Virtual Reality
https://doi.org/10.1007/s10055-021-00518-1

ORIGINAL ARTICLE

Towards estimating affective states in Virtual Reality based on behavioral data

Valentin Holzwarth[1] · Johannes Schneider[1] · Joshua Handali[1] · Joy Gisler[2] · Christian Hirt[2] · Andreas Kunz[2] · Jan vom Brocke[1]

Received: 31 October 2020 / Accepted: 17 March 2021

Abstract

Inferring users' perceptions of Virtual Environments (VEs) is essential for Virtual Reality (VR) research. Traditionally, this is achieved through assessing users' affective states before and after being exposed to a VE, based on standardized, self-assessment questionnaires. The main disadvantage of questionnaires is their sequential administration, i.e., a user's affective state is measured asynchronously to its generation within the VE. A synchronous measurement of users' affective states would be highly favorable, e.g., in the context of adaptive systems. Drawing from nonverbal behavior research, we argue that behavioral measures could be a powerful approach to assess users' affective states in VR. In this paper, we contribute by providing methods and measures evaluated in a user study involving 42 participants to assess a users' affective states by measuring head movements during VR exposure. We show that head yaw significantly correlates with presence, mental and physical demand, perceived performance, and system usability. We also exploit the identified relationships for two practical tasks that are based on head yaw: (1) predicting a user's affective state, and (2) detecting manipulated questionnaire answers, i.e., answers that are possibly non-truthful. We found that affective states can be predicted significantly better than a naive estimate for mental demand, physical demand, perceived performance, and usability. Further, manipulated or non-truthful answers can also be estimated significantly better than by a naive approach. These findings mark an initial step in the development of novel methods to assess user perception of VEs.

Keywords Virtual Reality · Affective VR · Sensor Data · Affective States

1 Introduction

Virtual Reality (VR) is a technology that has been heavily researched since the 1980s, but only recently became affordable and accessible (Slater 2018). This sparked the development of applications in diverse contexts such as entertainment, marketing, and training. Additionally, decades of VR research have established a rich knowledge base on how to generally intensify the VR experience, e.g., by raising users' presence (Cummings and Bailenson 2015).

Given the example of a Virtual Training Environment (VTE), it is essential to assess users' mental demand. A too low mental demand could cause users' boredom, whereas excessive mental demand could be tiring and overtaxing. However, each user perceives the same Virtual Environment (VE) differently, due to a combination of individual traits, such as prior VR exposure or context-specific vocational experience (Hirt et al. 2019). Thus, a one-fits-all approach regarding the design of a VE will not be expedient and

✉ Andreas Kunz
kunz@iwf.mavt.ethz.ch

Valentin Holzwarth
valentin.holzwarth@uni.li

Johannes Schneider
johannes.schneider@uni.li

Joshua Handali
joshua.handali@uni.li

Joy Gisler
gisler@iwf.mavt.ethz.ch

Christian Hirt
hirtc@iwf.mavt.ethz.ch

Jan vom Brocke
jan.vom.brocke@uni.li

[1] Institute of Information Systems, University of Liechtenstein, Vaduz, Liechtenstein

[2] Innovation Center Virtual Reality, ETH Zurich, Zurich, Switzerland

Published online: 09 April 2021

best practices in VTE development are difficult to derive (Jensen and Konradsen 2017). A potential solution could be the measurement of a user's affective state, such as mental demand, to adapt the VE accordingly. In this context, considerable prior research has investigated both on the induction and measurement of affective states in VR (e.g., Zhang et al. 2017; Hirt et al. 2020; Marín-Morales et al. 2020). Traditionally, affective states are assessed through user studies, wherein a representative sample of participants discloses their traits, uses a VR application, and finally self-reports on their experience after using the application, e.g., how present they felt. Such self-reports are usually formalized through standardized questionnaires, such as the NASA Task Load Index (TLX) (Hart and Staveland 1988).

Questionnaires are well-established and thus part of many user studies in the field of VR and more broadly in the field of human–computer interaction (MacKenzie 2013). This popularity is due to various advantages such as easy administration, scoring and coding of questionnaires (Newsted et al. 1998). Further, questionnaires eliminate interviewer bias, i.e., interviewer traits such as sex and age, may result in different interviewee answers on the same question (Gillham 2007). However, questionnaires also come with numerous issues. First, answering questionnaires is time-consuming and thus increases the user study's duration. Second, data obtained through questionnaires is subjective and thus prone to be unreliable, e.g., participants in a user study might be unmotivated or bored, which could lead them to not carefully answer each question (Mertens et al. 2017). Third, study participants have little incentive to invest effort and answer genuinely, since their financial compensation is generally not related to the quality of questionnaire responses. Thus, it is not uncommon that some participants of a study have to be removed due to seemingly bogus answers (Chmielewski and Kucker 2019). Fourth, questionnaires are deployed after exposure to a VE. As there is always an inherent time-span between the VR sensation and the administration of the questionnaire by the user, this might also impact the results being entered in the questionnaire, e.g., a stressful or frustrating event in the end of a VR session might have a disproportionate effect on users' judgment of their overall sensation. Fifth, questionnaires are unsuitable for most deployed or commercial VR applications, such as VR games, where the end-users' affective states shall be assessed. While certain groups, e.g., professional-level users, are motivated to contribute to the improvement of an application by providing feedback or administering questionnaires, most end-users, e.g., players of a VR game, have little motivation in investing time to administer a questionnaire. In related contexts such as customer surveys, response rates are often as low as 10% only, resulting in substantial non-response bias (Lambert and Harrington 1990). This non-response bias could lead to sample bias, i.e., the affective states of those users, who chose to administer a questionnaire are not representative of all users' affective states (Sivo et al. 2006). Despite elaborate techniques on questionnaire design, assessing the truthfulness of answers and detecting biases in reporting is difficult (Gillham 2007). Therefore, it is beneficial to use additional measures that provide at least some information on whether an answer deviates from the truth or not. Here, non-intrusive data collection would be favorable, since data can be collected without bothering users, which reduces effort and costs for researchers.

A recent addition to the measures obtained through questionnaires are implicit measures such as eye tracking, which record a participant's physiological response to the VR sensation. These methods are characterized through high precision and technical reliability, while eliminating some of the aforementioned shortcomings of questionnaires. However, eye tracking might come with other challenges, such as the VE scenery's brightness affecting pupil dilation (Pomplun and Sunkara 2003) or fast moving virtual objects affecting eye gaze (Vidal et al. 2013). Besides privacy concerns, it is questionable whether end-users are willing to carry out the required calibration procedures without receiving any obvious benefits. Furthermore, many low-cost end-user VR devices, such as the Oculus Quest 2, are not yet equipped with eye tracking devices. Consequently, while eye tracking is without doubt a valuable tool, it also adds technical complexity and end-user acceptance might be low.

