Measuring Cognitive Load and Insight: A Methodology Exemplified in a Virtual Reality Learning Context

Jonny Collins*
Information Science
University of Otago

Holger Regenbrecht†
Information Science
University of Otago

Tobias Langlotz ‡
Information Science
University of Otago

Yekta Said Can§
Computer Engineering
Department
Bogazici University

Cem Ersoy¶
Computer Engineering
Department
Bogazici University

Russell Butson‖
Higher Education
Development Center
University of Otago

ABSTRACT

Recent improvements of Virtual Reality (VR) technology have enabled researchers to investigate the benefits VR may provide for various domains such as health, entertainment, training, and education. A significant proportion of VR system evaluations rely on perception-based measures such as user pre- and post-questionnaires and interviews. While these self-reports provide valuable insights into users' perceptions of VR environments, recent developments in digital sensors and data collection techniques afford researchers access to measures of physiological response. This work explores the merits of physiological measures in the evaluation of emotional responses in virtual environments (ERVE). We include and place at the center of our ERVE methodology emotional response data by way of electrodermal activity and heart-rate detection which are analyzed in conjunction with event-driven data to derive further measures. In this paper, we present our ERVE methodology together with a case study within the context of VR-based learning in which we derive measures of cognitive load and moments of insight. We discuss our methodology, and its potential for use in many other application and research domains to provide more in-depth and objective analyses of experiences within VR.

Index Terms: Human-centered computing—Human computer interaction—Interaction paradigms—Virtual Reality; Applied computing—Education—Interactive learning environments; Human-centered computing—HCI theory, concepts and models

1 INTRODUCTION

How can we find out how a user is perceiving, experiencing, and emotionally responding to a virtual environment? *Virtual Reality* (VR) systems are commonly evaluated for various factors such as usability, cognitive workload, stress, simulator sickness, and the sense of presence, to name a few. These evaluations extract information about the performance of a given system, but also provide insight into the different aspects of users' experiences. There are three, commonly used forms of evaluation: (1) Self-report questionnaires are the most commonly used, subjective, quantitative measure for evaluating these phenomena and are presented to users after they have had exposure to a system [8, 28, 34, 58]. (2) Interviews are a subjective, qualitative form of evaluation often used to attain domain specific information [33, 54]. And, (3) Observational techniques provide an objective, qualitative form of measurement where experimenters watch or record an environment and users' actions, and behaviours within the environment. Observational measures are known to be used for studying the sense of presence in VEs [64] and are encouraged in some cases over the use of self-report questionnaires [60]. Physiological measurements provide us with an objective, quantitative measurement of users' physiological reactions to stimuli in VR environments. Examples of these measures are found in evaluations of VR-based anxiety and phobia treatment systems [76], and again, in studying the sense of presence [20].

Many of these forms of measurements, particularly self-report questionnaires and interviews, are a measure of users' perceptions of their experiences at certain instances in time, usually post-experience. While these are useful and valid measurements for assessing different phenomena, they rely on users' collective recollection of their experiences. Observational measurements are not relying on users' recollection as evaluators can see and record when specific behaviours emerge with respect to their presented stimuli, but what is missing from that analysis is a detailed measure of why behaviours occur. Physiological measurements can provide more depth by measuring the sub-conscious states (in addition to other factors) of a user throughout exposure.

In this paper we are using VR learning as our exemplary domain to explain and illustrate our methodology as VR has demonstrated high promise for application in the education domain. The research community has been able to apply VR within multiple topic spaces such as physics [78], chemistry [40, 49], biology [59], and geometry [65]. The flexibility, controllability, and more recently also the cost effective nature of immersive VR environments are leveraged to facilitate the learning experiences which lead to observed increases in learning outcomes [22, 47]. While these applications show promising results, learning outcomes, and the experiences that lead to them, are largely measured through self-report measures or examination style tests after a given exposure time [44]. These types of evaluations are a good indication of the potential efficacy of VR learning environments, however, the length and the process of the learning experience is largely overlooked. It is desirable for us in the research and development community to attain information regarding the continuous nature of users' learning experiences, as this can provide more detailed analyses and highlight relevant aspects of immersive VR learning applications.

Here, we propose a methodology which targets the elements of a user's experience in VR environments, exemplified in the context of immersive VR learning environments. Our methodology called *Emotional Responses in Virtual Environments* —ERVE, is primarily centred around users' physiological states which serves as a continuous measure of their experience. Physiological data is then combined with other forms of data (such as self-report or observational data) in a custom built analysis which can be used to

*e-mail: jonny.collins@otago.ac.nz
†e-mail: holger.regenbrecht@otago.ac.nz
‡e-mail: tobias.langlotz@otago.ac.nz
§e-mail: yekta.can@boun.edu.tr
¶e-mail: ersoy@boun.edu.tr
‖e-mail: russell.butson@otago.ac.nz

derive measures of various phenomena. In the case of our exemplary study, we are able to measure participants' cognitive load and the emergence of moments of insight in a VR context.

This work provides three primary contributions: 1) A detailed description of ERVE, a novel methodology to empirically detect emotional responses in virtual environments as a continuous measure of user experience, and analyze response data against event data to generate measures of varying types of phenomena. 2) A case study on learning in VEs demonstrating an implementation of the ERVE methodology. We effectively detect relevant emotional responses which are analyzed against naturally occurring events resulting in a measure of cognitive load and insight. 3) A critical discussion of results and applications that would benefit from measuring emotional responses and the application of ERVE to these scenarios.

2 Background

This work touches two categories of previous work that we will briefly introduce in the following. Firstly, we will discuss how Virtual Reality systems have been empirically investigated. We focus in particular on common measurements such as questionnaires but also discuss observations and their use in VR and the use of physical readings and their use when investigating VEs. Given the sheer number, we will only cover the most relevant measures together with some examples on how they have been used in VR. Secondly, we will introduce the reader to some key works in learning within VEs to provide context to our case study.

2.1 Measurement in VR

Questionnaires A common way for software developers to determine which features and interaction metaphors to integrate into their applications is through usability assessments [18]. A commonly used tool for usability testing is a questionnaire [9,36] which usually has a subject conduct tasks with a given application and after having completed the tasks they fill out a usability questionnaire. Other work has built on the concept of usability assessment by generating "user experience" assessment by questionnaire [39]. These usability assessments have also been used in the context of VR system development to evaluate immersive user interfaces [68]. Similarly, researchers commonly assess task workload by utilizing questionnaires. Here the NASA-TLX assessment tool [28] is one of the most well known and was developed to assess user workload for tasks ranging from simple cognitive tasks to more complex ones. Another example for assessing workload using questionnaires is the Cooper-Harper scale [77]. The sense of presence in VR environments is referred to as one of the, if not *the*, defining component of VR [54,66]. It is known as the subjective psychological sense of "being there" in the virtual environment. The most common technique for measuring presence is by questionnaire [42] which are, as with most other forms of questionnaires, applied after a given exposure time to assess a user's sense of presence in that environment.

Although questionnaires as a form of measurement have shown to be robust, overall measures of various phenomena, they are still a composite measure of users' recollections of their overall experience. There is a lot of potential insight to be gained from more continuous evaluation throughout users' exposure to VR systems. It would be possible to present VR users with questionnaires throughout an exposure, however in that case it requires the user to remove themselves from the VR environment to complete a questionnaire resulting in a complete break of the sense of presence. To solve this problem, Frommel et al. recently attempted to present questionnaires within VR environments to assess the users sense of presence throughout immersion and found that presence in the environment was retained when the questionnaires were presented virtually [23]. The issue with this technique is that if a user of a system is mentally invested in a task, i.e. a learning activity, and they are interrupted with a virtual questionnaire, their sense of presence may be retained, but their engagement with the environmental task will be broken. Therefore passive, continuous measures are still advantageous over prompting users' for their reports, particularly in certain scenarios.

Observational measures A less invasive way is to use observational measures which are less-commonly used than questionnaires due to the difficulty associated with first applying the technique, and then analyzing and applying the results. The research community has started to adopt measures that are more objective and are using them in conjunction with existing subjective measures [31]. Some presence work has used a technique called "Behaviour Presence Test in Threatening VEs" [43] which utilizes observation of users' reactions to environmental stimuli as one of its primary measurement tools. Other forms of observational measures include task results such as completion times and scores. These measures are common across many domains such as health (rehabilitation) [12], immersion/presence [6], and industry applications [25]. They are also often used to validate system usability. However, observational studies cannot fully capture users' emotional states and require experience in running these studies as they will otherwise risk a wrong interpretation.

