the same dimension throughout), the participants sorted the stimuli according to colour only. In the mixed block (higher load on inhibition and flexibility as they can be asked to sort on colour and shape interchangeably and randomly), the two dimensions of colour and shape were used interchangeably. The participants were instructed to press one of the two keys to denote their response as quickly as possible. In both tasks, the reaction time and accuracy were calculated.

Procedure

The study was conducted in a quiet room in the participants' school to ensure familiarity of surroundings. The experimenter built rapport with every participant by participating in their classroom activities during play and art lessons. Following this phase, the participants were recruited for the study. The experimenter conducted one-on-one sessions with every child. Each participant wore the E4 wristband prior to the session. They first completed a baseline period of sitting quietly and relaxing for 3 minutes. Following this, they were asked to press some random keys on the keyboard repeatedly to check if the movement affected the measurements. After ascertaining that the key press did not affect the readings, they completed the Johnson Woodcock Tests and the EF tasks (both were counter-balanced). The tasks were completed over a span of two sessions each lasting around 30 minutes. Other ambient conditions such as room temperature and lighting were controlled across participants and sessions.

Findings

Feasibility of Physiological Measurement: The pilot study revealed that it is indeed possible to collect physiological data from kindergartners as they perform tasks with varying cognitive load. E4 wristband was convenient to collect physiological data as the sensors are wireless and the same wristband offers heart rate sensing as well as skin conductance sensing. However, in spite of the adjustable strap and getting the tightest fit, the PPG sensor for heart rate measurement sometimes did not achieve good contact with the wrist of the participant owing to their small wrist size, which resulted in data loss. In order to avert this, we used a small pad of cloth near the strap to enable better contact of the PPG sensor on wrist.

Selection of Physiological Parameters: Based on a preliminary analyses, we found that body temperature did not change much across tasks as compared to baseline. Therefore, we dropped this measure for the main experimental study. The GSR measures, particularly, the number of Skin Conductance Responses (SCRs) and average amplitude of SCRs showed variations across baseline and tasks. In HRV measures, we found a change in mean heart rate (HR), low frequency power of HRV and high frequency power of HRV in tasks as compared to baseline. We therefore decided to use both the GSR and the HRV measures for the main experimental study.

Importance of Baseline and Consistent Ambient Conditions: We realized the need for baseline in similar ambient conditions and between the tasks for every participant. We noticed that at the end of one task, the measures in GSR and HRV shift from a resting baseline. Therefore, there is a need to take a break and bring the values back to resting baseline before proceeding to the next tasks in order to obtain a true measure of the change in physiology for every task.

Main Study: Triangulating Performance, Observational and Physiological Measurements

Incorporating the findings from the pilot study, the objective of this study was to triangulate performance measures, observational data and physiological measurements to explore whether behavioural analysis and physiological data can indeed reveal more insights into the cognitive-affective state of the participant beyond just performance scores / accuracy.

Participants, Stimuli and Procedure

Fifteen English-Mandarin bilingual preschoolers participated in this phase of the study (M_{age} = 5.23, SD = 0.73, age range:

4-6 years; 9 males, 6 females). They were recruited from the same school. The stimuli used were the same as those used in the pilot study and the physiological measures (HRV and GSR) were measured with E4 wrist band. In addition, we collected performance data (response time and accuracy) and video-recorded the sessions for later analysis of emotions and behaviour.

Dependent Variables

We compared the following dependent variables for all the participants across the baseline and experimental (task) conditions:

- *Performance Measures*: This included the percentage correct responses for all tasks. For the EF tasks, we also calculated the response time.
- *Galvanic Skin Response*: This includes the number of Skin Conductance Responses (SCRs), average amplitude of SCRs and the cumulative amplitude of the SCRs for baseline and experimental conditions.
- *Heart Rate Variability*: This includes mean heart rate (HR), heart rate variability (HRV), low frequency (LF) component of HRV and the high frequency (HF) component of HRV.
- *Observable Behaviour*: The coded behaviours included emotion (such as happiness, sadness, anger, disgust, fear, surprise, contempt and neutral [14]), response latency, vocalizations / comments, head movement, postural change, gazing / eye movement and other signs.

Results

Performance Measures

a) Johnson Woodcock IV Tests (JW)

For the four sub-tasks of the JW (see Figure 2), correct responses were assigned a score of "1" (denoted in green) while incorrect responses were assigned a score of "0" (denoted in red) across tasks. The scores were recorded manually by the experimenter. Each sub-test was administered until the participant made six consecutive errors. The last column denotes the performance score (number of correct responses and number of stimuli presented). If one were to merely look at the performance scores and performance accuracy (Figure 1), they do not reveal much about the performance pattern. While some participants had a series of all correct responses followed by six consecutive incorrect responses (e.g. P12 for the Number Series sub-task in Figure 2a), others had some incorrect responses right at the start that was then followed by correct responses (e.g. P11 for the Synonyms sub-task in in

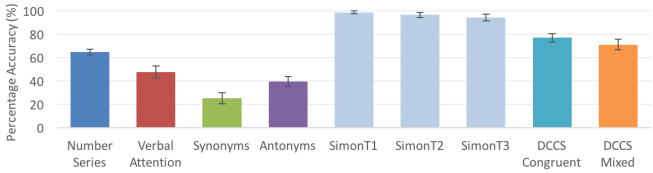

Figure 1. Graph showing the mean accuracy in percentage + standard error (SE) across tasks