Acknowledging the challenges related to questionnaires and currently available implicit measures, when assessing users' affective states, we argue that using behavioral characteristics to assess affective states in VR becomes even more appealing. That is, it would be highly desirable to obtain information on a user's affective state based on behavioral data without or in addition to administrating a questionnaire. The behavioral data can be recorded either through the VR system's tracking capabilities or dedicated motion tracking technology, such as motion suits. Through these means, a detailed data set of a users' full body movements can be carried out in all six degrees of freedom. However, our aim is to provide a method that can be applied to low-cost, end-user VR technology and thus be also useful beyond research settings, e.g., for developers of end-user VR applications. Therefore, we intentionally focus on head movements, which have been deemed readily interpretable and thus favorable compared to other body parts in the context of experimental psychology (Yaremych and Persky 2019). We expect head movement to be motion that is mostly affected by a user's affective state, compared to, e.g., hand movements which are predefined by the given task in the VE. In contrast, head movements can be recorded by most VR devices available today and have been related to certain affective states in prior research, e.g., Slater et al. (1998) and Won et al. (2016). Acknowledging the issues related to questionnaires and implicit measures, as well as the potentials of behavioral data, we develop the following research question: *Do VR users' head movements indicate their affective states?*

In addressing this research question, we contribute to the body of knowledge through providing empirical evidence for the relation between affective states and user behavior using standard regression analysis. We also show how this relationship can be exploited using machine learning methods to predict a user's affective state.

We organize the remainder of this paper as follows: In Sect. 2, we outline the related work for this study, including relevant work from cognitive sciences, psychology, and affective computing. Subsequently, we describe the approach we developed for utilizing data to infer affective states of VR users in Sect. 3. In Sect. 4, we describe the user study in detail. In Sect. 5, we perform a regression analysis showing the relationship between behavioral characteristics and user judgment. We also utilize this relationship for two prediction tasks, i.e., to coarsely predict a user's affective state as well as to identify potentially wrong replies in a questionnaire. We proceed to discuss our findings in Sect. 6. Further, we outline future work in Sect. 7, while we conclude this paper in Sect. 8.

2 Related work

The related work encompasses papers that analyze user behavior in VR, as well as articles from other disciplines, such as affective computing and cognitive sciences, which utilize human motion analysis to infer affective states.

2.1 Body expression of affective states

A human's affective state refers to a large variety of both, short-term phenomena, e.g., emotion, as well as long-term phenomena, e.g., mood (Karg et al. 2013). When the human's affective state changes, a corresponding body expression can be observed (Mehrabian and Friar 1969; Wallbott 1998). According to Ekman and Friesen (1967), this body expression could be due to either a neurophysical link, the verbalization of an affective theme, or a discharge mechanism, aiding to cope with the affective state. In the past, research has intensely focused on facial expressions and speech to infer affective states (Karg et al. 2013). However, there is strong evidence that other nonverbal expressions such as postural changes are at least as revealing (Kleinsmith and Bianchi-Berthouze 2013). In particular, head movement plays a key role as it can be utilized to infer various affective states, such as satisfaction, arousal, and interest, i.e., people tend to tilt their head, when they are interested (Noroozi et al. 2019).

2.2 Behavioral analysis in VR

The interpretation of users' behavioral data to infer user specific cognitive phenomena, such as affective states or traits has yet gained little attention among VR researchers. However, studies were published recently, which utilize human motion analysis in VR for the cases of social anxiety identification (Won et al. 2016), spherical video streaming optimization (Wu et al. 2017), behavioral biometrics (Pfeuffer et al. 2019), and trait prediction (Mu et al. 2020).

Won et al. (2016) investigated on identifying the affective state social anxiety among VR users in a virtual classroom setting. They found that head rotation correlates with users' social anxiety, i.e., users who scanned the virtual classroom more intensely also showed higher levels of social anxiety. Wu et al. (2017) recorded a user's head orientation during spherical video streaming and found that head movement patterns in VR are both content- and user-specific. Furthermore, they describe various applications that could benefit from head motion analysis, such as reducing the required bandwidth for streaming VR content by predicting a user's gaze, and user identification. Pfeuffer et al. (2019) developed a user identification mechanism in VR, which yielded a 30% accuracy out of 19 participants for overall head motion. However, head rotation alone has an identification accuracy between 25 and 35%. Another application of human motion analysis in VR is studied by Mu et al. (2020), who infer user traits, e.g., gender and age, based on head motion and eye tracking data.

2.3 Self-assessment of affective states

Affective states are commonly evaluated through self-assessment questionnaires (Marín-Morales et al. 2020). In the context of VR research, commonly employed questionnaires—assessing the users' affective states—include the System Usability Scale (SUS) (Brooke 1996), NASA TLX (Hart and Staveland 1988), and Presence (Slater et al. 1994). Yet, few prior works have investigated on the correlation between these commonly employed questionnaires and head movements. This includes papers from the field of VR, but also research in other domains, such as a real world office setting. Slater et al. (1998) investigated on the relation of body motion and presence in a VE. They found that users' sense of presence correlates with their motions. Thus, users who are more present in a VE show larger amounts of movement. In turn, this greater amount of movement further increases their sense of presence. More recent work has devoted attention to workload assessment based on human body posture (Qiu and Helbig 2012). Further assessment in an office setting has shown that the head movement data of office workers can be utilized to assess their task load (Chen and Epps 2019a, b). Regarding usability, Harms (2019) have investigated on the automated assessment of a VE's usability, based on the logging of user actions, e.g., a user grabbing virtual objects. Furthermore, Jacob and Karn (2003) have investigated on inferring a system's usability from eye tracking data.

3 Inferring affective states in VR based on behavioral data

Traditionally, researchers in the field of VR and designers of VR applications assess users' affective states through standardized self-assessment questionnaires. Thus, they apply a methodology, which requires a certain sample of users to undergo the following steps in a controlled setting:

1. Self-assessment on demography
2. Exposure to a VE
3. Self-assessment on affective states through standardized questionnaires
4. Analyze self-reports including assessing correctness of users replies; commonly, using statistical tests such as t tests

Our approach utilizes mainly behavioral data to infer VR users' affective states. External information, such as theoretical knowledge on the relation of behavior and affective states, as well as other empirical studies or potential moderating factors related to demographics (culture, age and gender) might be leveraged but are not necessary. For example, we might observe that a user moves more quickly than the average user. Given sufficient empirical and theoretical evidence that fast moving users exhibit a higher degree of presence than slow moving users (see Slater et al. 1998), we might conclude that the user under investigation is also likely to experience or feel a similar degree of presence. However, a precise inference of affective states is difficult, since most VR applications differ considerable in terms of tasks, composition of the VE, and user groups, making it difficult to reuse detailed quantitative information from prior studies. Thus in the initial phase, the approach utilized in this paper still requires the administration of self-assessment questionnaires—reflecting users' affective states—to at least a subgroup of users. The data from the subgroup is then utilized to (a) test basic relationships between behavioral data and questionnaire replies and assess whether they are conforming with established works, (b) utilize behavioral data and replies to obtain prediction models, possibly based on black-box machine learning models. Consequently, the initial phase consists of the following steps in a controlled environment with a certain sample of participants:

1. Self-assessment on demography
2. Exposure to a VE, where behavioral data is collected.
3. Self-assessment on affective states
4. Assess validity of relationship between behavioral data and self-assessed affective states
5. Train prediction model using self-assessments and behavioral data