Physiological measurement Different physiological data have been collected for evaluating various phenomena within VR applications including electrodermal activity (EDA) [2,21,76], heart rate (HR) [63,76], brain activity (functional magnetic resonance imaging (fMRI) [30]), and heart activity (electrocardiography (ECG) [7,27,62], electroencephalograph (EEG) [41]). Measuring emotional activity by monitoring physiological states is not a novel idea [10,35,37,53]. A common measure of emotional states used within various contexts is that of EDA, otherwise known as skin conductance. This measures the sympathetic nervous system (SNS) and is a sensitive index of sympathetic arousal which is integrated with emotional and cognitive states [14]. This measure has prior application in VR environments in the context of anxiety and therapy treatment [17,20,75,76], emotional replication studies [21], behavioural studies [24], and presence [16,45,46,51]. These measures have been very successful for measuring strong responses to those stimuli in VEs [45,48,51], however, what is lacking is the long term, continuous evaluation using physiological measurement in situations that are not designed to evoke responses at designated points in time. Saha et al. report on a study focused on measuring the emergence of negative emotional responses induced within a VR environment and are able to conclude "techno-stress" as a contributing factor to negative reaction [55]. We measure emotional responses through physiological measurements which are the central component of the methodology we report on in this work. The emotional responses we measure are detectable changes in the SNS and are known to be indicative of emotional activity. We use the physiological data representative of emotional activity together with other types of data (psychological reporting, or observational outcomes) to achieve measurements of different phenomenon in a VR environment.

Sanchez-Vives et al. argues that VR experiences are relevant for the neuroscience space by presenting multiple studies which utilize physiological measurements [56]. The prior work of Guger et al. and Slater et al. utilize physiological measures against events in VEs such as intentional breaks in presence, or stimuli from virtual characters [27,61]. The similarity between theirs and our work is that we both measure physiology about events throughout an exposure. In their approach, they take either averaged measures about instances in time, or averaged measures over a whole exposure and measure against a baseline. In our case, we conduct an in-depth analyses on our EDA data separating the event-driven (phasic) and long-term (tonic) components of EDA. After we extract features from both the EDA and heart rate variability (HRV) data, we use machine learning algorithms with event data collected from the exposure. With our methodology, we are focused not only on significant events within a VE, but we also consider subtle changes in the emotional activity

of users. With this approach we provide a continuous measure of a user's state throughout which we can measure against "naturally occurring" events (i.e. events not triggered by experimenters).

We identify the need to measure more natural situations in which emotional responses emerge as a result of users' own experiences with the environment, i.e. not instances in time designated by evaluators, but instances that can occur at any point according to the natural processes of the user. Our ERVE methodology allows for the application of physiological measures in a continuous transferable manner which facilitates its use in more natural situations. Rather than evaluating against specific points in time that are determined by the delivery of some arousing stimuli, we measure against other measurable aspects of the users' experiences (i.e. achievements or self-reported moments). This way researchers can evaluate user's emotional response to, and engagement with, more general and less provocative scenarios. One such scenario is a VR learning environment where a user is given a concept to learn over a period of time. There are no immediate stimuli intended to provoke a user in such a space, yet emotional response to that environment is as relevant as in any other VR context.

2.2 Measuring Learning Experiences in VR

We implement our ERVE methodology in a case study which lies in the context of VR-based learning. Below we give a brief overview of literature related to the space of educational VR research, and the particular type of learning we are investigating in our case study —insight learning.

A core imperative of education is to prepare students with the knowledge and skills they will need to be successful in meeting their academic and professional goals. Central to this imperative, is the development of problem-solving skills, in particular the ability to formulate inferences from observations. In a time of change and innovation, traditional approaches to learning are coming under increasing scrutiny from contemporary, disruptive thinking. For example, the work of Perkins [52], Ohlsson [50] and Weisberg [74], promote breakthrough or insight thinking as a legitimate solution to solving the growing landscape of wicked problems. At the heart of breakthrough thinking is the capability to secure insight through creative, engaged thinking. This approach relies heavily on a creative process rather than on analytic procedures and typically results in a higher level of engagement.

The use of VR in education is not new [29, 47], although many focus on replicating traditional approaches to education. Existing examples of education-based VR implementations cover topics such as high-school chemistry and physics [3, 49, 78]. The quality of VR learning environments is typically measured through questionnaires and/or interviews [3]. While these methods can help uncover a learners perception of their experience, they can be prone to inaccuracies due to the reliance on memory or post-event recollection. Another common method includes the use of tests to assess learner's progress (usually relative to a standard control group) [67]. While these methods can provide initial results for efficacy, they generate rigid outcomes that aren't necessarily representative of the true impact that VR environments provide. Many assessments are over short periods with small exposure times which makes it further difficult to accurately evaluate the educational potential of VR systems. This is reflected in the current research outcomes where little consistency is ascertained, especially in terms of users learning experiences and therefore the supposed advantages for learning provided by the VR systems [22].

To exemplify the methodology we describe in our work we have produced a VR learning environment which is based on a problem-solving approach. We use this environment to investigate the concept of insight learning, or Aha! moments, which are sudden moments of clarity with respect to some problem or concept [70]. Such moments are considered rare in the education space however if the conditions for breakthrough thinking are met or facilitated, then insight moments are more likely to emerge. As stated earlier, breakthrough or insight oriented approaches are on the rise, so investigating VR as a potential medium for such approaches to learning is timely. The phenomenon of the insight moment has been studied for nearly a century through behavioural methods, but has more recently been investigated using reported and physiological measurements [15, 38, 57]. We wish to investigate the emergence of Aha! moments in an exemplar study and our problem-based VR environment is designed and built to facilitate such a process. Reported measures of users' Aha! moments have recently been applied in which participants are asked to report in the very moment they experience such a moment of insight [15]. Upon pressing a button, they are presented several questions on a screen to assess the various factors of the Aha! moment such as suddenness, certainty, and relief. We opt for a similar approach, however presenting users with questionnaire items mid-immersion in a VR environment can be detrimental to a users involvement with the task at hand. We discussed the possibility to use integrated questionnaires, however this does not solve the problem of breaking task engagement. Aha! moments, or insight learning, has yet to be investigated in the context of VR. Due to many of its characteristics, VR could provide a viable platform for this approach to learning.

3 Emotional Responses in Virtual Environments Methodology

The *Emotional Responses in Virtual Environments* (ERVE) methodology is the process of identifying and collecting data relevant to a user's emotional responses within a VR environment, and performing an analysis such that we can isolate significant emotional events relative to environmental stimuli (of differing extremities). At the core of our methodology sits the measure of objective, physiologically measurable states of users' emotional responses. We call this the Sensed Reality dimension, and we augment this measure with two other data dimensions which we label as Reported Reality and Observed Reality. As discussed previously, current analyses of VR systems mainly rely on and utilize two of these dimensions, namely reported reality and observed reality, but not sensed reality.

The remainder of this section will introduce the conceptual overview highlighting the importance of the Sensed Reality domain and how it can be complemented by the other domains when investigating emotional responses. We follow this conceptual overview with a high level implementation that can be generalized to different application areas within VR. We later show how we applied the methodology in the specific context of learning in VR and more specifically identifying Aha! moments in VEs.

3.1 Measurement Dimensions

As we have mentioned, in VR evaluations the forms of data collection are most often reported, and observational measures. Users conduct a task in the VR environment and then after the experience they complete questionnaires. The quantitative results of the questionnaires are measured against the observed task results (completion times or scores for example) and an answer to the question is formed. We introduce below the conceptual components of our proposed ERVE methodology.

3.1.1 Sensed Reality Dimension

The sensed reality dimension of data is at the center of our methodology and refers to measured states of the body of a subject throughout an experience. It is entirely objective and it represents the subject's internal physiological state —something that is usually not consciously controlled by a subject. For example, when a subject is immersed in an experience and is presented with some form of stimuli, physiological measurements will likely measure the true impact of the event on the subject, even if the subject does not report

the event as significant later on. This makes physiological measures a very strong tool for evaluating emotional response and engagement in VR environments. This is the reason we place high priority on the implementation of such measures as the core aspect of VR analyses.

3.1.2 Observed Reality Dimension

Observed reality measures the resulting output from the subject interacting with the environmental stimuli. Data can come in multiple forms and can be appropriate for either qualitative or quantitative measurement. A common example of this data is filming of an experiment. There would be an experiment room with cameras in each corner filming the entire room including all actions and reactions (behaviours) of a subject and the changes in the environment. This is an example of data that requires qualitative analysis for any formal evaluation. Other forms of observational data include event records of anything considered a significant observation throughout an experiment. These can include measures of task performance, or decision points. VR environments are especially useful for collection of this form of data because we are able to retain control of the environment the user is immersed in.

3.1.3 Reported Reality Dimension

The reported reality dimension uses psychological measures which provide data representing one's perceptual reality. The most common means of measuring this dimension are questionnaires and interviews, each of which ask the user to reflect on their experience. They provide subjective data, however most of their value lies in the possibility to quantitatively analyze the data, if some form of numeric scales are used. The issue with questionnaires is that they are a measure of subject perception at a point in time. It relies on subject memory of (usually) multiple small experiences over a period of time. While the period of time is often short, users can have multiple different emotional responses to an environment and when answering a questionnaire after the entire experience, we are likely seeing a subject's internal average evaluation of the experience with respect to the system. Slater argues that questionnaires are a sub-optimal apparatus for measuring the presence construct and they developed a concept called RAVE (real actions in virtual environments) [60]. The argument is that only by observing the behaviour of subjects within VR environments can we make any robust deductions about whether a person is truly present in that environment. The RAVE approach to measuring presence has been since superseded by a concept called "respond as-if-real" (RAIR) [32] which explicitly includes physiological responses to stimuli in VEs. In addition, there are well known problems with questionnaires and interviews outside the realm of virtual reality, like experimenter bias, or unintentionally leading questions.

