Design and Validation of a Library of Active Affective Tasks for Emotion Elicitation in VR

Jason W. Woodworth* Christoph W. Borst†

CACS VR Lab
University of Louisiana at Lafayette

ABSTRACT

Emotion recognition models require datasets of physiological responses to stimuli designed to elicit targeted emotions, preferably stimuli similar to the experience during which the models will be used. Many libraries of such stimuli have been created to ease this data collection process, most of which involve passive media such as images or videos. Virtual Reality, however, offers an opportunity to investigate uniquely active emotion elicitation stimuli that directly center the user in the experience with an increased feeling of presence and potential to elicit stronger emotions. We leverage this to introduce a set of four active affective tasks in VR designed to quickly elicit targeted emotions without need for narrative understanding common to passive stimuli. We compare our tasks with selections from an existing affective library of passive 360° videos and validate our approach by comparing self-reported emotional responses to the stimuli. Results indicate that these types of active task stimuli can reliably elicit strong emotions comparable to other passive media and provide the basis for building a larger library of training- and education-relevant tasks.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Applied computing—Law, social and behavioral sciences—Psychology

1 INTRODUCTION

We are investigating the use of active affective task stimuli in VR as a tool for enabling emotion recognition. Automated emotion recognition has long been understood to be critical to building intelligent systems that can naturally interact with human beings [31]. Implications of accurate and generalizable recognition models are widespread: cars could detect and react to stressed [47] or drowsy [18] drivers, assisted living homes could adjust living conditions automatically [12], games could adapt difficulty based on player stress [13], and teachers could be alerted when students have trouble with material [14].

Emotion recognition models are built using large training datasets mapping sensed behavioral and physiological responses to felt emotions elicited by affective stimuli [37, 39]. To ease the burden of needing every researcher or developer to create unique stimuli for their dataset, many libraries and databases of affective stimuli have been created to be reused in future works. Traditionally, many of these libraries have been composed of passive media such as video [15] or images [20], relying on empathy or environmental reactions. While these have proven successful in some instances, they face several limitations. Relying on the user empathizing with a character or narrative may prove to be problematic for sensitive users or be culturally dependent [29], and relying on reactions to static environments may lead to less extreme emotions [24].

*e-mail: jason.woodworth1@louisiana.edu
†e-mail: cwborst@gmail.com

This is an author-formatted version. Original publication: J. W. Woodworth and C. W. Borst, "Design and Validation of a Library of Active Affective Tasks for Emotion Elicitation in VR," 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR), Orlando, FL, USA, 2024, pp. 398-407, doi: 10.1109/VR58804.2024.00061.

At the same time, virtual reality is increasingly recognized as an effective tool for eliciting emotion [25, 39]. The use of VR imposes some unique challenges and benefits. For example, the use of a headset restricts visual access to the face, making facial expressions more difficult to read and reducing the ability to remotely measure heart rate through cameras [27]. However, virtual reality's immersive nature is known to increase users' felt presence in the experience, leading us and many researchers to believe it can also heighten emotional responses [21, 22], making up for these limitations.

However, despite the inherently active nature of VR, many recent works on VR for emotion elicitation still present passive stimuli. For example, exposing users to 360° videos [21], static environments [24], or dynamic, non-interactable scenes [22]. Physiological data collected during these experiences may not translate well to VR tasks that require movement, such as training or immersive education. Other works exploring active VR stimuli have often used commercial games [4, 38] or very domain-specific tasks [23] that may limit wider use, require long data collection times, or not generalize beyond a specific scope. To our knowledge, there are no current public datasets showing physiological responses to active VR tasks, nor is there study on the difference in reactions to active and passive VR stimuli.

To help address this gap, we introduce a set of four active affective tasks meant to elicit a set of emotions we expect to see in VR training or classroom settings. We envision their contribution to a larger library of generic (not domain-specific) affective tasks, analogous to affective film libraries, that researchers and developers could reuse to create models meant to recognize emotion in an active VR scenario. In this paper, we present a user study meant to validate these tasks as affective stimuli. In this study, we consider the following questions:

Do the stimuli precisely elicit intended emotions? Are the elicited emotions diverse between tasks? An imprecise stimulus may be unreliable, and two stimuli that do not have different emotional responses may be redundant. To this end, we collect self-reported emotional responses at the end of each task for statistical comparison.

Are the emotional responses due to movement? It is possible that requiring movement for the tasks inherently affects their emotional state, and since movement affects measured physiology it may obfuscate results. To this end, we implemented and compared neutral variants of each task which replicated the motions of their affective counterparts without added affective features.

How do results compare to those from passive stimuli? A library of tasks should be as reliable as one of videos. To this end, we also consider videos from a validated library of 360° videos [21] to compare consistency and spread of self-reported emotions.