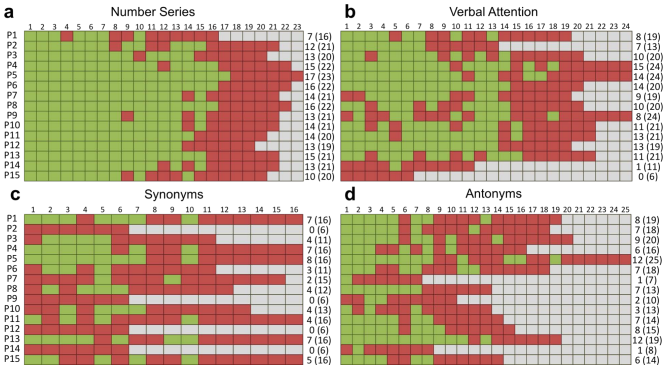

Figure 2. Graphs of performance pattern across the four sub-tasks in the Johnson Woodcock Battery: a) Number Series, b) Verbal Attention, c) Synonyms and d) Antonyms. The green boxes denote correct response, red boxes denote incorrect responses. The grey boxes denote the questions that were not presented.

Figure 2c). The performance plot reveals a better picture of where the errors are.

b) Executive Function (EF) Tasks

Similar to the JW tasks, the EF tasks were also scored “0" for incorrect and “1" for correct responses. In addition to the accuracy, the time taken to complete each sub-task was calculated. Figure 3 reveals the response pattern for each participant across trials. It was noticed that most children performed all the Simon Tasks with at least 95% accuracy (Figure 1). As is the case with JW tasks, the total correct score reveals nothing about where the participants made errors. As expected, the participants took a longer time to complete the DCCS mixed block which requires them to exert cognitive flexibility compared to the DCCS Congruent task (see Figure 4). The performance scores and accuracy in percentage (Figure 1) is aligned with the time taken to complete the task (Figure 4). However, as the difficulty is not increasing in order like the JW tasks, one cannot see a pattern here. While performance scores offer an overall picture of how "well" a participant performed, it does not tell us much about what the participant experienced as they went through the tasks, the pain points and their emotions as they faced easy compared to difficult test items.

Galvanic Skin Response Measures

We analyzed three measures of skin conductance across the tasks: a) the number of Skin Conductance Responses (SCRs) (Please refer to Figure 5), b) the cumulative amplitude of SCRs (Please refer to Figure 6) and c) average amplitude

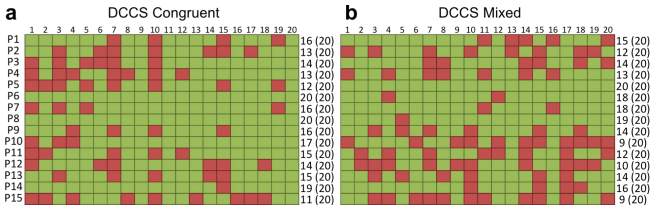

Figure 3. Graphs of performance pattern across the EF tasks: a) DCCS Congruent and b) DCCS Mixed. The green boxes denote correct response and red boxes denote incorrect response.

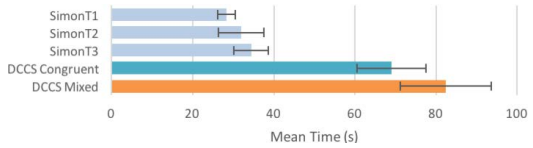

Figure 4. Graph showing the total mean time taken + SE to complete the EF tasks

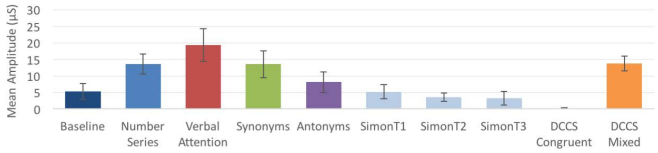

Figure 6. Plot of mean cumulative amplitude of SCRs + SE (in μS) across tasks

of the SCRs. Any change in skin conductance greater than 0.01 microsiemens (μS) was considered as an SCR [4]. A continuous decomposition analysis was done using Ledalab toolkit for Matlab 2016b [2]. Paired t-tests were conducted to compare these measures across the tasks and baseline.

a) Number of SCRs

There was data loss from one participant due to issues with the sensor contact and we excluded that participant from GSR analysis. Among the Johnson Woodcock sub-tasks, the number of skin conductance responses (see Figure 5) was significantly higher than baseline for synonyms (t(13) = -2.608, p = .022), verbal attention (t(13) = -2.352, p = .018) and number series (t(13) = -3.97, p = .008) sub-tasks of the Johnson Woodcock tests as compared to the baseline. Among the EF tasks, there were significantly more SCRs in DCCS-mixed block as compared to the baseline (t(13) = -7.128, p = .00). There was a marginally significant higher number of SCRs in the mixed Simon block than the baseline (t(13) = -1.970, p = .061). No such differences were observed for the DCCS Congruent block as well as the congruent and incongruent trials of Simon Task. The DCCS Congruent block has relatively less cognitive load as compared to the mixed block that may have resulted in no significant difference between the groups. The skin conductance may not have been sensitive to the demands placed by the Simon task as the blocks were of very short duration and easier as compared to the mixed block for which differences were found. When compared to the performance measures, it can be noted that when a task is easy, the number of SCRs is lower. This is best illustrated with the DCCS tasks where the number of SCRs for mixed block (challenging) is higher than the congruent (easier) block.

b) Cumulative SCR Amplitude

The cumulative GSR amplitude refers to the sum of all the SCRs. There was a significant increase in the cumulative SCR amplitude compared to the baseline in the mixed DCCS block (t(13) = 2.199, p = .001); verbal attention task (t(13) = -2.123, p = .0524) and the Number Series task (t(13) = -1.842, p = .056). Even though synonyms had a significantly higher number of SCRs than the baseline, they were probably not of very high amplitude. Although graphs of SCR amplitude (Figure 6) emphasizes the magnitude of difference, the emotional response to the demands of the task is unknown.

c) Average Amplitude of SCRs

We did not find any significant difference in the average value of SCRs across tasks.