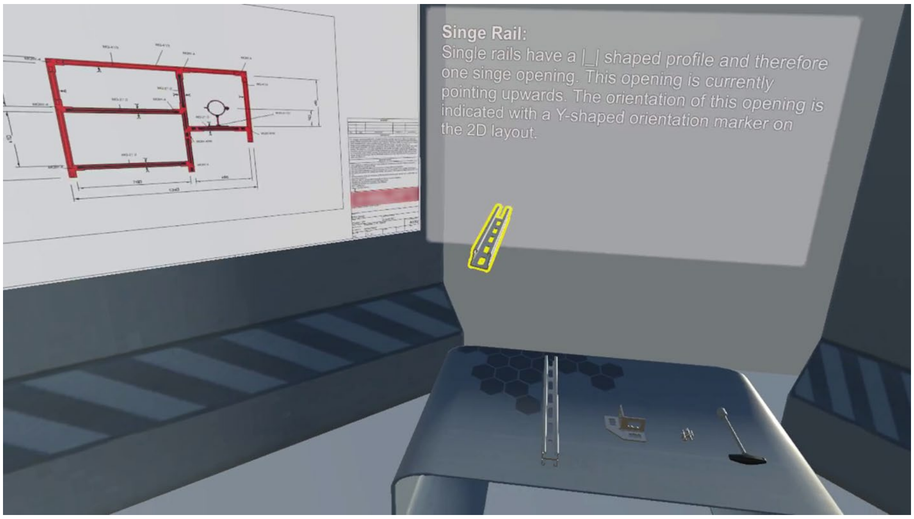
If the prediction model is sufficiently trained, a real-time system assessment of users' affective states can be established. This second phase does not require a specific, additional user action anymore, e.g., administering self-assessment questionnaires. It can even be carried out without the user being aware of it, though for ethical and legal reasons user consent should be obtained:

5. Exposure to a VE, where behavioral data is collected
6. Real-time inference of affective states with prediction model based on behavioral data
7. Real-time system adaption (optional, if required)

More concretely, in this study we identify relationships from literature that relate behavior to affective states. Here, we consider head movements—particularly head rotation—as highly promising (see Slater et al. 1998; Won et al. 2016; Pfeuffer et al. 2019). Furthermore, head rotation can be recorded by nearly any VR system, which ensures the wide applicability of the approach and independence from specific VR hardware. Behavioral characteristics, often referred to as features, in particular, in machine learning, are computed based on raw time series data after standard pre-processing, e.g., cutting off initial recordings due to calibration. They can be obtained from prior research, human expertise in modeling and domain knowledge as well as (automated) feature extraction from machine learning, using complex and data-demanding techniques such as deep learning. Currently, understanding these models still poses many challenges Meske et al. (2020). Well-interpretable, simple features, such as average accelerations and speeds, as well average displacements are preferable to ensure validity. For example, features that had been used in prior research are the mean and standard deviation (SD) of head rotation around the vertical axis, also referred to as head yaw (Won et al. 2016).

We conduct the first three steps of our approach through a user study, wherein users' head yaw during exposure to a VE is measured. Subsequently, each user self-assesses their affective state through questionnaires (TLX, SUS, presence). Third, we compute features based on the participants head orientation data.

In the fourth step of the approach, we investigate if there is a relationship between the computed features and the users' affective states represented through questionnaire scores. To this end, we perform a regression analysis to assess if there are significant and strong correlations between the features extracted from users' head yaw and questionnaire scores, such



Singe Rail:
Single rails have a |_| shaped profile and therefore one singe opening. This opening is currently pointing upwards. The orientation of this opening is indicated with a Y-shaped orientation marker on the 2D layout.

Fig. 1 Setup in the VTE including table with assembly components (lower right), assignment board (top right), and 2D plan of final assembly (top left)

as presence, usability and TLX items, such as physical and mental demand. This implies that we need a mathematically concise definition of behavioral characteristics to assess their relevance and explanatory power.

We then proceed to investigate, if the identified measures and relationships are useful for prediction tasks. While establishing a connection between behavioral data and affective states is interesting for the conceptual understanding of theory building, we are also interested whether observational data can be used for predictive tasks. For example, adaptive systems might monitor user behavior and act upon user sentiments. For instance, if a user seems to be bored during a training task in VR, the system might choose a more difficult task. To this end, we aim to estimate at least coarsely users' affective states. That is, we predict, based on behavioral data, whether a user is above the median in terms of presence, physical and mental demand or not. Here, we choose simple and well-interpretable prediction models that only require few data samples, i.e., linear regression, regression tree, and decision tree. The prediction of users' affective states is particularly relevant in the context of adaptive systems or when it is difficult and costly to utilize self-assessments. However, even if self-assessment questionnaires can be administered, the developed approach of this paper could be useful to support the identification of possibly incorrect answers, i.e., detect a mismatch between users' self-assessed affective states and their behavior in a VE. We investigate this under the assumption that the majority of users reply correctly. That is, we intentionally manipulate user judgments in the data set and assess if we can identify the alterations.

4 User study

The goal of the user study in this paper is to create a data set that allows to illustrate and validate the developed approach. The data set consists of recorded head yaw during exposure to a VE and corresponding user self-assessed affective state, such as TLX for each study participant.

4.1 Virtual training environment

The VE that the study participants were exposed to is a VTE for the manual assembly of modular support systems, which was developed in collaboration with an industry partner. The VTE consists of three steps that each participant fulfills to complete a training session. All training steps have a similar setup, which is shown in Fig. 1. This setup consists of a table with assembly components on the lower right-hand side, a board with written instructions on the top right-hand side, and a 2D plan of the final assembly on the top left-hand side. In a first step, participants familiarize themselves with the components to be assembled. In the second step, participants are instructed to identify the correct orientation of the assembly components according to the 2D plan. In the third step, the assembly logic is presented to the participants. Throughout the entire session, interaction with the VTE is limited to proceed through the training session by using the trackpad on the HTC Vive controller.