3.2 General Implementation of ERVE

We present our ERVE methodology in two primary steps (1) system design and data collection and (2) data analysis. We will describe in general the process for each of the reality measures described above. It should be noted that while the sensed reality dimension is central and necessary for ERVE, one or both of the supporting dimensions can be applied for successful application of ERVE. We have written up and provided access to a detailed set of instructions with particular focus on the analysis. The instructions describe each step of the analysis process with links to example sets of data and the code which is produced to perform parts of the analysis. The instructions can be found at http://www.hci.otago.ac.nz/research_erve.html.

3.2.1 System Design and Data Collection

The implementation of the ERVE methodology requires the VR system to be designed in a particular way. The system implementation is most important for the observed dimension of data described above, but can also encompass the reported dimension depending on experimenters' application space. Observed data includes any data that can be captured from the environment, users' behaviours and interactions, and the resulting outcomes of interactions with the environment. There are two ways a virtual environment and users behaviours can be captured. The virtual environment can be designed in such a way that it can be viewed from one or more perspectives. A virtual camera can be set up from a third person perspective, much like how a camera in the real world would be positioned (i.e. in the corner of a room). If required, such a virtual camera in VR could be animated to follow the relevant scenario space if necessary (exemplifying the advantages of VR environments). An alternative way could be to record the entirety of the VR space as it is acted or played out so it can later be replayed where an external viewer can immerse themselves within the replayed environment. If necessary, and if ethically appropriate, a real camera could be set up in the real-world space where the user is immersed in the VR space. These forms of measurement are qualitative and entirely objective and can yield much relevant information about users' experiences, however they are more difficult in terms of analysis.

The occurrence of events are a further form of measurement to be considered during system design. There are two forms of events that one might want to measure: 1) externally triggered events, 2) internally generated events. Externally triggered events are generated from the environment, usually triggered by an experimenter or they are pre-programmed, and are designed to induce some response from system users. Internally generated events are more naturally occurring, and are generated either as a result of a users interactions with an environment or by the user themselves as some form of report. For either of these types of event, the system should be designed to record timestamps of these moments which are considered to be significant. Furthermore, details specific to the type of event should also be stored. For instance, if a user is completing a task, details of that task should be stored such as how long it took, and how difficult it is. If a user is reporting on a thought or feeling, these should be distinguishable from one another through the recorded data.

3.2.2 Physiological Data Collection

With the sensed reality component at the core we begin by describing physiological data collection. Various measurement tools are available for collecting physiological data as was described in Sect. 2.1. One of the most appealing aspects of physiological measurement is its continuity, so it should be ensured that where this measure is applied, it can do so without interruption. A requirement for all dimensions used in the ERVE methodology is measurement of time, so the measurement device should record the data point in its corresponding unit, along with the current timestamp.

Again, one of the advantages afforded by VR systems is the ease of recording events within the environment. We described above different methods of measurement for observed data that are implemented within a system. Some reported measures can also be measured through the system. Examples of reported measures which are readily suited for application in ERVE are self-reports which occur and can be recorded in a continuous manner.

3.2.3 Data Analysis

Sensed measures are used as the fundamental measure of physiological responses to VEs which we infer emotional responses from. We describe first the treatment of the sensed data to prepare it for further analysis, then we describe the classification of the supporting dimensions of data to be measured against the physiological component.

The type of physiological measure selected for experimentation will determine how that data is specifically processed, and which tools are used. Sometimes raw data from physiological measures contain a lot of noise which makes a formal and robust analysis

very difficult, so noisy data points need to be accounted for. For the purpose of feature extraction we have to be able to identify data points (or sets of data points) from a dataset that are indicative of significant moments (keeping in mind that identified data points have corresponding timestamps). The values used for feature extraction are also specific to the type of physiological measure. Examples of preprocessing for noise reduction and feature extraction for physiological data are available for different datasets such as ECG [27], EEG [41], and EDA/GSR [69].

The advantage of both observed and reported data when collected as described earlier is that cleaning is often not required. The only step that may need to be taken into account is to ensure the timestamps are of the same format and timezone as the physiological data. Unix-based UTC timestamps are a common recording format for time.

The final step of the analysis is classification of data through machine learning algorithms. Classification toolkits such as the Weka Toolkit [19] can be used to classify observed and reported data sets as described above. For example, when a participant reports moments they exhibit a specific feeling such as surprise (reported measure). The duration of the study exposure can be divided into time segments, e.g. 10 second segments over a 10 minute study. We assign a binary classification system marking each segment as either a 1 if subjects experienced such a feeling in that 10 second segment, or a 0 if not. These markers are passed into one or more classifiers as labels and we can then test the features extracted from the physiological data against those labels.

This type of analysis method can yield valuable information about the emotional states of users in VEs while exposed to any form of stimuli be it extreme and provoking, mild and calming, or anywhere in between. We can potentially predict individual user's needs such as whether they require more or less stimuli, or whether a task is too difficult or too easy. A strength of this methodology is that it is not limited to a single context but rather it is relevant to any context where emotional engagement is key. It should be noted that the set of instructions we provide (mentioned above) describe an implementation using EDA and heart rate (HR) measurements, although, the same process would work by replacing the EDA/HR data preprocessing steps with the appropriate preprocessing requirements of other physiological data (i.e. EEG or ECG).

We have designed and conducted an exploratory study to demonstrate examples of measurement in each dimension, and include an analysis of two of the interactions - Sensed Reality with Observed Reality, and Sensed Reality with Reported Reality. The results we provide are intended to be demonstrative of the possible outcomes, and aren't intended to answer any hypothesis-driven questions—this paper is about the ERVE methodology and not about the specific study. The following sections describe our study context, the study procedure, and detail the forms of measurement we used in our ERVE methodology.

4 Study Context, Environment, and Subject Matter

To illustrate our ERVE methodology, we use a certain, particularly chosen learning scenario as our study context. This study serves as an exemplification of our methodology and hopefully helps researchers to understand, replicate, and apply ERVE to their research contexts.

Evaluating learning environments is difficult and complex. In particular maintaining objectivity in a context where each learner brings with him/her individually very different past experiences. Those past experiences will undoubtedly influence our evaluation of the learning experience. Therefore, we try to minimize those effects by providing a novel-as-possible context for the learner. This is implemented in two ways: (1) originality of the environment which users act in, and (2) originality of the subject matter that users are attempting to gain mastery over.

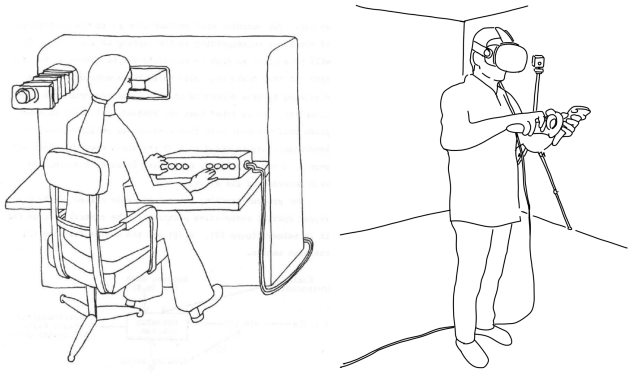

Figure 1: Overview over our experimental Virtual Reality learning system. (Left) The original hand drawn proof of concept implementation from [4] with permission for use granted from the Biological Computer Laboratory, University of Illinois, (Right) Our implementation of the original system using current state of the art fully immersive VR hardware.

For the originality of the subject matter, we utilize an interactive problem solving task allowing users to experience and manipulate objects in the fourth spatial dimension (i.e. a space where four axes lie mathematically perpendicular to each other). The system was initially proposed for the purpose of evaluating the process that learners go through when gaining mastery over novel concepts [4,73]. Very few people have knowledge of the theory behind the concept of four-dimensionality, and even fewer have ever interacted with it. This presents the opportunity of being able to give a problem to adults based on novel content, and monitoring their progress as they go through a learning process in an attempt to achieve an understanding or a grasp (German: begreifen) of the subject matter.

The original system allows users to experience a 4D construct called a hypercube (a four-dimensional cube) by rotating it in space. Figure Fig. 1 (left) depicts Arnold's hand drawn design plan which has a user looking into a 3D stereoscopic monitor and utilizing six analog dials on a board to manipulate the rotation of the hypercube around its six faces, which is updated on the monitors. This is the principle of the system —allowing a user to interactively experience a hypercube through rotations.