GSR analysis reveals whether a certain marker is sensitive to cognitive load in kindergartners and to some extent the amount of cognitive load imposed by different tasks that tap on different cognitive resources. However, they do not accurately pin-point to the nature of cognitive load and how the learner perceived them. For example, were the SCRs lesser because the participant found them too easy and hence boring? Or, were there lesser SCRs because they found the tasks easy and therefore comforting? Since GSR can be mapped to variety of emotional states such as excitement, frustration and engagement, there is a need to supplement this with behavioural observations to get a complete picture of cognitive-affective state.

Heart Rate Measures

We analyzed various measures using Kubios HRV 2.2 [54] on Matlab 2016b. The analysed measures included: mean heart rate (HR), mean inter-beat-intervals (RR), heart rate variability (HRV) and low frequency (LF) and high frequency (HF) components of heart rate variability. Overall, we found that the low frequency component of HRV as shown in Figure 7 (modulated by sympathetic and parasympathetic activity) was significantly higher than baseline values for synonyms (t(14) = -2.361, p = .015), antonyms (t(14) = 2.437, p = .0168), and DCCS Mixed task (t(14) = -3.378, p = .005) and marginally significant for number series (t(14) = 1.922, p = .054). We did not find any significant changes in the parasympathetic activity measured through HF component (Figure 8). This may be owing to the age group of the participants as they may not truly experience relaxation when on a task. While our analysis shows that HRV measures maybe sensitive to cognitive load in kindergartners, they alone do not offer complete picture of how this load was perceived and whether the load resulted in frustration or encouraged them to be more curious and explore.

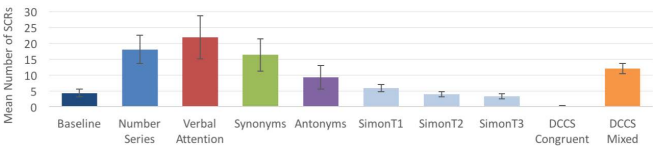

Figure 5. Plot of mean number of SCRs + SE across tasks

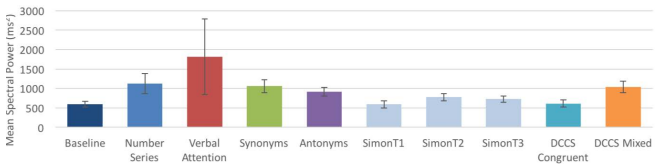

Figure 7. Plot of mean Low Frequency power + SE (in ms^2) across tasks

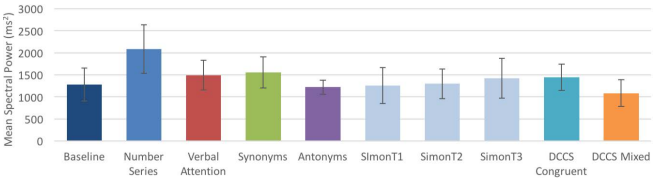

Figure 8. Plot of mean High Frequency power + SE (in ms^2) across tasks

Behavioural Video Analysis

In addition to the manual coding, video recordings of the sessions were analyzed using the Microsoft API[2] that recognizes and outputs scores for 8 emotions (happiness, anger, sadness, fear, contempt, surprise, disgust and neutral). We wanted to explore the possibility of deriving emotions non-manually and integrating it with the rest of the data from the physiological sensors. The Emotion API returned the results in a .json file, which we parsed and combined using Tableau[3]. Microsoft Excel was used to convert 'ticks' to time stamps. In order to ascertain that the emotions were recognized accurately, the video recordings were independently coded by two researchers who have experience working with children. In addition to coding for emotion, the coders also coded every performance trial for facial expression, head movements, body language, comments made and other overt behaviours. The coded behaviour was first organized into those for correct responses and incorrect responses. These were then further categorized into emotion, response latency (time taken to respond after a stimulus was presented), head movements, utterances, eye gazing and other overt behavioural signs [15] (Table 1).

We illustrate three sample outputs from the Emotion API. Figure 9a shows the breakdown of emotions over time for number series task from participant P8. As the task involves pattern recognition, there are a lot of surprise peaks as new stimuli are presented. Similarly, Figure 9b shows the emotions for participant P1 for the DCCS Mixed block. Since the difficulty is mixed across trials, there are a lot of surprise peaks as she switches between conditions imposed by the task but there are a lot of happiness peaks as she gets her answer right and makes comments throughout the session. Similarly, participant P13 (Figure 9c) appears quite calm overtly as she attempts the verbal attention task. The API is able to detect micro-expressions of sadness that increases as the task difficulty increases.

The Emotion API and manual coding of emotions were in-sync with each other. While such behavioural analysis reveals insights into the course of emotions as the participants faced different tasks with different difficulties, they alone do not offer insights into whether the child displayed an emotion in the presence of cognitive load.