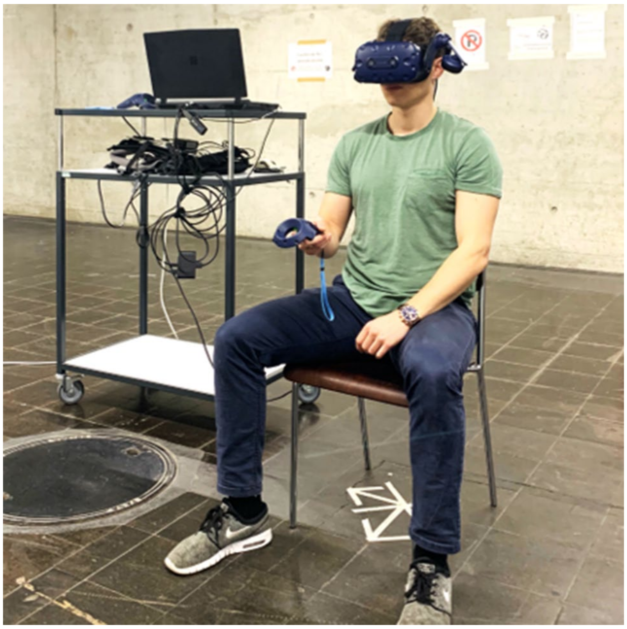


Fig. 2 A study participant wearing the HMD and holding one controller to interact with the VE, while sitting on a chair

4.2 Participants

Participants were recruited through public posters and online announcements. To enroll for the study, participants had to comply with two requirements. First, a basic understanding of English was required, as the experiment was conducted in English. Second, the participants had to be unfamiliar with the content of the training application that they would be exposed to. Forty-two participants (26 male, 16 female) enrolled, with an average age of 25.26 (SD = 3.44) years. All participants had normal or corrected to normal vision. Thirty-eight participants were university students, on bachelor's, master's or PhD level, whereas 5 participants were employees. Thirteen participants were using VR technology for the first time, 27 participants had used VR technology before for less than five hours, while 2 participants had more than five hours of VR experience. All participants received a financial compensation for enrolling in the user study and additionally a performance-based reward to maintain their motivation throughout the entire study session.

4.3 Apparatus

The experimental setup is shown in Fig. 2 and consists of the HTC Vive Pro VR System, including a Head-Mounted Display (HMD), two controllers, and four Steam VR 2.0 base stations for tracking. The VTE was created using the Unity 3D Engine.[1] During the VR session, participants were sitting on a chair, while using the controller only to proceed through the application. This type of setup has proven to be effective to infer users' affective states in another study, which investigated the context of a VR science lab simulation (Makransky et al. 2019).

4.4 Measures

Within each step of the user study, measurements were taken. Before the VR exposure, participants filled out a questionnaire regarding basic demographic questions, such as gender, age, and prior VR experience. During VR exposure, the participants' head yaw was recorded in the form of time-stamped trajectories with a sampling rate of 2 Hertz. After VR exposure, each participant filled out the Presence (Slater et al. 1994), NASA TLX (Hart and Staveland 1988), and SUS (Brooke 1996) questionnaires.

4.5 Procedure

Each study session was conducted with one participant and at least one of the authors of this paper, who led the session. First, the participant was welcomed to the study and informed about its procedure. Furthermore, the participants read and signed a consent form, wherein they declared their agreement to the terms of the study, as well as that they were informed about the risks. Subsequently, each participant filled the demography questionnaire on a digital form on a computer that was provided by the session leader. After the demography questionnaire, the participants were prepared for VR exposure by the session leader, which involved explaining the controllers, as well as putting on the HMD. Each VR session started with a tutorial, wherein participants familiarized themselves with the VTE and the controls. Subsequently, the participants were exposed to the VTE, wherein their head yaw was recorded. After the VR session, the participants proceeded to the final questionnaires, wherein they disclosed their judgment of their VR experience. Finally, the participants received their payment and left the study session. We note that the user study has been conducted between December 2019 and February 2020.

5 Results

After providing descriptive results of the user study, we provide outcomes of the regression analysis to estimate the relationship of head yaw on the questionnaires' results, which reflect the participants' affective states. In turn, we used the discovered measures to derive a prediction model to assess the suitability of the relationship to perform predictions.

[1] https://unity.com/.

Table 1 Descriptive statistics

Measure	Range	Median	Mean	Std. deviation
Presence	1–7	4.75	4.79	1.07
Mental demand	0–10	3.0	3.14	2.01
Physical demand	0–10	1.0	.86	1.0
Temporal demand	0–10	2.0	1.88	1.77
Perceived performance	0–10	9.0	8.17	2.02
Effort	0–10	2.0	2.60	1.93
Frustration	0–10	2.0	2.36	2.5
Usability	0–100	83.75	82.68	11.02

5.1 Descriptive results

In the user study, we collected eight self-assessed measures (see Table 1). The results for presence were on average 4.75/7 (SD = 1.07), which we consider reasonable. This is also confirmed by other recent VR studies yielding presence values in the same range (see Chang et al. 2019; Zenner et al. 2020). The average mental demand was 3.14/10 (SD = 2.01), which is rather on the lower end. The elevated SD for mental demand could be explained by the fact that each user was perceiving the VE differently, which is due to a combination of individual traits (Jensen and Konradsen 2017). This is further supported by subjective feedback from the user study, where some participants reported that they had to concentrate considerably while being exposed to the VE, while others reported that they did not have to concentrate at all. The results for the average physical demand were 0.86/10 (SD = 1.00), which is unsurprisingly low. As the participants proceeded through the VE in a seated setup and only interacted with the VE by pushing the trackpad on the HTC Vive's controller, they were limited to mainly head movements. This could be also one of the reasons for the participants' low average effort of 2.60/10 (SD = 1.93). The average temporal demand was 1.88/10 (SD = 1.77), which is also rather low. This could be explained by the VE's design, as the participants were able to navigate through the training in their own pace without any time constraints. The average perceived performance was 8.17/10 (SD = 2.02), which is on the upper end, implicating that participants felt successful in completing the virtual training session. This can be well explained through the VE's task, which required solely to push the HTC Vive's controller trackpad to proceed through the application. Consequently, there was little possibility for wrong actions that might decrease a user's perceived performance. Finally, the participants' average frustration was 2.36/10 (SD = 2.5), which was also rather low. However, the SD for frustration is considerable, which could be explained by the impossibility to return to a previous step during the training, since some participants provided the subjective feedback on being frustrated about not being able to return to a previous step in the VE. The VEs average SUS was 82.68/100 (SD = 11.02), which can be generally classified as good usability (Bangor et al. 2009).

5.2 Regression analysis

We conducted a regression analysis to investigate on the relationship of head yaw and the users' affective states, measured through the questionnaires. As the dependent variables, we use the following head movement features computed from head yaw, which can be also described as the head's rotation around the vertical axis (z-axis): mean (Me) and standard deviation (Cv) of angular displacement Won et al. (2016), as well as mean of the angular speed (Vh), which was also used in Pfeuffer et al. (2019). We then created a linear regression model for each of the independent variables: presence, the six NASA TLX items (mental demand, physical demand, temporal demand, perceived performance, effort, and frustration), and the SUS. These eight linear regression models are shown in Table 2.

Results from the regression analysis indicate our metrics, the dependent variables, to be significant predictors in five models, i.e., mental demand, physical demand, perceived performance, presence, and usability. The exceptions are models for temporal demand, effort, and frustration. Possible reasons on why these three models did not show significant relationships could be that there is only a weak relationship being significant for larger sample sizes or that there is simply no relationship, e.g., due to the setup of the task. In our study, the self-paced nature of the task might reduce the relevance or severity of temporal demand (also indicated by a low average). We found that for four out of five of the independent variables at least two dependent variables showed a significant influence. The adjusted R^2 of the regression models shows that head movement features are shown to be able to explain up to 25% of the variance in users' answers. In general, users' head movement features appear to be most predictive of users' perceived performance in the task done in the VTE. The models also showed indications of relationships between the metrics with usability, mental and physical demand, on lesser degrees. Presence is shown to only have a weak relationship with the selected metrics. While this might be due to the metric selection, another factor could be that perception of presence among users varies more than their perceptions on other measure. For example, users might agree more on the meaning of a 3 out 10 in terms of physical demand, rather than what 3 out of 7 in terms presence means. Further discussions of the five models are presented in the following paragraphs.