For most users, exposure to a virtual environment with a meaningful task is novel. In combination with the 4D hypercube representation we provide originality of the environment. One of the advantages of VR is the flexibility it provides which allows us to place users in unique locations or situations. If we place users in a different situation than usual, it can allow them also to think outside of their usual paradigms. We provide this element by presenting users with a minimal and unique VR environment containing the original subject matter. The minimal environment, without the distraction of other external stimuli, provides us with a useful scenario for evaluating emotional responses. We want to ensure the emotional responses we get are relative to users' tasks, and therefore, the minimal environment is way to control for this.

The following section describes in detail our second contribution: the exploratory study which implements our ERVE methodology in the context of a VR learning environment. Following that is a discussion on how the methodology is, or can be, useful.

5 Methodology by Example: An Exploratory Study

This study is originally designed to identify the emergence of learning trajectories that underpin the process inherent in problem-solving. We are particularly interested in potential moments of insight or Aha! moments that we discussed in section 2. We use an exploratory design where a single stream of participants perform problem-solving tasks within a VR environment. We have implemented a system which requires a subject to gain mastery over the original concept of 4D space, in particular, mastery over manipulation of a 4D construct —the hypercube.

5.1 Participants

Participants were recruited from the University of Otago student and staff populations. In total, 24 participants completed the study (17 male, 7 female) with ages ranging from 18 to 45. There were no inclusion criteria with respect to domain of expertise. Participants were compensated for their time with a $50 voucher. The study was approved by the University of Otago ethics committee.

5.2 Apparatus

We implemented a system based on the principle described in Sect. 4. It involves a learner interactively manipulating rotations of a 4D cube (hypercube) and attempting to gain mastery over it. Rather than using the originally proposed visual and interactive mediums (Fig. 1 left), we implement it for use with modern immersive technology. Our implementation runs a fully immersive VR environment with an HTC head-mounted display (HMD) for visualization, and two HTC Vive controllers are used together to manipulate the rotations of the hypercube (see Fig. 1 right). Fig. 2A demonstrates the user's view into the minimal environment where the two controllers are used to manipulate the rotations of the hypercube in space.

We introduce the task aspect to the system by presenting the user with two hypercubes, one which they manipulate, and one which is static and pre-rotated. The task for the participant is to manipulate their hypercube to match (with some built-in error tolerance) the rotation of the second static hypercube. Fig. 2B demonstrates this task.

5.3 Procedure Overview

Participants put on the HMD and are given the two controllers. For the first three minutes of the experience, participants are exposed to a single hypercube which they can manipulate to get used to the environment, the controllers, and movements within the space (Fig. 2A). After three minutes, the participants were presented with a second hypercube and their goal was to try to match the hypercubes in terms of rotations (Fig. 2B). There were 30 puzzle cubes to match in total, and subjects were able to switch through the list of hypercubes by using a panel with forward and backward arrows on it. If subjects completed all 30 of the hypercube puzzles, the VR segment of the study would end, otherwise they were asked to remain for the full hour within the puzzle system to complete as many puzzles as they could. Upon completion of the study, participants were thanked, and remunerated.

5.4 Measurements

In this section we describe the measurements used in our exploratory study. We have one form of measure for each of our described dimensions of data. Emphasis is given to our ERVE methodology, as this is our main contribution.

5.4.1 Physiological

We employed the E4 wristband sensor produced by Empatica (www.empatica.com) to measure physiological signals (Fig. 3). The E4 sensor measures: electrodermal activity (EDA), blood volume pulse (BVP), heart-rate (HR), peripheral skin temperature, motion through accelerometer, and it contains an internal real-time clock. We were particularly interested in the EDA, HR, temperature, and accelerometer data. EDA and HR signals are used as our primary measure of emotional response and represent the sensed reality dimension of data in our ERVE methodology. Accelerometer and temperature data were used to wrangle our measured raw data, in particular to detect and clean artifacts. More detail will be given below on how these measures were analyzed and used.

Participant's are required to wear the E4 bracelet for the duration of the study. The bracelet was firmly fitted to the wrist/forearm of the participant. They can be worn on the ankle/leg if needed, but the wrist was appropriate for our study. Upon conclusion of each participant's session, the device is plugged into a PC containing the E4 software and the data is automatically uploaded.

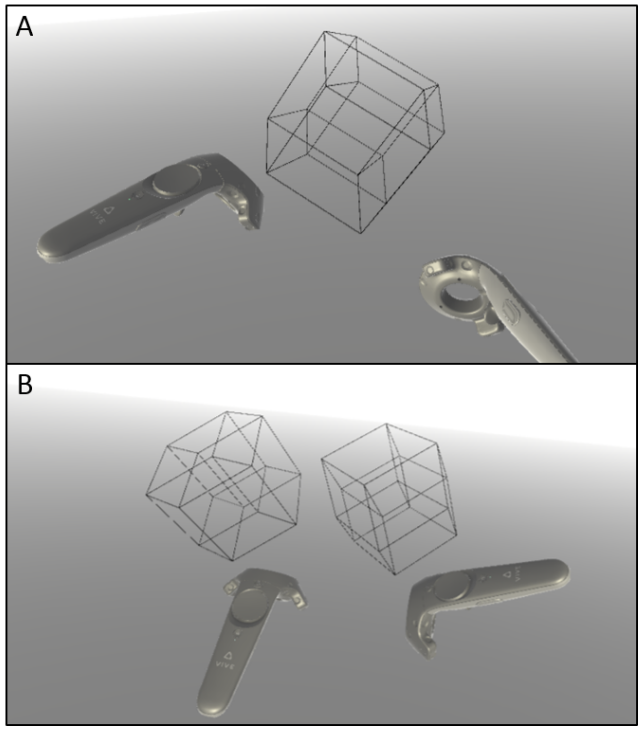

Figure 2: A) the phase where the participant uses the controllers to manipulate the hypercube alone, and B) the phase where the user manipulates their hypercube to match the rotation of a static pre-randomized hypercube.

Figure 3: Empatica E4 wristband device for collecting physiological data used in our ERVE methodology. The exposed sensors can be seen in each view of the device.

5.4.2 Observational

Our observational measure was participant successes —when they solved a hypercube puzzle. When this happened the system recorded which hypercube was solved (and therefore the hypercube difficulty) and the time-stamp of the event. In this case, the computer acts as an automatic observer of the environment. Hypercubes were categorized into easy, medium, or hard difficulty bins of which there were approximately equal numbers of each in the list of 30 hypercube puzzles.

Hypercube Difficulty. Rotational complexity refers to the combinations of 4D rotational planes (xw, yw, zw, xy, xz, yz) that are rotated, and the extent of the rotation. The rotational complexity in 4D space in terms of rotational planes is not intuitive, i.e. we expect that rotation in only one plane must be easier than rotations in all six. In fact, certain rotational planes compliment each other. For instance, if the 'xw' plane only is rotated 75 degrees, it will be a much harder ghostcube to solve than if the 'xw' and 'xy' planes are both rotated 75 degrees. Based on these differences, three categories of difficulty ratings were assigned to the hypercubes resulting in nine easy difficulty, 11 medium difficulty, and 10 hard difficulty. Rotational complexity was also rated by an expert user of the system whose ratings were similar to the difficulties established by the rotational analysis. There were a small number of hypercubes the expert rated as easy difficulty that were determined as medium by the rotational analysis, and the same from medium to hard.

5.4.3 Psychological

We are interested in contributing factors to the process of learning in VR environments. In particular, we wonder what leads learners to gain insights and experience Aha! moments. We include in our study the measure of self-perceived Aha! moments, or moments of insight. Given these moments emerge as a result of one's cognitive state, this measure is representative of the psychological (reported) dimension in our methodology. At the beginning of the study, participants are provided with a definition of an Aha! moment and are asked that if, at any point throughout their exposure, they experienced such a moment they should perform a specific action (i.e. press a button on their controllers). When this happens, the system records a time-stamp of the event. This is a naturally occurring continuous measure due to that it is an event not generated by external stimuli. It is rather a psychological event generated from within the participant. We keep the task minimal (as opposed to Danek et al. [15]) such that participants do not become too distracted from their primary task of problem solving.

5.5 Analysis

We explore with the following analysis what insights can be gained from our implementation of ERVE. We describe the required steps for first processing the data, and then how we analyze it in the context of the other dimensions of data. We will firstly briefly report on participants' performance and reporting of Aha! moments during the experiment.

5.5.1 Results

Participant solutions from the ghostcube task are expected to be a strong indicator for learning achievements. Out of 30 hypercubes presented to participants, they were able to achieve an average *13.54* with a S.D of *7.57*. In terms of difficulty, participants were able to achieve in total: 168 easy hypercubes, 138 medium hypercubes, and 39 difficult hypercubes.

13 out of 24 of the participants reported Aha! moments during the experiment where the mean number of reported moments is *1.79* with a S.D of *3.24*. The total number of recorded Aha! moments is 43. There were two observable outliers where one reported a total of 13 Aha! moments, and the other reported 11.