Results suggest that our affective tasks successfully elicited a diverse range of targeted emotions, with precision and spread similar to that of the video library and similar time requirements.[1]

2 RELATED WORKS

To enable computers to understand human emotion, it is useful to first understand it ourselves. To that end, we first discuss how

[1]Packages containing task implementations are made available at https://github.com/cacs-vr-lab/affective-tasks.

emotion can be modeled in ways humans and computers can understand. We then review works eliciting emotion in traditional and VR mediums, with a focus on comparing active and passive stimuli.

2.1 Modeling Emotion

In order to map physiological responses onto a set of emotions, we must first define how we want to model emotions. Models typically take one of two approaches: discrete or dimensional [37]. Discrete models attempt to give named labels to each emotion. One seminal model is Ekman's original six basic emotions which he proposed were universally experienced across cultures and could be recognized through body movements and facial expressions [30]. These emotions were happiness, sadness, fear, disgust, anger, and surprise, and all other more complex emotions, such as love or annoyance, were combinations or derivatives of these.

Other popular discrete models have followed and tried to expand upon the idea of a set of basic emotions, such as Plutchik's wheel of emotions [32]. This model proposes eight basic emotions (joy, fear, trust, sadness, anticipation, surprise, anger, and disgust). These emotions can vary in intensity, naming weaker and stronger versions of each, and can be blended with others to produce more specific emotions. These discrete models are easy to understand and map closely to our linguistic models of the world, but can be difficult to analyze quantitatively when there are many labels and omitted feelings that may be hard to express in words.

Dimensional models attempt to account for the lack of specificity in language by placing all emotions on a set of axes. Russell's valence-arousal model [35] uses a 2D space with valence levels ranging from negative (unpleasant) to positive (pleasant) and arousal levels ranging from passive (calm) to active (excited). All emotions then fall in this space, for example, tiredness may have neutral valence and low arousal, and happiness may have a high valence and high arousal. Despite its age, this model is still among the most common used to have people self-identify emotions.

Despite its strengths, the 2D model has difficulty distinguishing between similar emotions in the same quadrant. For example, fear and anger are thought of as fairly distinct emotions, yet lie close together on the model, and could be mistaken for one another if self-reporters were not very precise in their reporting. To address this, some models include a third dimension, dominance, referring generally to the amount of control a person has on the emotion [28].

Dimensional models are useful in part because of tools that allow users to easily report on them. For example, the Self-Assessment Manikin [7] has users pick a character that aligns with their current mood, then ascribes an associated position on the relevant axis. Despite its simplicity, it has been successfully used in many studies [39]. We employ both dimensional and discrete measures, using a within-VR SAM interface and a custom emotion wheel based on our provided stimuli, for the sake of a more comprehensive analysis and comparison of results in both models.

2.2 Eliciting Emotion

Emotion elicitation typically exposes users to a set of stimuli meant to elicit targeted emotions of interest to the researcher. This process can be performed in many ways; older works include methods such as hypnosis or memory recall [6], relying on the user to recreate the emotions internally. Modern methods typically use external stimuli, such as the International Affective Picture System (IAPS) [20], composed of 956 pictures meant to elicit a wide set of emotions, with each corresponding to a coordinate on the PAD emotion model verified initially through SAM. Film has also been a common method [26], especially more recently [39], with multiple libraries constructed of film clips [15]. While films often elicit stronger emotions than images, they have limitations. For example, they are typically composed of clips of commercial films, limiting their use, and can elicit varying emotions per film, with a single clip attempting to elicit multiple emotions [34].

While these methods are common and functional, the viewer is a passive observer in the experience. This, theoretically, makes them unreliable for eliciting the same set of physiological responses of active experiences [5, 17, 39]. As such, more recent works have attempted to use active experiences in which the user takes part in the stimulus directly.

For example, Bevilacqua et al. [5] created "calibration games" for a system aiming to recognize stress and boredom during video game play. Each calibration game had a increasing difficulty, starting at unloseable and becoming unbeatable, with the curve intending to elicit boredom and then frustration or stress, creating a dataset with physiological signals much more closely related to what they would want to test in a final product. McDuff et al. [27] similarly used game-like tasks to elicit a set of emotions to detect cognitive stress levels. Tasks included the Berg Card Sorting Task and a game where users had to keep a ball from touching the walls it was pulled towards. VR, with its more active and immersive nature, then seems like a natural fit for the emotion elicitation task.