Mental demand Users who let their gaze wander to cover a wide area of the VE, e.g., to find relevant information, and those who moved their head slowly indicate high mental demand. Covering a larger area, i.e., a high Cv value,

Table 2 Regression relating head movement features and users' questionnaire scores

Variables	Mental demand		Physical demand		Temporal demand	
	Coeff.	*P* value	Coeff.	*P* value	Coeff.	*P* value
Intercept	3.264	0.000***	1.000	0.002**	2.144	0.000***
Me	−0.014	0.879	0.045	0.344	−0.141	0.107
Cv	0.868	0.010*	0.421	0.013*	0.318	0.286
Vh	−3.364	0.006**	−1.722	0.005**	−1.547	0.153
R^2	0.198		0.190		0.138	
Adj. R^2	0.135		0.127		0.070	
AIC	175.4		117.4		167.9	
BIC	182.4		124.4		174.9	
Variables	Perceived Performance		Effort		Frustration	
	Coeff.	*P* value	Coeff.	*P* value	Coeff.	*P* value
Intercept	8.316	0.000***	2.772	0.000***	2.258	0.010*
Me	−0.297	0.002**	0.035	0.718	0.042	0.741
Cv	−0.878	0.006**	0.445	0.192	0.529	0.235
Vh	2.970	0.010*	−1.856	0.132	−1.819	0.257
R^2	0.299		0.059		0.039	
Adj. R^2	0.244		−0.015		−0.037	
AIC	170.5		178.7		201.4	
BIC	177.4		185.6		208.3	
Variables	Presence		Usability			
	Coeff.	*P* value	Coeff.	*P* value		
Intercept	4.760	0.000***	82.075	0.000***		
Me	−0.111	0.043*	−0.991	0.061		
Cv	−0.135	0.461	−4.505	0.015*		
Vh	0.501	0.449	17.123	0.011*		
R^2	0.106		0.198			
Adj. R^2	0.036		0.134			
AIC	127.4		318.5			
BIC	134.3		325.5			

relates to having to digest and/or gather more information, which leads to more information to process and in turn to an increase in mental demand. A higher angular speed (Vh), indicates that users were likely more certain where they want to direct their attention to. Such users are then more likely to report a lower mental demand, as they were more certain throughout the task than users who moved slower.

Physical demand The coefficients of predictors for physical demand show a similar pattern like for the mental demand. That is a positive coefficient for coverage (Cv) and a negative coefficient for angular speed (Vh). In this case, speed is less negatively related, which is intuitive, since in fact a larger speed might have caused an increased physical demand. On top of that, the range of the area covered by head tilting, as indicated by a high Cv value, seems less important in predicting physical demand than mental demand. The model for physical demand, however, shows a lower adjusted R^2 value than the one for mental demand. This could be due to the fact that physical demand may be better reflected in other types of movements, e.g., torso or limb movements, whereas increased head movements might reflect the mental activity of the users.

Perceived performance Users who move their head faster (large Vh) may reflect their certainty in what they are doing. This resulted in an increased confidence that they did well on the given task. A negative impact from having a large coverage (Cv) is the need to search for information extensively, which is indicating uncertainty. While having the need for more information is not necessarily negative, it might still indicate feeling more challenged, i.e., being less optimistic on their success. Finally, the users' average tilting direction (Me) is also shown to be a significant predictor for perceived performance. The negative coefficient indicates that users,

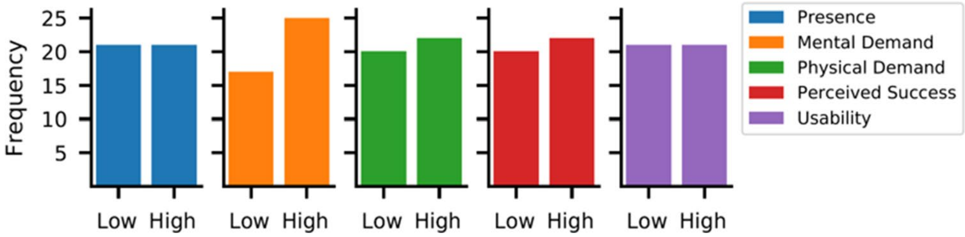


Fig. 3 Distribution of response groups for each measure

who tilted more to the right reported lower perceived performance. This could be a consequence of our VE setup, where the instruction board and table for components are located to the right of the 2D plan of the final assembly and the final assembly itself. Users who looked more to the right relative to their peers could be the ones spending less time on the task assembling than reading instructions or examining components on the table, and reported higher perceived performance.

Presence Our regression model indicated that the *Me* is a significant predictor for their sense of "being there," i.e., the presence score. The negative sign of the coefficient indicates that users who tilted more to the right tend to report lower presence score. The corresponding coefficient of −0.111 shows that one degree of yaw towards the right contributes to a lower presence score of about a tenth of a point in the presence. Thus, users who on average tilted their heads ten degrees to the right-of-center, are estimated to report one point lower in presence compared to those who tilted their head left and right equally, i.e., their averaged tilting would be at the center. As shown in Table 1, presence scores range from 0 to 7 with a mean and standard deviation of 4.79 and 1.07, respectively. This implies that a user, who tilted 10 degrees more towards the right shoulder than the average user is likely to report a presence score one point lower than the average user. With the training setup within the VE having points of interest on the users' left-hand side and none on their right-hand side, tilting towards the right shoulder might indicate users' decreased attention and missing involvement during the training.

In a review by Yaremych and Persky (2019), findings from studies which used tracing data of physical behavior, such as ours, could be influenced by the specifics of the VE used for the study. Thus, while there might exist general relationships between behavioral data and affective states, they must be carefully assessed in multiple settings to allow for generalizations. In the following paragraphs, we set forth a few possible explanations and relationships that might hold beyond our study. These explanations could serve as starting points for future studies on tracing physical behavioral data, in particular that of head yaw.

Usability A large positive coefficient on *Vh* possibly indicates that users who moved faster are more at ease with using the VE. Slower movements, on the other hand, could be due to users having difficulties in processing information presented through the VE. This is aligned with the negative coefficient on coverage. That is, users' who felt less need to look around and gather information within the VR would deem that the system has high usability.

5.3 Prediction task

We evaluate the suitability of the features extracted from head yaw data in predicting coarsely users' questionnaire responses as follows. First, we selected the questionnaires, i.e., the independent variables, of interest based on our regression in Sect. 5.2. That is, we took the five independent variables which are shown to have significant predictors: presence, mental demand, physical demand, perceived performance, and usability. We then divided the response values into two groups based on the median. This is done in order to achieve groups with balanced sizes. The distribution of the resulting groups, namely less than median value and greater or equal to median value are presented in Fig. 3

To evaluate the prediction models, we employed the leave-one-out cross-validation approach. That is, we leave one sample out, train the model with the rest $n - 1$ samples, and test the model using the left-out sample. This procedure is done for each sample, i.e., n times. We report the cross-validation accuracy, i.e., the average accuracy of all n samples.