DISCUSSION

Putting Them All Together: Triangulating Physiological, Performance and Behavioural Measures

We observed that while every measure offers a different perspective to the mental effort, they do not offer an entire picture of the individual or the process when used alone. For example, finding a one-to-one mapping of GSR to an emotional / psychological response is impossible as GSR can be high for positive and negative valence emotions. However, when GSR is evaluated in conjunction with HRV, it narrows down the list of possible emotions. For example, orienting, startle and defensive responses both elicit a high GSR. However, only startle and defensive responses are accompanied by increased heart rate [6]. Hedman [21] calls for a thick psychophysiological approach to understand events in the world using quantitative measures, external influences that may cause a physiological response and internal influence that refers to the meaning of that measure. In his studies, he uses video recordings in conjunction with skin conductance responses. In our study, we triangulated the physiological measure and performance data with our observations to explore if such an approach would offer us more holistic insights to the cognitive-affective state of the children. In the following two paragraphs, we describe two exemplary cases.

Case 1

Figure 10a shows the response of participant P13 as she went through the Antonyms sub-task of the JW. The antonyms sub-task taps onto the recall / retrieval of previously stored knowledge from long-term memory. Thus, the mental effort on part of the child is concerned with retrieval of pre-existing

[2] https://azure.microsoft.com/en-us/services/cognitive-services/emotion/
[3] https://www.tableau.com

Feature	Correct Response	Incorrect Response
Emotion	Happy, calm, neutral, confident	Sad, frustrated, irritated, bored, confused, anxious, curious
Response Latency	Fast, occasional pauses	Filled pauses "Oh no", "How much more?", "Can we play another game?", "When can I go?"
Head movements	Gentle leaning in	Head tilt towards floor lay head down on table looking away repeated head shakes
Postural Change	Straight and alert sometimes casual	Rigidity move to edge of seat standing up leaning all the way back pressing palms against table
Gazing	Looking towards experimenter for affirmation, good eye contact	Looking to experimenter for affirmation, gazing away, looking elsewhere as an attempt to disengage from stress
Response Time	Usually fast except when child tried justify an answer	Slow and laboured sometimes

Table 1. Manual Video Coding of Behaviour

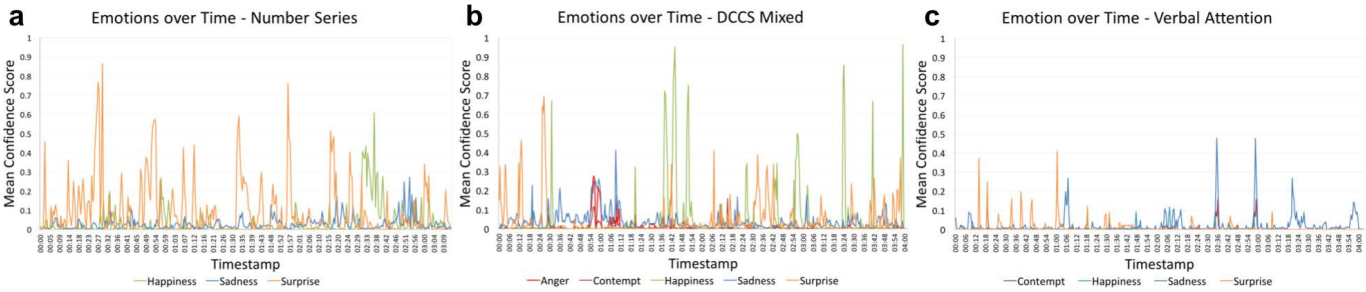

Figure 9. Graphs of Emotion over Time for: a) Number Series, b) DCCS Mixed and c) Verbal Attention. Analysed for three different participants.

knowledge. The green shaded areas with the item numbers on top denote the correct responses while the area in red denoted the incorrect responses. We can see that neither all correct responses nor incorrect responses share the same physiological and emotional characteristics. In addition, P13 had no SCRs (every change in the skin conductance level was less than 0.01 microsiemens). The first high peak of surprise at around 15 seconds of the procedure is characterized by surprise accompanied by a high LF component of HRV, indicating mental load. This may have been due to processing instructions and the novelty of the task. The LF/HF ratio is also high, indicating a higher sympathetic activation and cognitive stress [35]. However, as she gets familiar with the task there is still an element of surprise around item 10-11 but we can notice that the mental effort / load imposed by the task has reduced as shown by the LF/HF ratio and reduced LF power. As she gets comfortable with the task and has a spate of correct response, she demonstrates happiness and an even lower LF power and LF/HF ratio. It is here that the emotions start changing towards sadness. When she faces a difficult question to which she is unsure of the answer, she demonstrates sadness and a very small proportion of contempt with LF of 1407, HF of 1195 and LF/HF of 1.17. Following this, there are more periods of sadness but they are characterized by different heart rate measures. For instance, the sadness at Item 15 has a high LF, indicating higher cognitive load, and an attempt to think and solve the question. Towards the flag end of the test, even though the emotion is still sadness, she seems to have given up which is also reflected in the reduced LF power as well as sympathetic activity that highlight that she is not stressed or even exerting much mental effort anymore. The presence of HRV measures compensated for the lack of GSR values. This emphasizes the need to triangulate the measures and understand them in light of the performance, the emotions and approach adopted by the child. This also finds affirmation with her body posture that seems to be more alert with a gentle leaning in and then becoming rigid as test items become more challenging. Towards the end, her posture becomes more relaxed as she realizes that she does not seem to know the answers to the questions anymore. By looking at this combined representation, we get a much clearer idea of what P13 went through during the procedure.