5.5.2 Electrodermal Activity Signal Preprocessing and Feature Extraction Tools

The clarity of the Electrodermal Activity (EDA) signal is affected by the varying intensity of physical activity and alterations in skin temperature. To mitigate these effects, we applied the EDA-Explorer tool developed by Taylor et al. [69] to filter the raw signal data. In their work, they employ two experts to manually label EDA data collected from 32 participants. Data points are labelled as either clean or artifacts based on a specified set of criteria. A total of 1301 labelled data points are given to a Support Vector Machine (SVM) to train a model (Radial basis function (RBF), β=0.1, C=1000, 60/20/20 split). This tool has a classification accuracy of 95.67% for artifact detection with those labels [69].

After removing artifacts from the signal, the two common components of EDA (tonic and phasic) are extracted. The tonic component refers to long-term skin conductance levels and slow changes, whereas the phasic component refers to short-term (event-related) changes in the signal [5], both of which are used for the analyses. We employed a convex optimization approach to decompose the EDA signal into phasic and tonic components by applying the cvxEDA-tool [26]. We extracted features from both components. Peak related features (peak, strong peak) are calculated from the phasic component whereas long-term signal features (mean, std and percentile features) are extracted from the tonic component. Seven features that are extracted from the EDA signal can be listed as: mean, standard deviation, peak, strong peak, 20th percentile, 80th percentile and quartile deviation (75th percentile, 25 percentile). These features are noted in the literature as the most discriminative for the EDA signal [1].

5.5.3 Heart Activity Signal Preprocessing and Feature Extraction Tools

The heart activity is also exposed to signal contamination due to the movement of subjects. To address this, a preprocessing tool has been developed in MATLAB which employs the 20 percent rule on data and a local average. In this rule, every data point is compared with the local average and detected as an artifact if the difference is higher than 20%. We delete the data points that do not satisfy this rule. The 20% rule for artifact detection is commonly used in the literature [13]. We have implemented parameters that can be used to either remove the artifact points, or adjust artifact points by applying shape preserving cubic spline interpolation. If the artifacts are removed and not interpolated, the cleaning tool can impose new constraints on the remaining clean data. It requires N consecutive data points, or it can set a minimum consecutive time rule similarly (non-interrupted with deleted artifacts) to evaluate the segment worth processing. The percentage parameter of the artifact detection rule and window length of local mean calculation for data point comparison can also be adjusted from the preprocessing tool.

MATLAB's built-in functions along with Marcus Vollmer's HRV (heart rate variability) toolbox[1] [71, 72] was applied to extract heart activity features. The following time domain features were extracted: HR mean, standard deviation of the inter-beat interval, mean value of the inter-beat intervals, root mean square of successive difference of the inter-beat intervals, the percentage of the number of successive inter-beat intervals varying more than 50ms from the previous interval, the total number of inter-beat intervals divided by the height of the histogram of all inter-beat intervals measured on a scale with bins of 1/128s (HRV triangular index), and triangular interpolation of inter-beat interval histogram. We applied a Fast Fourier Transform (FFT) approach to isolate the separate frequencies within the data. We were able to determine low frequency power (LF), high frequency power (HF), very low frequency power, prevalent low frequency, prevalent high frequency, and the ratio of LF to HF (LF/HF).

[1] marcusvollmer.github.io/HRV/

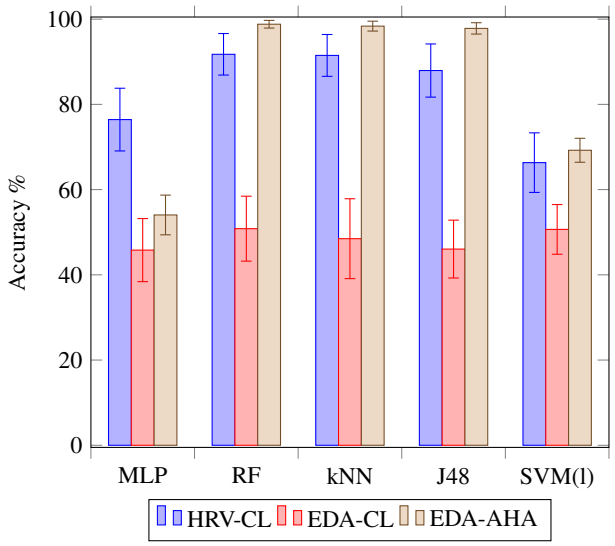

Figure 4: Bar plot showing all five classifiers, indicating that Cognitive Load classification using HR features (HRV-CL) and Aha moment (insight) classification using EDA features (EDA-AHA) yield the highest prediction accuracy. Cognitive Load classified using EDA features (EDA-CL) yields low prediction rates.

Lomb-Scargle periodogram was applied to analyze the periodicity of the LF, HF and LF/HF time series data. The features that were extracted are the most widely used, and the most discriminative according to the literature [1].

With the EDA and Heart Activity data preprocessed, we conduct the classification with respect to observed data (ghostcube solutions) and reported data (Aha! moments).

5.5.4 Classification of Cognitive Load

Electrodermal Activity and Solution Events. We first analyze the relationship between EDA data, and solution events. We estimate cognitive load trends based on the solution difficulties established earlier. During easier solutions, participants would be experiencing less cognitive difficulty, and harder solutions require more cognitive effort. The difficulty levels of the ghostcubes (easy, medium, and hard) solved by participants in our experiment were used as labels for machine learning algorithms (1, 2, and 3 respectively). In order to classify three cognitive load levels, we have used the Weka toolkit [19]. These classes were imbalanced due to the nature of the data where "hard" labels represent the minority class. We employed a re-sample method from Weka toolkit to balance the data (i.e added samples of the minority class) to prevent classifiers from biasing towards the majority class. To ensure we are providing an exhaustive analysis, we test our EDA features (extracted earlier) using multiple classifiers. We have applied five different classifiers on our cognitive load data:

A. PCA and SVM with linear basis function

B. MultiLayer Perceptron (7-5-3) (MLP)

C. K-nearest neighbours (k=1)

D. J48 Decision Tree

E. Random Forest (RF, 100 trees)

These classifiers are selected due to their common application for physiological signal processing in the literature. All classifiers in the Weka toolkit were run with the algorithms' default values. For each classifier, results are validated with 10-fold cross validation. Results have been provided for 3-class classification of low, medium and high cognitive load levels. We tried the selected machine learning algorithms on our EDA data (see Fig. 4 (EDA-CL)). The resulting average classification accuracy across the five different classifiers was 48.36% when discriminating using three difficulty levels of cognitive load. The most successful classifier for EDA data was the Random Forest approach which achieved 50.83% accuracy with 7.62% variance.

We applied the same process to the heart activity data collected from the Empatica E4 device using the same cognitive labels. We were able to discriminate the three cognitive load levels with a resulting average classification accuracy of 82.79%. For the heart activity data, the most successful classifier is the Random Forest approach which achieved 91.75% with variance 4.87% (see Fig. 4 (HRV-CL)).

5.5.5 Classification of Aha! Moments

Aha! experiences (moments of insight) are times the participants felt that they made a conceptual breakthrough in determining a solution. These events occur rarely. In the duration of our experiment, less than two of these moments occur in sixty minutes on average. This means that the detection of the events is not trivial. EDA is excellent for determining the arousal level, since it can assess changes in the SNS (sympathetic nervous system) [79]. Since these moments represent a form of arousal, we have used the EDA signals which can detect instant arousal of individuals. We have employed the same EDA tools as specified above in Sect. 5.5.2 for feature extraction of the EDA data.

We were able to classify against multiple classes for the cognitive load classification due to our difficulty categorization. Aha! moments can not be categorized in such a way, so a binary classification is required. We divided each 60 minute session into 60 segments, each one minute long. A one or a zero was assigned to a segment depending on whether a participant reported an Aha! moment within that minute or not. These values are then given to the classifier.

By applying the Weka toolkit with the same machine learning algorithms, we achieved an average accuracy of 83.65%. The most successful classifier was again the Random Forest approach achieving an accuracy of 98.81% with a variance of 0.9 (see Fig. 4 (EDA-AHA)).

5.6 Interpretation

The average classification accuracy for the cognitive load with the EDA was 48.36% with the lowest classifier yielding an accuracy of 45.8%, and the highest classifier yielding 50.83%. The primary reason the EDA classification has returned low accuracy is due to the short consecutive segments of cognitive load data. Puzzle completion times were often quite close together (in the order of 10 seconds). The more delayed nature of the EDA makes it more difficult to isolate significant emotional responses against such frequent cognitive events. For instance, a user's EDA will rise as they spend 3 minutes on a level-3 puzzle and then upon completion they are presented with a level-1 puzzle which they complete in 10 seconds. The EDA takes longer to stabilize making a correct decision difficult for the classifier.

The HR data does not have the issue of delay as it is one of the first responses to the sympathetic nervous system. This explains the consistently higher classification accuracy of cognitive load against the HR data of 82.78%. By looking at the HR data alone, we are able to predict the difficulty of environmental subject matter at the approximate rate of the achieved accuracy. For example, we could tell when a subject is struggling with a particular problem, or when they are finding particular problems easy.