Our evaluation included two methods to perform the prediction task, namely linear regression and decision tree. For the decision tree we used the scikit-learn (Pedregosa et al. 2011) implementation. We set the maximum depth at 3 to improve the model's simplicity, and left the other hyperparameters at their default setting. A commonly applied

Table 3 Prediction task

Measure	Baseline (%)	Linear regression (%)	Decision tree (%)
Presence	50.0	47.6	52.4
Mental demand	59.5	52.4	64.3
Physical demand	52.4	59.5	66.7
Perceived performance	52.4	66.7	52.4
Usability	50.0	50.0	61.9

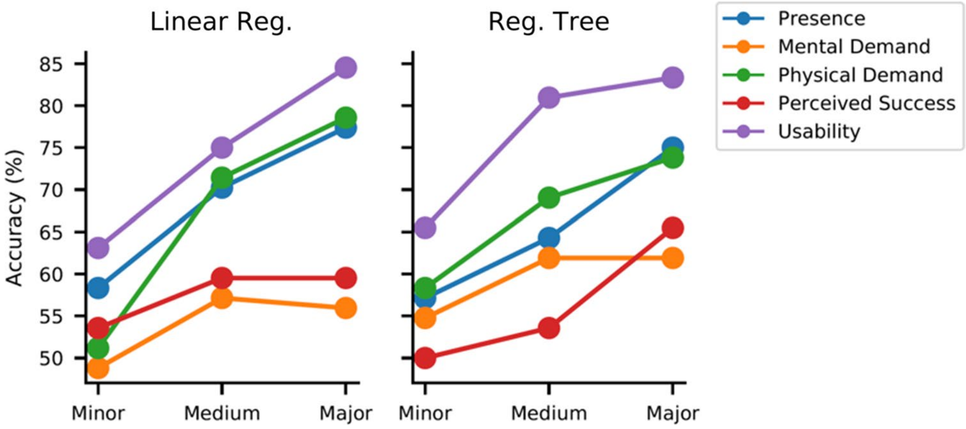


Fig. 4 Results of the detection task

baseline to infer whether a user belongs to one group or the other is to guess based on the proportion of group sizes. That is, every user is predicted to belong to the majority group. We used this baseline to evaluate the performance of both prediction models.

Outcomes Table 3 shows the cross-validation accuracy for each model on each measure. Overall, the decision tree performed best. It outperformed the baseline for 4 out of 5 measures, whereas the difference was larger than 10% for physical demand and usability. For perceived performance, simple linear regression did better than the baseline and clearly outperformed the decision tree. This outcome is aligned with our regression analysis, where the model on perceived performance had the highest adjusted R^2 value. Presence and mental demand did not yield satisfactory prediction accuracy compared to the baseline. It is also no surprise that no single model performs best at predicting all five measures given the "no free lunch" theorem (Wolpert and Macready 1997), saying that there is no single best model for all tasks.

5.4 Detection task

While most users administer questionnaires genuinely, some might not answer properly (Mertens et al. 2017). This manifests in the intentional or unintentional provision of incorrect answers, which decreases data quality and could even lead to drawing wrong conclusions. While the intentional provision of wrong answers can be related to users' low motivation to invest time and effort, e.g., in a crowdsourcing setting participants might be monetary driven and data screening might be needed to eliminate poor responses (Chmielewski and Kucker 2019). However, even the motivated users might unintentionally provide wrong answers. For example, users might exhibit an optimism bias that manifests in answers by underestimating mental effort and overestimating their performance. That is, users might require a lot of mental effort to address a task in a VE, but report low scores. Furthermore, users might also interpret scales differently. This would result in some users answering, such that the answers across many questions are Gaussian distributed, i.e., only few answers will consist of the minimum or maximum value. Other users might tend towards more binary answers, where answers tend to lean more towards extreme values. Some users might also exhibit central tendency bias.

To assess the ability to detect manipulated answers, we assume that (most) users answered correctly. That is, if we alter answers of users and there is a relationship between behavior and users' self-reported affective states, we should be able to detect it. Accordingly, we assess if we can detect when the actual reported answer of a user is altered. The smallest unit of possible change depends on the range of the VR measure. We defined it as follows:

- Presence scores (ranged 1–7): 1 unit corresponds to 1 point.
- Mental Demand, Physical Demand, and Perceived Performance scores (ranged 0–10): 1 unit corresponds to 1 point.
- Usability scores (ranged 0–100): 1 unit corresponds to 15 points.

The difficulty of this task depends heavily on the magnitude of manipulations, e.g., changing an answer from score of 10 to 0 is much easier to notice than from 10 to 9. Therefore, we investigate three levels of manipulation: (1) minor: answers are increased or reduced by the smallest amount, i.e., just one unit; (2) medium: changed by two units, and (3) major: changed by three units.

It is also ensured that the manipulated answers are neither above the maximum possible value nor lower than the minimum possible value, e.g., the smallest possible is always increased.

To assess if a user's answer is manipulated, we assess the prediction of a model (trained without the user data) and compute the difference in prediction and user answer.

We classify an answer as incorrect if the prediction error is larger than a threshold. For each user, we generate one manipulated answer and also use the actual reported answer. That is, there is an equal number of manipulated answers and "truthful" answers. A "guessing" approach would therefore obtain 50% accuracy. Two types of models are evaluated for this detection task, namely linear regression and regression tree. For the regression tree, we kept the maximum depth at three to maintain model's simplicity. Figure 4 shows the cross-validation accuracy of both models for each manipulation level. In general, the larger the manipulations, the better the models perform better than guessing, i.e., 50%.

Results Both models behave qualitatively similar. Minor manipulations could only be detected clearly above baseline for presence and usability. However, the larger the manipulations the higher the accuracies that can be yielded. For large manipulations, the accuracy ranged from about 60% up to 85% depending on the measure and model. This indicates the suitability of the approach, in particular, since the data used to infer the model might be "noisy," i.e., contain user replies that are wrongful but assumed to be correct. It is also no surprise that perceived performance and mental demand are the most difficult to predict. Table 1 one indicates the standard deviations of responses. There, these two measures exhibit relatively (compared to their min-max possible reply) the largest standard deviation. That is, these replies are most "noisy," indicating that it is fairly plausible that a user's reply might fluctuate within the order of manipulation. In turn, this makes it difficult to distinguish manipulation from natural variation.

6 Discussion

Our work is among the first to utilize behavioral data in VR to assess users' affective states measured through common questionnaires (TLX, SUS, presence). Therefore, more empirical evidence is needed that can be combined into general theories that allow to relate general behavioral patterns to affective states.