Case 2

Now consider participant P8 (Figure 10b) as he attempted the Number Series sub task of JW. The number series task tests the working memory and pattern recognition of the participant. As the task progresses, it requires the user to hold more information in their working memory as they process and derive the number that should follow the series presented. The green and the red shaded regions represent correct and incorrect responses respectively. Unlike P13, P8 demonstrates SCRs throughout. He starts off with a very overt expression of surprise characterized by a high LF/HF ratio that again can be attributed to the excitement that accompanies the onset of a new task. This is corroborated by presence of a strong SCR which is also an indicator of arousal and excitement. Since no other emotion such as fear has a high value, one can ascertain

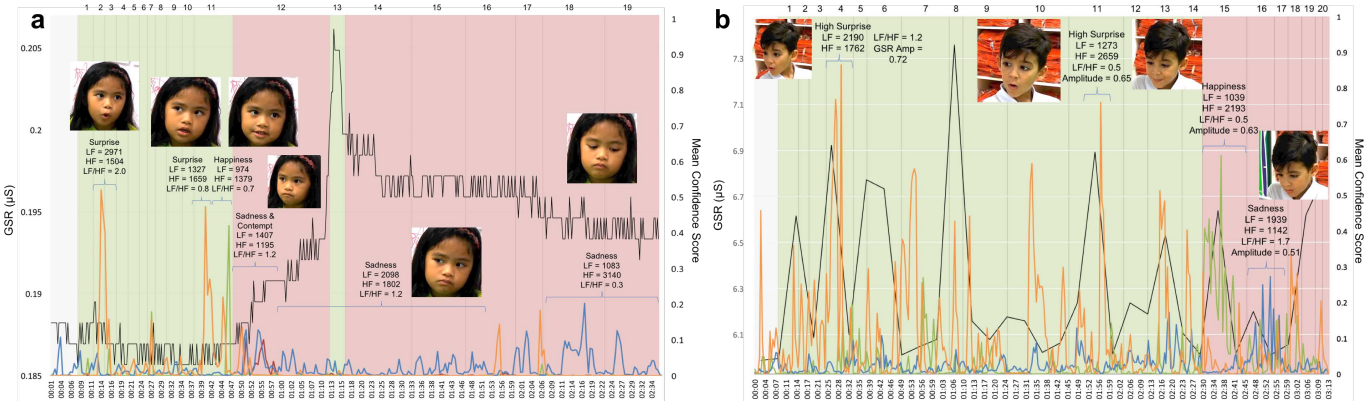

Figure 10. Triangulation of measures (—- representing GSR, other colours representing emotions and HRV values annotated. Orange line represents surprise, green line represents happiness and blue line represents sadness. The green and the red shaded regions represent correct and incorrect responses respectively.) for: a) Participant 13 during the Antonyms task and b) Participant 8 during the Number Series task

that this surprise maybe more of excitement. Somewhere midway into the task, there is another surprise event that is also followed by an almost equal SCR like the previous one. However, the HRV LF and LF/HF measures which are an indicator of cognitive stress have now dropped lower. Once P8 reaches the incorrect response region, there is an event of "happiness" right at the start that may be attributed to having had a series of correct responses or being unaware of the first incorrect response he makes. This is to some extent corroborated by the HRV measures that still show very low cognitive stress on part of the child. The SCR remains almost the same. However, towards the end of the tasks after a series of incorrect / no responses, he seems less sure of his answers and the emotion of sadness is quite strong. That he is aware of his wrong answers and finds the test items difficult is shown by the increase in the LF and LF/HF measures of HRV. SCR is present as well albeit slightly lower in amplitude. Expression of emotions is highly subjective and also varies across emotions. While the expression of surprise in P8 is very obvious, P13's surprise is not as overt. But, P13's expression of sadness is more marked than that of P8. The emotion API is able to draw these emotions out to a good extent and they correspond well with the physiological measures.

No matter how efficiently an emotion is recognized, the information of mental effort is important to understand the child's approach according to the task difficulty. Therefore, tagging behavioural events during, and even before and after the physiological responses may facilitate a better understanding of the child's state. If a child approached a difficult problem with curiosity and happiness in spite of experiencing cognitive stress, then it is exploratory and needs to be encouraged. But if the child approaches a problem with sadness and shows a high cognitive stress, there may be a need for some feedback / intervention. Such insights are best attained by triangulating measures from different sources.

Implications of Triangulation on Learning and Usability Evaluations

Learning and Pedagogy

As seen from the above examples, the triangulation of findings from different findings contributed to a better picture of the child's cognitive-affective state. If one were to apply such a framework in classroom, whereby, such responses are collected from students as the teacher / facilitator goes on with their lessons or as children perform tasks in the classrooms, it offers a multifaceted understanding of whether the child is exerting mental effort. If a child's SCRs are high in number or are of high amplitude, with a higher LF component of HRV, but demonstrates a more curious or engrossed look, it may signal that the child is exploring or trying to understand a problem. Now if one were to contrast this with that of a child who exhibits the same SCRs and HRV values but has a sad expression or frustrated expression, then, it shows that the child is probably finding the content too challenging. The objective values are still needed because not every child has overt expressions and sometimes, an expression may go unnoticed. But with a big change in the objective measures accompanied by the observations, it becomes easier to identify some of the key moments. At this point, as deemed appropriate by the teacher, the child may need some intervention in the form of feedback or a re-evaluation of the pedagogy on part of the teacher. Of course, once an intervention or remedial action is implemented, the same measures may offer an insight into whether this was effective at all.