The Aha! moment classification was higher again with an average accuracy of 83.66%, meaning that we have identified emotional responses in the EDA data which indicate the subject has had, is having, or is about to have an Aha! moment. Classifier performances are aligned with recent investigations of classification algorithms for wearable sensor data [11].

6 Discussion

Using our proposed ERVE methodology, we conducted an exploratory study which doubles as a description of how to apply our methodology. Included is an analysis of two interactions between the three independent data dimensions. We were able to isolate with high accuracy physiological determinants of cognitive load, and subjective moments of insight with users in a VR learning context. With more refinement and testing, analyses using ERVE can yield informative results which can be used in the delivery of virtual content across a more broad spectrum of applications. We could reliably identify the difficulty users were experiencing while trying to learn a new concept in a minimal environment with no introduction of emotionally provocative environmental content other than the subject matter. More can be gained from the ERVE method such as investigating other possible interactions in a wide range of VR application scenarios.

6.1 VR Application Spaces

Our ERVE methodology and analysis technique is not only useful in the narrowly defined context of learning in VR. Actually, it can and perhaps should be applied in all VR evaluations where emotional responses matter. VR environments are the primary target for our methodology due to the controlled and flexible nature of VR applications. We are able to control, record, and measure variables within that space which is what makes the VR context immediately suitable for the methodology. Application of our methodology outside the VR space is a focus for future work. The following scenarios are examples for such applications of ERVE.

Training and instruction often fall under the umbrella of education, and it is certainly the case that something is being learned, though it is most often hard skills that are learned through drill and repeat practices. In some cases training can mean providing a trainee with situational experiences. VR has the ability to produce such experiences for trainees that are difficult and/or expensive to simulate in the real world. An example of such an experience is an emergency room situation where a person in critical condition is rushed into the hospital. These experiences will be most valuable if the trainees are engaged with the environment and genuine emotional responses are generated.

The gaming and entertainment industry benefits from having highly engaged users. This includes (arguably more so) serious games. Game developers generally aim to create experiences that players will not forget. Many game development concepts feed into maximizing that experience, although in the context of VR, game development processes are still being established. The set of conventions that underlie game development as we know it do not all necessarily apply to VR game development, so as the industry works towards a new framework for immersive development, our ERVE methodology will be a useful tool in validating that the experiences being generated are producing the emotional engagement developers want for their players.

Health treatments and well-being applications are common areas of interest in the VR research community and, as with training and entertainment systems, these applications can benefit from ERVE throughout their development. Presence studies have been conducted in the context of phobia treatment and have established the importance that a user achieves a sense of presence for the treatment to be effective [54]. Presence has been shown in many cases to be a pivotal factor for the success of VR applications. Generally speaking, VR can be described as a system which is computer-generated, three-dimensional, and is interactive in real-time such that users experience a sense of presence. Given that presence is one of the defining characteristics of successful VR applications, we need to work to maximize that construct. Involvement was found to be a significant factor for the sense of presence in virtual environments [58], and it is a factor which requires emotional engagement. Therefore our ERVE methodology becomes relevant in any given VR context that wishes to maximize a user's level of engagement or involvement with the virtual environment. We can not only measure a user's emotional responses during exposure for later analyses, but we can potentially use that data in real-time to tailor the environment for each user to maximize their engagement levels, resulting in an increased sense of presence and therefore a maximized VR experience.

6.2 Limitations and Future Work

We have been successful in implementing our ERVE method in our context of learning. There are however a number of outstanding questions and observations which demonstrate the potential of ERVE.

The EDA data has both tonic and phasic components where the tonic is the long-term component. It would be interesting and potentially beneficial to analyze the tonic forms of the EDA data before and after significant observational events. To take the data sets we used in this work - the subject successes and Aha! moments, what would the tonic forms of the EDA say about the process leading to those moments? Are there any emerging patterns? Furthermore, it could be possible to classify a combination of HR and EDA signals to improve the classification accuracy. These are interesting directions for future work.

As our methodology is applied in more contexts, we also have to consider what different kinds of reported, observed, and sensed data might be measured. We have described one analysis technique that will work for a broad range of datasets, however other data will be more difficult to analyze against physiological measures such as interview and questionnaire results. Perhaps new methods of collecting reported measures can be devised for the purposes of application in the ERVE methodology. Furthermore, reported data collected through post-experience questionnaires can be difficult to use in the context of ERVE. Interview and questionnaire data can be a useful indicator of what kind of sensed data is being collected due to the potential confounding factors that could impact physiological measures. For instance, our measures of emotional response originate in the sympathetic nervous system which is responsible for the fight or flight instinct. Therefore, more subjective data can help to further contextualize sensed data. This is an interesting space for future work.

There are a number of potential issues that arise upon reflection of this methodology. Physiological data, particularly EDA and HR data, are susceptible to noise from physical movement. One of the benefits afforded by VR environments is their interactive capabilities requiring movement of the user, often using their arms and hands. Should a wristband be used for data collection, this could result in noisy data. We have employed techniques to mitigate for this in the preprocessing step of the analysis, however this should be investigated and perhaps compared to participants in more static scenarios.

Given the emphasis of emotional response in our analyses, we should consider the novelty of VR use. Users' first experiences in VR often produce novelty effects which could have an impact on physiological measures of their sub-conscious emotional states. It was for this reason that we generated a minimalist VR environment, and also provided participants with a long exposure time of one hour. If novelty effects are occurring, they should likely be reduced after

the participant has settled into the environment and are focusing on the task. There is potential here though for more investigation to confirm these speculations.

With respect to our study context, we should be careful that certain tasks which are presented to participants will not become detrimental to their performance in other tasks or result in a bias on any other measurements. In our case, it was tasking users to report any Aha! moments they might experience during the exposure. It is possible that asking participants to keep this in mind could result in their divided attention. Despite this potential issue, we observed participants more often realizing they are having such a moment and remembering then to report it rather than always being distracted by it. Future work could consider alternative or additional observational or sensed measures of insight occurrences to mitigate these potential issues.

More generally, the study of insight or Aha! moments is difficult due to their often rare occurrence. This has made it difficult to investigate however as was reported earlier, breakthrough thinking approaches to learning are gaining increasing attention and if we can facilitate the required conditions, we can improve users' chances of achieving moments of insight more frequently. We have been able to provide users with a problem space and an environment to achieve these moments and our methodology provided us with a structured approach to measuring this phenomenon. Further investigation is encouraged in this space to solidify findings, to improve our understanding of how we can better implement VR learning environments, and to fine-tune analysis techniques.

Two potential benefits become apparent when thinking about the outcomes of our analysis. The first is the impact it will have on current education-based VR system evaluations that rely on pre- and post-exposure reported, and observed measures only. With more implementations of this kind and more data over time, we will be able to increase the reliability of these analyses. This means educators could analyze data obtained from learners exposures to subject matter and determine, for instance, which content in particular they are finding difficult in the VR environment. As in the real world, different learners take different approaches to problems. This analysis could be extended to help clarify the emotional processes behind different approaches and how they might have an effect on the learning outcomes. How we can use the insights of those analyses to further facilitate education in virtual environments is up for discussion and will likely become more clear as the data emerges.

Our approach also presents the opportunity for a real-time user feedback implementation of this measure. A common issue people encounter when attempting to solve problems (or resolve habits) is they exhibit mal-adaptive behaviours which hinder them from accomplishing their goals. By applying this methodology in a real-time feedback loop, it would be possible to have a VR environment alter the way is it delivering subject matter according to the emotional responses and mal-adaptive behaviours exhibited by the subjects. This would hopefully drive them toward improving their behaviours, therefore impacting the way they approach problems. A further real-time approach of interest would be to present users with a representation of their own physiological data (which is representative of their sub-conscious emotional state) while immersing them in various situations and scenarios. This approach could have potential application in VR mental health treatments.

The wider space of mixed and augmented reality applications can also benefit from these forms of analyses. As in the case of VR, MR is applied across multiple different domain spaces where user's engagement is a desirable outcome. AR has been considered in the space of education, a space we have shown in this work can benefit from our methodology. Other spaces such as commercial and industrial, medical, entertainment, and collaborative domains have received focus from augmented and mixed reality developers, for all of which psychological involvement of their users is desirable.

7 Conclusion

We have designed and implemented a new methodology based on measuring physiological responses from which we infer emotional states in virtual environments. We implemented the methodology in a VR learning context to demonstrate its implementation and possible outcomes. By sensing users' physiological responses and conducting our analyses we have been able to predict with high accuracy the degree to which VE and task stimuli has impacted the emotional states of subjects. We can determine through heart activity the cognitive load a user is under in VR, and using electrodermal activity we can predict when users are approaching, experiencing, or have had an Aha! moment. These results are indicative of the potential of the ERVE methodology described in this work. We concluded with a discussion on the impact of the methodology on modern VR systems development and potential applications of these measures in other VR contexts. ERVE is relevant in any VR context, especially for researchers and developers intending to maximize the engagement and response levels of their users.

Acknowledgments

We would like to thank the members of the Otago Human-Computer Interaction group, the Otago University Information Science department, and Brendon Woodford for their support and input on the project. Thanks also to the study participants for their time.