While our approach to derive affective states based on behavioral data has clear advantages, we believe that for some time to come, self-assessment questionnaires might still be valuable, if not preferable. As stated in prior work (Yaremych and Persky 2019), the choice of appropriate measures and interpretation of behavioral data is highly context dependent. Computing measures using head rotation is appropriate for our task, as it requires the users to largely stay in one spot while surrounded with relevant virtual objects, e.g., instruction boards and 2D final assembly plan. Head positions, on the other hand, could be more appropriate in deriving affective states, where users are required to walk between different workstations. Furthermore, the VE and task also affect the interpretation of results. For example, for a metric one used in our study, i.e., mean of the angular displacement (Me). Results akin to "users who on average looked more towards the left reported higher perceived performance," need to be put into context by considering questions such as: What virtual objects are on the users' left? Does the task require more attention on the area left to the users than other areas? A more subtle impact of the task design should also be considered. Our task requires users to switch between the different virtual objects throughout. As such, the SD of angular displacement (Cv) and mean of the angular speed (Vh) are useful proxies for users' activity. It could be plausible that for tasks, which require users to focus on a specific point in the VE, a large Cv could mean that they are distracted. As such, some metrics are more generalizable than others, e.g., Cv and Vh is less dependent on the absolute locations of the virtual objects as compared to Me. On the other hand, the more context dependent metrics may still provide value by informing us how the users solve a particular task and/or behave in a particular VE.

Qualitatively, our work is also aligned with work dating more than two decades back, e.g., Slater et al. (1998), who found that presence correlates positively to Vh in a VE. In our case, correlation is only fairly weak. While this might be attributed to the task at hand, it might also be due to technological difference, i.e., today's VR technology invokes much higher presence values. To test the significance of relationships between behavioral characteristics and affective states, our work utilizes both a "classical" approach based on linear regression and statistical measures, such as P values as well as a more modern "machine learning" based approach to assess the usefulness of the data based on prediction accuracy. The overall results are mostly aligned, e.g., prediction performs well given stronger statistical relationships.

Another aspect that is not relevant for this study, but could be generally important for other studies identifying affective states based on behavioral measures, is the cultural background of the participants. In this context, prior research has identified a significant influence of culture on the recognition of emotion from body posture (Kleinsmith et al. 2006). Therefore, studies involving emotionally arousing scenes, should consider balancing out the participants' cultural backgrounds within their sample.

7 Future work

Future research should focus on conducting more studies utilizing behavioral data to infer users' affective states. These studies could vary the context of the VE in terms of: setup (e.g., standing, real walking), the task (e.g., high mental demand, low usability), participants (e.g., other age groups). Furthermore, it might be promising to combine behavioral data with other implicit measures, such as eye tracking, to strengthen the correlation and prediction accuracy. For larger sets of participants and higher sampling rates, machine learning techniques, such as deep learning methods might become feasible to extract features, using supervised, as well as unsupervised techniques. While understanding these models is still a concern Meske et al. (2020) hindering their applicability in scientific fields, more and more models are being developed with interpretability in mind, e.g., Fusco et al. (2019). Ultimately, these efforts could support adaptive VR systems, which synchronously measure and influence users' affective states, e.g., if a user's mental demand in a VTE is too low, the system could react by automatically increasing the difficulty. Thus, the human sensory data from VR serves as input to a machine learning model, which alters the VE. In the more distant future, human-to-AI coaches Schneider (2020) might even provide feedback to a user, so that (mutual) understanding of the machine learning component and the human improves.

Utilizing behavioral data could also be used to shed new light on the correlation between presence and performance, which has been a controversial topic among VR researchers for decades (Barfield et al. 1995). In particular, in the vocational use of VTEs, using behavioral data to predict performance would be immensely beneficial, since traditional alternatives are often laborious and costly (Gisler et al. 2020). The fact that similar approaches for analyzing the usability of a VE yielded promising results (Schroeder et al. 2006) should further motivate this endeavor.

8 Conclusion

In this paper, we mark initial steps towards the assessment of VR users' affective states based on their behavioral data. We show significant correlations between users' head yaw and their self-assessed affective states in a VE, i.e., presence, mental demand, physical demand, perceived performance, and usability. Additionally, we demonstrate that head yaw can be further utilized to predict coarsely a user's affective state in a VE and to detect incorrect reporting of user judgment.

Acknowledgements This work was supported by a project financed by the Hilti Family Foundation in Schaan, Liechtenstein and RhySearch in Buchs SG, Switzerland. Additionally, the authors would like to thank Hilti AG for the fruitful collaboration and the equipment provided.

Funding Open Access funding provided by ETH Zurich.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Bangor A, Kortum P, Miller J (2009) Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 4(3):114–123

Barfield W, Zeltzer D, Sheridian T, Slater M (1995) Presence and performance within virtual environments. In: Barfield W, Furness TA (eds) Virtual environments and advanced interface design. Oxford University Press, New York, pp 473–513

Brooke J (1996) SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL (eds) Usability evaluation in industry, chap 21. Taylor & Francis Ltd, London, pp 189–194. https://doi.org/10.1201/9781498710411

Chang CW, Yeh SC, Li M, Yao E (2019) The introduction of a novel virtual reality training system for gynecology learning and its user experience research. IEEE Access 7:43637–43653. https://doi.org/10.1109/access.2019.2905143

Chen S, Epps J (2019a) Atomic head movement analysis for wearable four-dimensional task load recognition. IEEE J Biomed Health Inform 23(6):2464–2474. https://doi.org/10.1109/JBHI.2019.2893945

Chen S, Epps J (2019b) Task load estimation from multimodal head-worn sensors using event sequence features. IEEE Trans Affect Comput 1–13. https://doi.org/10.1109/TAFFC.2019.2956135

Chmielewski M, Kucker SC (2019) An MTurk crisis? Shifts in data quality and the impact on study results. Social Psychol PersonSci 11(4):464–473. https://doi.org/10.1177/1948550619875149

Cummings JJ, Bailenson JN (2015) How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. Media Psychol 19(2):272–309. https://doi.org/10.1080/15213269.2015.1015740

Ekman P, Friesen WV (1967) Head and body cues in the judgement of emotion: a reformulation. Perceptual Motor Skills 24(3):711–724. https://doi.org/10.2466/pms.1967.24.3.711

Fusco F, Vlachos M, Vasileiadis V, Wardatzky K, Schneider J (2019) Reconet: an interpretable neural architecture for recommender systems. In: International joint conferences on artificial intelligence (IJCAI)

Gillham B (2007) Developing a questionnaire, 2nd edn. Continuum/Bloomsbury Academic, London

Gisler J, Hirt C, Kunz A, Holzwarth V (2020) Designing virtual training environments:does immersion increase task performance? In: 2020 International conference on cyberworlds (CW), pp 125–128. https://doi.org/10.1109/CW49994.2020.00026

Harms P (2019) Automated usability evaluation of virtual reality applications. ACM Trans Comput Human Interact 26(3):1–36. https://doi.org/10.1145/3301423

Hart SG, Staveland LE (1988) Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv Psychol 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hirt C, Holzwarth V, Gisler J, Schneider J, Kunz A (2019) Virtual learning environment for an industrial assembly task. In: 2019 IEEE 9th international conference on consumer electronics (ICCE-Berlin), IEEE. https://doi.org/10.1109/ICCE-Berlin47944.2019.8966169