Usability Evaluations with Kindergartners

When conducting usability evaluations with children, especially in the Kindergarten age group, it has been shown that verbalized responses and surveys maybe tricky as they are confounded by the developing language abilities as well as the child's need to please the experimenter. Many behavioural responses such as yawns, sighs, turning away, frowns are more reliable indicators than the ratings / verbal feedback. Further, depending on the goals, the experimenter wants to understand what are the aspects that are easy to use, interesting enough to hold the child's interest, the parts that are hard to understand, the aspects that bore the child, the points where an adult intervention helped and so on. For example, if when testing a construction toy or an interactive multimedia game, the child displays a neutral / almost bored expression complemented by lack of SCRs and HRV and shows no difficulty in trying the product, it may suggest that it is within comfortable limits. Depending on what the product aims to accomplish, it is up to the designer to think if they want to make this more exciting, add more challenges and get the children to explore more and with excitement. On the other hand, if this is a platform to learn new content, then navigating the platform and getting used to it should be done with ease as indicated by low SCRs, low LF HRV and a general calm / neutral expression. Or, if we were to imagine a child tester displaying mild frustration with high SCRs and high LF HRV, the experimenter could more closely evaluate the point where this occurred and think of what aspect of the interaction / product feature may have brought about this response - did the toy have too many instructions or were there too many elements to remember (taxing working memory) or in the case of a website for children, did the child have a lot of distracting features like colours and cartoon typography (affecting inhibition and making it hard to focus) or did it require them to constantly shift between different features (taxing flexibility)? If data from different sources are triangulated, they may offer the experimenter a better understanding of the child's cognitive-affective state as they go through the usability testing. This coupled with what the child responded verbally, rated on a response scale may point to some pain points or good aspects of a design. This can aid the experimenter / designer to closely evaluate the causes and possible ways to rectify them and re-evaluate the new design.

Other Takeaways for Implementing such Triangulated Frameworks

Tasks Design

We adopted a combination of tasks based on two main aspects: a) Time allowed for task completion and, b) Organization of challenges in the tasks.

Time taken to complete a task has an effect on the measurements. For example, a very short task may not capture the

fluctuation in the physiological parameters well. A very long task on the other hand may bore or tire the child, thereby discouraging him from even continuing participation. Using a combination of short and longer tasks may best mimic learning in real-life situations.

Tasks can also be categorized as being consistently easy at the start and increasing in difficulty towards later parts or have a mix of easy and difficult items interspersed. Having mixed difficulty tasks may give a good insight into whether the measures are truly responsive to randomly occurring difficulties / cognitive load and not just build up over time. In our case, the JW tests were arranged in ascending order and the EF tasks were interspersed.

Using the Appropriate Instrumentation

The choice of the wearable depends on the age group, task and the site of testing. When electrodes are placed on fingers, and when an individual moves their knuckles, it could produce artifacts [21]. Therefore, for Kindergarteners, we experienced that wearables with minimum instrumentation and distracting buttons or wires work best.

Ideally, if one wearable can accommodate most sensors, it would cause least distraction and also reduce setup time. However, some of these wearables like the E4 may need some modification to facilitate best contact with the sensors. Placement of sensors is important and one may need to test a few sites before getting a site that is sensitive, least obtrusive to the task at hand and comfortable for the wearer. For us, this site turned out to be the wrist, given the E4 affordances as well as ease of access. We did not use gaze-tracking or pupillometry measures as some tasks require the participant to bend their head as they work on the computers. Despite a lot of research on using EEGs as a tool for cognitive load, we refrained from using them owing to the set up and discomfort it would cause to a Kindergartner as well as the difficulty of obtaining measurements due to the child's movements during the tasks. We did not use the fewer channel EEGs even though they involve less set-up as they were still bulky and not offer very good resolution. Furthermore, we also wanted to use instruments that would be relatively easier to employ in a real-classroom or play context. Even though E4 does not exactly offer the highest resolution for skin conductance as compared to Shimmer (both of which were tested in the Pilot), E4 was easier to use and less obstructive / intrusive. This was also true for collecting heart rate variability. Along with the fact that the PPG sensor was also in the same watch, it provided fairly accurate data for the situation and physical involvement in the activities used in the study. If there was a lot of movement and intense physical play / learning, it would be a challenge. For older groups of children, there may be more flexibility in the choice of instrumentation with the luxury of choosing a device with higher resolution even in presence of physical movements. With these constraints, using video recording to track eye movements, gaze direction and facial expressions with body language, along with the physiological data from a wearable sensor offered a more complete behavioural picture with minimum intrusion. Having said that, we acknowledge that collecting any form of physiological data is invasive to some extent and needs to be exercised with consent and caution.

Procedure

Any procedure with children, especially those that look at emotional / physiological responses must begin with rapport building and data must be collected from a very familiar environment. In this study, the experimenter spent a week with the participants in their classroom in art and play activities.

Another major consideration in the procedure is the need for baselines. If the participants undergo different tasks, it is recommended that a baseline be established before every task to estimate change in physiological measures. After ascertaining that key press did not create any artifacts, we went ahead with a resting baseline before every task.

In addition, baselines change over time and across tasks. If the baseline has been measured at the start of a session before Task 1, the baseline before Task 2 will be the GSR from Task 1. If the participant's GSR has already reached the limit, then even though Task 2 imposes a cognitive load, no observable SCR may be seen. Therefore, it advisable to record resting baselines before every task. It is recommended that the entire procedure be conducted under controlled ambient conditions, as room temperature and humidity may influence the physiological and to some extend emotional responses.

Analysis

One of the most important criteria for analysis is to normalize the data to overcome inter-subject variability and analyse the difference between baseline measures and the task measures when comparing as a group. However, a true triangulation happens only when every subject's data is individually analysed and all measures are studied in relation to each other as shown in Figure 10.

Selecting the section for analysis is also important. Usually, for GSR, one determines a minimum threshold and analyses responses above the threshold. If there are too many SCRs, an alternative way would be to look at the top 10% or top 20 SCRs [21]. However, analysing HRV may require segmenting the data into fractions, and run a time and frequency domain analysis on them. Such an analysis would help detect the part of the task that contributed to cognitive stress if any.