References

[1] A. Alberdi, A. Aztiria, and A. Basarab. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of Biomedical Informatics*, 59:49 – 75, 2016. doi: 10.1016/j.jbi.2015.11.007

[2] M. Alghamdi, H. Regenbrecht, S. Hoermann, and N. Swain. Mild stress stimuli built into a non-immersive virtual environment can elicit actual stress responses. *Behaviour & Information Technology*, 36(9):913–934, Sept. 2017. doi: 10.1080/0144929X.2017.1311374

[3] N. Ali, S. Ullah, A. Alam, and J. Rafique. 3d Interactive Virtual Chemistry Laboratory for Simulation of High School Experiments. *Proceedings of EURASIA GRAPHICS*, 2014.

[4] P. Arnold. A Proposal for a Study of the Mechanisms of Perception of, and Formation of Internal Representations of, the Spatial Fourth Dimension. *Accomplishment Summary*, 71(72):223–235, 1972.

[5] M. Benedek and C. Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1):80–91, June 2010. doi: 10.1016/j.jneumeth.2010.04.028

[6] D. A. Bowman and R. P. McMahan. Virtual Reality: How Much Immersion Is Enough? *Computer*, 40(7):36–43, July 2007. doi: 10.1109/MC.2007.257

[7] A. Brogni, V. Vinayagamoorthy, A. Steed, and M. Slater. Variations in Physiological Responses of Participants During Different Stages of an Immersive Virtual Environment Experiment. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, VRST '06, pp. 376–382. ACM, New York, NY, USA, 2006. doi: 10.1145/1180495.1180572

[8] J. Brooke. SUS - A quick and dirty usability scale. p. 8, 1996.

[9] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[10] R. A. Calvo and S. D'Mello. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, Jan. 2010. doi: 10.1109/T-AFFC.2010.1

[11] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy. Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, 19(8):1849, 2019.

[12] C. Christiansen, B. Abreu, K. Ottenbacher, K. Huffman, B. Masel, and R. Culpepper. Task performance in virtual environments used for cognitive rehabilitation after traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, 79(8):888–892, Aug. 1998. doi: 10.1016/S0003-9993(98)90083-1

[13] B. Cinaz, B. Arnrich, R. Marca, and G. Tröster. Monitoring of mental workload levels during an everyday life office-work scenario. *Personal Ubiquitous Comput.*, 17(2):229–239, Feb. 2013. doi: 10.1007/s00779-011-0466-1

[14] H. D. Critchley. Review: Electrodermal Responses: What Happens in the Brain. *The Neuroscientist*, 8(2):132–142, Apr. 2002. doi: 10.1177/107385840200800209

[15] A. H. Danek and J. Wiley. What about False Insights? Deconstructing the Aha! Experience along Its Multiple Dimensions for Correct and Incorrect Solutions Separately. *Frontiers in Psychology*, 7, 2017. doi: 10.3389/fpsyg.2016.02077

[16] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in Psychology*, 6, 2015. doi: 10.3389/fpsyg.2015.00026

[17] J. Diemer, A. Mühlberger, P. Pauli, and P. Zwanzger. Virtual reality exposure in anxiety disorders: Impact on psychophysiological reactivity. *The World Journal of Biological Psychiatry*, 15(6):427–442, Aug. 2014. doi: 10.3109/15622975.2014.892632

[18] J. S. Dumas and M. C. Salzman. Usability Assessment Methods. *Reviews of Human Factors and Ergonomics*, 2(1):109–140, Apr. 2006. doi: 10.1177/1557234X0600200105

[19] F. Eibe, M. Hall, and I. Witten. The weka workbench. online appendix for" data mining: Practical machine learning tools and techniques. *Morgan Kaufmann*, 2016.

[20] A. Felnhofer, O. D. Kothgassner, T. Hetterle, L. Beutl, H. Hlavacs, and I. Kryspin-Exner. Afraid to Be There? Evaluating the Relation Between Presence, Self-Reported Anxiety, and Heart Rate in a Virtual Public Speaking Task. *Cyberpsychology, Behavior, and Social Networking*, 17(5):310–316, Mar. 2014. doi: 10.1089/cyber.2013.0472

[21] A. Felnhofer, O. D. Kothgassner, M. Schmidt, A.-K. Heinzle, L. Beutl, H. Hlavacs, and I. Kryspin-Exner. Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. *International Journal of Human-Computer Studies*, 82:48–56, Oct. 2015. doi: 10.1016/j.ijhcs.2015.05.004

[22] L. Freina and M. Ott. A Literature Review on Immersive Virtual Reality in Education: State Of The Art and Perspectives. *eLearning & Software for Education*, (1), 2015.

[23] J. Frommel, M. Weber, K. Rogers, J. Brich, D. Besserer, L. Bradatsch, I. Ortinau, R. Schabenberger, V. Riemer, and C. Schrader. Integrated Questionnaires: Maintaining Presence in Game Environments for Self-Reported Data Acquisition. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*, pp. 359–368. ACM Press, London, United Kingdom, 2015. doi: 10.1145/2793107.2793130

[24] M. Garau, M. Slater, D.-P. Pertaub, and S. Razzaque. The Responses of People to Virtual Humans in an Immersive Virtual Environment. *Presence: Teleoperators and Virtual Environments*, 14(1):104–116, Feb. 2005. doi: 10.1162/1054746053890242

[25] N. Gavish, T. Gutiérrez, S. Webel, J. Rodríguez, M. Peveri, U. Bockholt, and F. Tecchia. Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments*, 23(6):778–798, Nov. 2015. doi: 10.1080/10494820.2013.815221

[26] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, April 2016. doi: 10.1109/TBME.2015.2474131

[27] C. Guger, G. Edlinger, R. Leeb, and G. Pfurtscheller. *Heart-Rate Variability and Event-Related ECG in Virtual Environments*. 2004.

[28] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. H. a. N. Meshkati, ed., *Advances in Psychology*, vol. 52 of *Human Mental Workload*, pp. 139–183. North-Holland, 1988.

[29] K. F. Hew and W. S. Cheung. Use of three-dimensional (3-D) immersive virtual worlds in K-12 and higher education settings: A review of the research. *British Journal of Educational Technology*, 41(1):33–55, 2010. doi: 10.1111/j.1467-8535.2008.00900.x

[30] H. G. Hoffman, T. L. Richards, B. Coda, A. R. Bills, D. Blough, A. L. Richards, and S. R. Sharar. Modulation of thermal pain-related brain activity with virtual reality: Evidence from fMRI. *NeuroReport: For Rapid Communication of Neuroscience Research*, 15(8):1245–1248, 2004. doi: 10.1097/01.wnr.0000127826.73576.91

[31] W. A. IJsselsteijn, H. d. Ridder, J. Freeman, and S. E. Avons. Presence: concept, determinants, and measurement. In *Human Vision and Electronic Imaging V*, vol. 3959, pp. 520–530. International Society for Optics and Photonics, June 2000. doi: 10.1117/12.387188

[32] J. Jordan and M. Slater. An Analysis of Eye Scanpath Entropy in a Progressively Forming Virtual Environment. *Presence: Teleoperators and Virtual Environments*, 18(3):185–199, June 2009. doi: 10.1162/pres.18.3.185

[33] H. Kaufmann, D. Schmalstieg, and M. Wagner. Construct3d: A Virtual Reality Application for Mathematics and Geometry Education. *Education and Information Technologies*, 5(4):263–276, Dec. 2000. doi: 10.1023/A:1012049406877

[34] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.

[35] S. Khalfa, P. Isabelle, B. Jean-Pierre, and R. Manon. Event-related skin conductance responses to musical emotions in humans. *Neuroscience Letters*, 328(2):145–149, Aug. 2002. doi: 10.1016/S0304-3940(02)00462-7

[36] J. Kirakowski and M. Corbett. SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology*, 24(3):210–212, Sept. 1993. doi: 10.1111/j.1467-8535.1993.tb00076.x

[37] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, Jan. 2012. doi: 10.1109/T-AFFC.2011.15

[38] J. Kounios and M. Beeman. The Aha! Moment: The Cognitive Neuroscience of Insight. *Current Directions in Psychological Science*, 18(4):210–216, Aug. 2009. doi: 10.1111/j.1467-8721.2009.01638.x

[39] B. Laugwitz, T. Held, and M. Schrepp. Construction and Evaluation of a User Experience Questionnaire. In A. Holzinger, ed., *HCI and Usability for Education and Work*, Lecture Notes in Computer Science, pp. 63–76. Springer Berlin Heidelberg, 2008.

[40] M. Limniou, D. Roberts, and N. Papadopoulos. Full immersive virtual environment CAVETM in chemistry education. *Computers & Education*, 51(2):584–593, Sept. 2008. doi: 10.1016/j.compedu.2007.06.014

[41] C. Lin, I. Chung, L. Ko, Y. Chen, S. Liang, and J. Duann. EEG-Based Assessment of Driver Cognitive Responses in a Dynamic Virtual-Reality Driving Environment. *IEEE Transactions on Biomedical Engineering*, 54(7):1349–1352, July 2007. doi: 10.1109/TBME.2007.891164

[42] M. Lombard, T. B. Ditton, and L. Weinstein. Measuring Presence: The Temple Presence Inventory. *Proceedings of the 12th Annual International Workshop on Presence*, p. 15, 2009.