Hirt C, Eckard M, Kunz A (2020) Stress generation and non-intrusive measurement in virtual environments using eye tracking. J Ambient Intell Human Comput: 1–13. https://doi.org/10.1007/s12652-020-01845-y

Jacob RJ, Karn KS (2003) Eye tracking in human-computer interaction and usability research. In: Hyönä J, Radach R, Deubel H (eds) The mind's eye. North-Holland, Amsterdam, pp 573–605. https://doi.org/10.1016/B978-044451020-4/50031-1

Jensen L, Konradsen F (2017) A review of the use of virtual reality head-mounted displays in education and training. Educ Inform Technol 23(4):1515–1529. https://doi.org/10.1007/s10639-017-9676-0

Karg M, Samadani AA, Gorbet R, Kuhnlenz K, Hoey J, Kulic D (2013) Body movements for affective expression: a survey of automatic recognition and generation. IEEE Trans Affect Comput 4(4):341–359. https://doi.org/10.1109/t-affc.2013.29

Kleinsmith A, Bianchi-Berthouze N (2013) Affective body expression perception and recognition: a survey. IEEE Trans Affect Comput 4(1):15–33. https://doi.org/10.1109/t-affc.2012.16

Kleinsmith A, Silva PRD, Bianchi-Berthouze N (2006) Cross-cultural differences in recognizing affect from body posture. Interac Comput 18(6):1371–1389. https://doi.org/10.1016/j.intcom.2006.04.003

Lambert DM, Harrington T (1990) Measuring nonresponse bias in customer service mail surveys. J Bus Logistics 11:5–25

MacKenzie IS (2013) Designing HCI experiments. In: Human-computer interaction. Morgan Kaufmann, Waltham, pp 157–189. https://doi.org/10.1016/b978-0-12-405865-1.00005-4

Makransky G, Terkildsen TS, Mayer RE (2019) Adding immersive virtual reality to a science lab simulation causes more presence but less learning. Learn Instruct 60:225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007

Marín-Morales J, Llinares C, Guixeres J, Alcañiz M (2020) Emotion recognition in immersive virtual reality: from statistics to affective computing. Sensors 20(18):5163. https://doi.org/10.3390/s20185163

Mehrabian A, Friar JT (1969) Encoding of attitude by a seated communicator via posture and position cues. J Consult Clin Psychol 33(3):330–336. https://doi.org/10.1037/h0027576

Mertens W, Pugliese A, Recker J (2017) How to start analyzing, test assumptions and deal with that pesky p-value. In: Quantitative data analysis. Springer, Cham, pp 135–156. https://doi.org/10.1007/978-3-319-42700-3_8

Meske C, Bunde E, Schneider J, Gersch M (2020) Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. Inform Syst Manage. https://doi.org/10.1080/10580530.2020.1849465

Mu M, Dohan M, Goodyear A, Hill G, Johns C, Mauthe A (2020) User attention and behaviour in virtual reality art encounter. arXiv:2005.10161

Newsted PR, Huff SL, Munro MC (1998) Survey instruments in information systems. MIS Quart 22(4):553–554. https://doi.org/10.2307/249555

Noroozi F, Kaminska D, Corneanu C, Sapinski T, Escalera S, Anbarjafari G (2019) Survey on emotional body gesture recognition. IEEE Transactions on Affective Computing pp 1-1, https://doi.org/10.1109/taffc.2018.2874986

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Pfeuffer K, Geiger MJ, Prange S, Mecke L, Buschek D, Alt F (2019) Behavioural biometrics in VR: identifying people from body motion and relations in virtual reality. In: Proceedings of the 2019 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, pp 1–12. https://doi.org/10.1145/3290605.3300340

Pomplun M, Sunkara S (2003) Pupil dilation as an indicator of cognitive workload in human-computer interaction. In: Smith M, Duffy V, Harris D, Stephanidis C (eds) Human-centered computing: cognitive, social, and ergonomic aspects, vol 3. CRC Press, Boca Raton, pp 542–546

Qiu J, Helbig R (2012) Body posture as an indicator of workload in mental work. Hum Factors 54(4):626–635. https://doi.org/10.1177/0018720812437275

Schneider J (2020) Human-to-AI coach: improving human inputs to AI systems. In: International symposium on intelligent data analysis. Springer, pp 431–443

Schroeder R, Heldal I, Tromp J (2006) The usability of collaborative virtual environments and methods for the analysis of interaction. Presence Teleoperat Virtual Environ 15(6):655–667. https://doi.org/10.1162/pres.15.6.655

Sivo S, Saunders C, Chang Q, Jiang J (2006) How low should you go? Low response rates and the validity of inference in IS questionnaire research. J Assoc Inform Syst 7(6):351–414

Slater M (2018) Immersion and the illusion of presence in virtual reality. Brit J Psychol 109(3):431–433. https://doi.org/10.1111/bjop.12305

Slater M, Usoh M, Steed A (1994) Depth of presence in virtual environments. Presence Teleoperat Virtual Environ 3(2):130–144. https://doi.org/10.1162/pres.1994.3.2.130

Slater M, McCarthy J, Maringelli F (1998) The influence of body movement on subjective presence in virtual environments. Hum Factors 40(3):469–477. https://doi.org/10.1518/001872098779591368

Vidal M, Pfeuffer K, Bulling A, Gellersen HW (2013) Pursuits: eye-based interaction with moving targets. In: CHI '13 extended abstracts on human factors in computing systems. Association for Computing Machinery, New York, pp 3147–3150. https://doi.org/10.1145/2468356.2479632

Wallbott HG (1998) Bodily expression of emotion. Eur J Social Psychol 28(6):879–896. https://doi.org/10.1002/(SICI)1099-0992(199811)28:6<879::AID-EJSP901>3.0.CO;2-W

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82. https://doi.org/10.1109/4235.585893

Won AS, Perone B, Friend M, Bailenson JN (2016) Identifying anxiety through tracked head movements in a virtual classroom. Cyberpsychol Behav Social Netw 19(6):380–387. https://doi.org/10.1089/cyber.2015.0326

Wu C, Tan Z, Wang Z, Yang S (2017) A dataset for exploring user behaviors in VR spherical video streaming. In: Proceedings of the 8th ACM on multimedia systems conference. Association for Computing Machinery, New York, NY, USA, pp 193–198. https://doi.org/10.1145/3083187.3083210

Yaremych HE, Persky S (2019) Tracing physical behavior in virtual reality: a narrative review of applications to social psychology. J Exp Social Psychol 85:103845. https://doi.org/10.1016/j.jesp.2019.103845

Zenner A, Makhsadov A, Klingner S, Liebemann D, Kruger A (2020) Immersive process model exploration in virtual reality. IEEE Trans Visual Comput Graph 26(5):2104–2114. https://doi.org/10.1109/tvcg.2020.2973476

Zhang W, Shu L, Xu X, Liao D (2017) Affective virtual reality system (AVRS): design and ratings of affective VR scenes. In: 2017 International conference on virtual reality and visualization (ICVRV), IEEE. https://doi.org/10.1109/icvrv.2017.00072

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.