FUTURE WORK

As subsequent step to this study, we wish to explore a wider variety of cognitive tasks that test a wider range of cognitive skills. We have also commenced a series of studies to explore the application of this framework to understand learning of various concepts through different pedagogical approaches in real-life contexts as well as supplementing evaluations of learning and game content using this triangulated approach.

CONCLUSION

Current methods of learning assessment in Kindergartens follow lesson delivery and are limited in their ability to capture the true cognitive-affective state of the learner during the process. We explored the feasibility of obtaining direct measures of cognitive load using physiological sensors from Kindergarteners and then used observational data to make sense of

the physiological measurements. We found potential GSR and HRV markers of cognitive load that are applicable to Kindergarteners. While objective measures help better understand the extent of cognitive load, they are always accompanied by emotions. Furthermore, emotions determine how a child faces a task. Therefore, tagging behavioural events during, and even before and after the physiological responses may facilitate a better understanding of the child's state. By triangulating these markers with our observations, we were able to better explain how the child perceived the cognitive load, thereby enabling us to differentiate the actual state in spite of two almost similar physiological measures. While we triangulated the measures for Kindergarteners, we believe that such an approach can be applied across age groups for learning and task performance. Given that there is a rapid proliferation of interactive educational and play applications, designing a learning situation that reflects a child's cognitive-affective state by acknowledging it and responding to it effectively would create a more meaningful and enriching interactive learning experience. Furthermore, collecting a child's state during interaction with the application, could reveal insights into the application itself. In this direction, we are exploring learning behaviours as Kindergarteners learn concepts over longer time frames across pedagogical approaches and different interactive media. We believe that such an understanding paves way for designing pedagogies, learning tools and other adaptive learning interfaces that are responsive to the learner's cognitive and affective states.

SELECTION AND PARTICIPATION OF CHILDREN

18 children, aged 4-7, were recruited from a Kindergarten in a middle-class neighborhood. Prior to the study, ethical approval was obtained from the Institutional Review Board (IRB). Children as well as their parents were told about the aims of the research and signed a form giving their consent to their data being used.

ACKNOWLEDGMENTS

We would like to thank the participants, their parents and teachers for being so supportive throughout this study. We want to thank Niha who helped with the video coding and transcription. We would also like to thank Shanaka Ransiri who helped with setting up the Empatica E4 and building a platform to collect data locally at the place of testing.

REFERENCES

1. Carole Ames and Jennifer Archer. 1988. Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of educational psychology* 80, 3 (1988), 260.
2. Mathias Benedek and Christian Kaernbach. 2010. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods* 190, 1 (2010), 80–91.
3. Frédéric Bousefsaf, Choubeila Maaoui, and Alain Pruski. 2013. Remote assessment of the heart rate variability to detect mental stress. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*. IEEE, 348–351.
4. Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* 49 (2013), 1017–1034.
5. Roland Brunken, Jan L Plass, and Detlev Leutner. 2003. Direct measurement of cognitive load in multimedia learning. *Educational psychologist* 38, 1 (2003), 53–61.
6. John T Cacioppo, Louis G Tassinary, and Gary Berntson. 2007. *Handbook of psychophysiology*. Cambridge University Press.
7. William H Clune and Paula A White. 2008. Policy Effectiveness of Interim Assessments in Providence Public Schools. WCER Working Paper No. 2008-10. *Wisconsin Center for Education Research (NJ1)* (2008).
8. Mihaly Csikszentmihalyi. 1997. Finding flow. (1997).
9. Ton De Jong. 2010. Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional science* 38, 2 (2010), 105–134.
10. Angel De Vicente and Helen Pain. 1998. Motivation diagnosis in intelligent tutoring systems. In *Intelligent Tutoring Systems*. Springer, 86–95.
11. Allison Druin. 2002. The role of children in the design of new technology. *Behaviour and information technology* 21, 1 (2002), 1–25.
12. Carol S Dweck. 1986. Motivational processes affecting learning. *American psychologist* 41, 10 (1986), 1040.
13. Carol S Dweck and Ellen L Leggett. 1988. A social-cognitive approach to motivation and personality. *Psychological review* 95, 2 (1988), 256.
14. Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.
15. Sharon Eldar and Y Bar-Haim. 2010. Neural plasticity in response to attention training in anxiety. *Psychological medicine* 40, 4 (2010), 667–677.
16. Johan Engström, Emma Johansson, and Joakim Östlund. 2005. Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 8, 2 (2005), 97–120.
17. Stephen H Fairclough. 2008. Fundamentals of physiological computing. *Interacting with computers* 21, 1-2 (2008), 133–145.
18. Eija Ferreira, Denzil Ferreira, SeungJun Kim, Pekka Siirtola, Juha Röning, Jodi F Forlizzi, and Anind K Dey. 2014. Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. In *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2014 IEEE Symposium on*. IEEE, 39–48.

19. Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. 2014. Empatica E3? A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*. IEEE, 39–42.
20. Béatrice Susanne Hasler, Bernd Kersten, and John Sweller. 2007. Learner control, cognitive load and instructional animation. *Applied cognitive psychology* 21, 6 (2007), 713–729.
21. Elliott B Hedman. 2014. *Thick psychophysiology for empathic design*. Ph.D. Dissertation. Massachusetts Institute of Technology.
22. Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology* 92, 1-2 (2004), 84–89.
23. Curtis S Ikehara and Martha E Crosby. 2005. Assessing cognitive load with physiological sensors. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. IEEE, 295a–295a.
24. Ashish Kapoor, Selene Mota, Rosalind W Picard, and others. 2001. Towards a learning companion that recognizes affect. In *AAAI Fall symposium*. 2–4.
25. John M Keller. 1983. Motivational design of instruction. *Instructional design theories and models: An overview of their current status* 1, 1983 (1983), 383–434.
26. John M Keller. 1987. Strategies for stimulating the motivation to learn. *Performance Improvement* 26, 8 (1987), 1–7.
27. John M Keller and BH Keller. 1989. Motivational delivery checklist. *Florida State University* (1989).
28. Paul A Kirschner. 2002. Cognitive load theory: Implications of cognitive load theory on the design of learning. (2002).
29. Barry Kort, Rob Reilly, and Rosalind W Picard. 2001. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on*. IEEE, 43–46.
30. Mark R Lepper and Ruth W Chabay. 1988. Socializing the intelligent tutor: Bringing empathy to computer tutors. In *Learning issues for intelligent tutoring systems*. Springer, 242–257.
31. Jimmie Leppink and Angelique van den Heuvel. 2015. The evolution of cognitive load theory and its application to medical education. *Perspectives on medical education* 4, 3 (2015), 119–127.
32. George Mandler. 1984. *Mind and body: Psychology of emotion and stress*. WW Norton.
33. Yukihiro Matsubara and Mitsuo Nagamachi. 1996. Motivation system and human model for intelligent tutoring. In *Intelligent Tutoring Systems*. Springer, 139–147.
34. Richard E Mayer, Mary Hegarty, Sarah Mayer, and Julie Campbell. 2005. When static media promote active learning: Annotated illustrations versus narrated animations in multimedia instruction. *Journal of Experimental Psychology: Applied* 11, 4 (2005), 256.
35. Daniel J McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W Picard. 2016. Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4000–4004.
36. Ziev B Moses, Linda J Luecken, and James C Eason. 2007. Measuring task-related changes in heart rate variability. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. IEEE, 644–647.
37. Ebrahim Najafpour, Naser Asl-Aminabadi, Sara Nuroloyuni, Zahra Jamali, and Sajjad Shirazi. 2017. Can galvanic skin conductance be used as an objective indicator of children?s anxiety in the dental setting? *Journal of Clinical and Experimental Dentistry* 9, 3 (2017), e377.
38. Arturo Nakasone, Helmut Prendinger, and Mitsuru Ishizuka. 2005. Emotion recognition from electromyography and skin conductance. In *Proc. of the 5th International Workshop on Biosignal Interpretation*. 219–222.
39. Donald A Norman. 1980. Twelve issues for cognitive science. *Cognitive science* 4, 1 (1980), 1–32.
40. Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* 38, 1 (2003), 63–71.
41. Fred GWC Paas and Jeroen JG Van Merriënboer. 1994. Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of educational psychology* 86, 1 (1994), 122.
42. Rosalind W Picard, Seymour Papert, Walter Bender, Bruce Blumberg, Cynthia Breazeal, David Cavallo, Tod Machover, Mitchel Resnick, Deb Roy, and Carol Strohecker. 2004. Affective learning?a manifesto. *BT technology journal* 22, 4 (2004), 253–269.
43. Katherine Picho and Anthony R Artino Jr. 2016. 7 deadly sins in educational research. (2016).
44. Fredrick D Pociask and Gary R Morrison. 2008. Controlling split attention and redundancy in physical therapy instruction. *Educational Technology Research and Development* 56, 4 (2008), 379–399.

45. Jocelyne A Posthumus, KBE Böcker, MAJ Raaijmakers, H Van Engeland, and W Matthys. 2009. Heart rate and skin conductance in four-year-old children with aggressive behavior. *Biological Psychology* 82, 2 (2009), 164–168.
46. Janet C Read and Stuart MacFarlane. 2006. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 conference on Interaction design and children*. ACM, 81–88.
47. Mike Scaife and Yvonne Rogers. 2001. Informing the design of a virtual environment to support learning in children. *International Journal of Human-Computer Studies* 55, 2 (2001), 115–143.
48. Walter Schneider, Amy Eschman, and Anthony Zuccolotto. 2002. *E-Prime: User's guide*. Psychology Software Incorporated.
49. Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. 2010. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on information technology in biomedicine* 14, 2 (2010), 410–417.
50. Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 extended abstracts on Human factors in computing systems*. ACM, 2651–2656.
51. Herbert A Simon. 1967. Motivational and emotional controls of cognition. *Psychological review* 74, 1 (1967), 29.
52. J Richard Simon. 1969. Reactions toward the source of stimulation. *Journal of experimental psychology* 81, 1 (1969), 174.
53. John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.
54. Mika P Tarvainen, Juha-Pekka Niskanen, Jukka A Lipponen, Perttu O Ranta-Aho, and Pasi A Karjalainen. 2014. Kubios HRV–heart rate variability analysis software. *Computer methods and programs in biomedicine* 113, 1 (2014), 210–220.
55. Priscilla L Vail. 1994. *Emotion: The on/off switch for learning*. Modern Learning Press.
56. VH Vroom. 1964. Work and Motivation: Wiley New York. (1964).
57. Denise Whitelock and Eileen Scanlon. 1996. Motivation, media and motion: Reviewing a computer supported collaborative learning experience. *CITE REPORT* (1996).
58. Benjamin N York and Susanna Loeb. 2014. *One step at a time: The effects of an early literacy text messaging program for parents of preschoolers*. Technical Report. National Bureau of Economic Research.
59. Beste F Yuksel, Kurt B Oleson, Lane Harrison, Evan M Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5372–5384.
60. Philip David Zelazo. 2006. The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature protocols* 1, 1 (2006), 297.