[43] E. Malbos, R. M. Rapee, and M. Kavakli. Behavioral Presence Test in Threatening Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 21(3):268–280, Aug. 2012. doi: 10.1162/PRES_a_00112

[44] R. L. Mandryk, M. S. Atkins, and K. M. Inkpen. A Continuous and Objective Evaluation of Emotional Experience with Interactive Play Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pp. 1027–1036. ACM, New York, NY, USA, 2006. doi: 10.1145/1124772.1124926

[45] M. Meehan, B. Insko, M. Whitton, and F. P. Brooks, Jr. Physiological Measures of Presence in Stressful Virtual Environments. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, pp. 645–652. ACM, New York, NY, USA, 2002. doi: 10.1145/566570.566630

[46] M. Meehan, S. Razzaque, B. Insko, M. Whitton, and F. P. Brooks. Review of Four Studies on the Use of Physiological Reaction as a Measure of Presence in Stressful Virtual Environments. *Applied Psychophysiology and Biofeedback*, 30(3):239–258, Sept. 2005. doi: 10.1007/s10484-005-6381-3

[47] Z. Merchant, E. T. Goetz, L. Cifuentes, W. Keeney-Kennicutt, and T. J. Davis. Effectiveness of virtual reality-based instruction on students'

learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education*, 70:29–40, Jan. 2014. doi: 10.1016/j.compedu.2013.07.033

[48] K. Moore, B. K. Wiederhold, M. D. Wiederhold, and G. Riva. Panic and Agoraphobia in a Virtual World. *CyberPsychology & Behavior*, 5(3):197–202, June 2002. doi: 10.1089/109493102760147178

[49] M. Morozov, A. Tanakov, A. Gerasimov, D. Bystrov, and E. Cvirco. Virtual chemistry laboratory for school education. In *IEEE International Conference on Advanced Learning Technologies, 2004. Proceedings.*, pp. 605–608, Aug. 2004. doi: 10.1109/ICALT.2004.1357486

[50] S. Ohlsson. *Deep Learning: How the Mind Overrides Experience*. Cambridge University Press, Jan. 2011. Google-Books-ID: Qb33YMi1868C.

[51] H. M. Peperkorn, J. Diemer, and A. Mühlberger. Temporal dynamics in the relation between presence and fear in virtual reality. *Computers in Human Behavior*, 48:542–547, July 2015. doi: 10.1016/j.chb.2015.02.028

[52] D. Perkins. *The Eureka Effect: The art and logic of breakthrough thinking*. The Eureka Effect: The art and logic of breakthrough thinking. W W Norton & Co, New York, NY, US, 2000.

[53] M. Poh, N. C. Swenson, and R. W. Picard. A Wearable Sensor for Unobtrusive, Long-Term Assessment of Electrodermal Activity. *IEEE Transactions on Biomedical Engineering*, 57(5):1243–1252, May 2010. doi: 10.1109/TBME.2009.2038487

[54] H. T. Regenbrecht, T. W. Schubert, and F. Friedmann. Measuring the Sense of Presence and its Relations to Fear of Heights in Virtual Environments. *Internation Journal of Human-Computer Interaction*, 10(3):233–249, Sept. 1998. doi: 10.1207/s15327590ijhc1003_2

[55] D. P. Saha, R. B. Knapp, and T. L. Martin. Affective feedback in a virtual reality based intelligent supermarket. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pp. 646–653. ACM, 2017.

[56] M. V. Sanchez-Vives and M. Slater. From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, 6(4):332–339, Apr. 2005. doi: 10.1038/nrn1651

[57] S. Sandkhler and J. Bhattacharya. Deconstructing Insight: EEG Correlates of Insightful Problem Solving. *PLOS ONE*, 3(1):e1459, Jan. 2008. doi: 10.1371/journal.pone.0001459

[58] T. Schubert, F. Friedmann, and H. Regenbrecht. The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments*, 10(3):266–281, June 2001. doi: 10.1162/105474601300343603

[59] K.-C. Shim, J.-S. Park, H.-S. Kim, J.-H. Kim, Y.-C. Park, and H.-I. Ryu. Application of virtual reality technology in biology education. *Journal of Biological Education*, 37(2):71–74, Mar. 2003. doi: 10.1080/00219266.2003.9655854

[60] M. Slater. How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence*, 13(4):484–493, 2004.

[61] M. Slater, C. Guger, G. Edlinger, R. Leeb, G. Pfurtscheller, A. Antley, M. Garau, A. Brogni, and D. Friedman. Analysis of Physiological Responses to a Social Situation in an Immersive Virtual Environment. *Presence: Teleoperators and Virtual Environments*, 15(5):553–569, Oct. 2006. doi: 10.1162/pres.15.5.553

[62] M. Slater, P. Khanna, J. Mortensen, and I. Yu. Visual Realism Enhances Realistic Response in an Immersive Virtual Environment. *IEEE Computer Graphics and Applications*, 29(3):76–84, May 2009. doi: 10.1109/MCG.2009.55

[63] M. Slater, D.-P. Pertaub, C. Barker, and D. M. Clark. An Experimental Study on Fear of Public Speaking Using a Virtual Environment. *CyberPsychology & Behavior*, 9(5):627–633, Oct. 2006. doi: 10.1089/cpb.2006.9.627

[64] M. Slater, M. Usoh, and Y. Chrysanthou. The Influence of Dynamic Shadows on Presence in Immersive Virtual Environments. In M. Göbel, ed., *Virtual Environments '95*, Eurographics, pp. 8–21. Springer Vienna, 1995.

[65] K. S. Song and W. Y. Lee. A virtual reality application for geometry classes. *Journal of Computer Assisted Learning*, 18(2):149–156, 2002. doi: 10.1046/j.0266-4909.2001.00222.x

[66] J. Steuer. Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication*, 42(4):73–93, Dec. 1992. doi: 10.1111/j.1460-2466.1992.tb00812.x

[67] K.-T. Sun, H.-T. Chan, and K.-C. Meng. Research on the application of virtual reality on arts core curricula. pp. 234–239, 2010. doi: 10.1109/ICCIT.2010.5711063

[68] A. G. Sutcliffe and K. D. Kaur. Evaluating the usability of virtual reality user interfaces. *Behaviour & Information Technology*, 19(6):415–426, Jan. 2000. doi: 10.1080/014492900750052679

[69] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, , and R. Picard. Automatic identification of artifacts in electrodermal activity data. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 7:1934, 2015.

[70] S. Topolinski and R. Reber. Gaining Insight Into the Aha Experience. *Current Directions in Psychological Science*, 19(6):402–405, Dec. 2010. doi: 10.1177/0963721410388803

[71] M. Vollmer. Marcusvollmer/hrv toolbox.

[72] M. Vollmer. A robust, simple and reliable measure of heart rate variability using relative RR intervals. In *2015 Computing in Cardiology Conference (CinC)*, pp. 609–612, Sept. 2015. doi: 10.1109/CIC.2015.7410984

[73] H. Von Foerster, E. von Glasersfeld, and P. M. Hejl. *Einfhrung in den Konstruktivismus*. Piper, 1992.

[74] R. W. Weisberg. Toward an integrated theory of insight in problem solving. *Thinking & Reasoning*, 21(1):5–39, Jan. 2015. doi: 10.1080/13546783.2014.886625

[75] B. K. Wiederhold, R. Gevirtz, and M. D. Wiederhold. Fear of Flying: A Case Report Using Virtual Reality Therapy with Physiological Monitoring. *CyberPsychology & Behavior*, 1(2):97–103, Jan. 1998. doi: 10.1089/cpb.1998.1.97

[76] B. K. Wiederhold, D. P. Jang, S. I. Kim, and M. D. Wiederhold. Physiological Monitoring as an Objective Tool in Virtual Reality Therapy. *CyberPsychology & Behavior*, 5(1):77–82, Feb. 2002. doi: 10.1089/109493102753685908

[77] W. W. Wierwille and J. G. Casali. A Validated Rating Scale for Global Mental Workload Measurement Applications. *Proceedings of the Human Factors Society Annual Meeting*, 27(2):129–133, Oct. 1983. doi: 10.1177/154193128302700203

[78] Y. Wu, T. Chan, B. Jong, and T. Lin. A Web-based virtual reality physics laboratory. In *Proceedings 3rd IEEE International Conference on Advanced Technologies*, pp. 455–, July 2003. doi: 10.1109/ICALT.2003.1215178

[79] R. Zangróniz, A. Martínez-Rodrigo, J. Pastor, M. López, and A. Fernández-Caballero. Electrodermal activity sensor for classification of calm/distress condition. *Sensors*, 17(10):2324, 2017.