2.3 Eliciting Emotion in VR

Since the proliferation of affordable commercial VR devices, VR has become a much more common tool for eliciting emotion [25]. Many works focus on elicitation with immersive virtual environments or 360° videos. For example, Anwary et al. created a VR horror presentation and compared it to a neutral presentation, noting that fear was felt in a more pointed manner than in traditional systems. Marín-Morales et al. [24] used architectural principles when designing rooms in VR to elicit emotions in all four quadrants of the valence-arousal spectrum, controlling the lighting and geometry as affective parameters, suggesting that VR environments could focus more clearly on specific targeted emotions. Liao et al. [22] created and validated a set of VR scenes designed to elicit sadness, peace, happiness, distaste, and fear by distilling and combining affective parameters (color, motion, sound) from existing verified sources, with results showing emotions were felt potentially more strongly than those from non-VR stimuli. Li et al. [21] emulate the success of large emotional film libraries with a similar set of 73 360° videos.

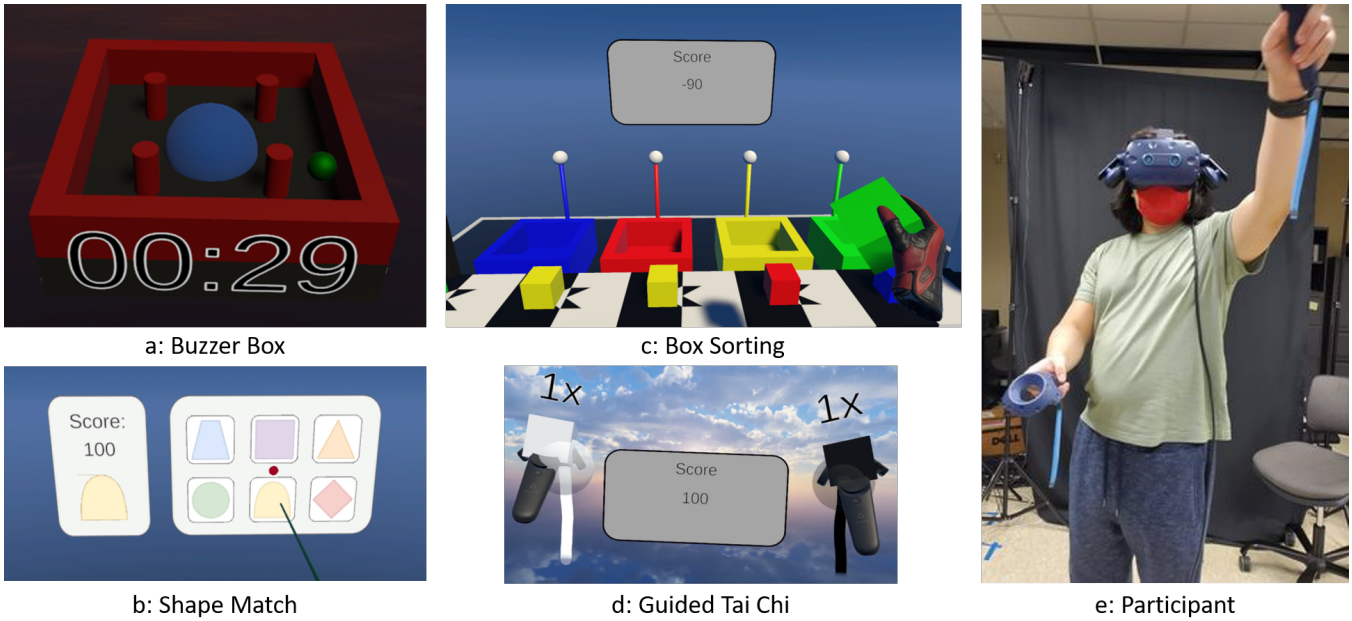
These and other similar works successfully demonstrate the strength of immersive VR, but crucially omit interactive elements that elevate the user beyond the observer role. Fewer works use fully interactive tasks or games. Those attempting to elicit emotion through games often use commercial games that are difficult to control in an experiment setting and have low subject counts that inhibit generalizing results [4, 38, 42]. The use of commercial games in general may be a hindrance to eliciting a full range of emotions as game mechanics often aim to make the player feel "good" [39]. Other studies have implemented custom interactive tasks with affective stimuli [33], but these have been very specific to a certain domain, such as a car crash simulation for firefighters [23], and take longer to complete relative to their video counterparts.

To our knowledge, there is no library of active affective tasks, similar to those for videos, that aims to elicit a diverse spread of emotions in a similar amount of time as is taken by videos. This hampers other researchers' abilities to investigate the effects of affective tasks, as they must be designed and built from scratch for each project, and limits our knowledge of how active tasks compare to passive videos. This work attempts to address this by designing and validating an initial set of four active affective tasks to be used for general emotion elicitation.

3 Task Design

3.1 Task Descriptions

We describe our four tasks by the named emotion they are designed to elicit, the type of motion required, and their desired location on



a: Buzzer Box c: Box Sorting

b: Shape Match d: Guided Tai Chi e: Participant

Figure 1: Affective Tasks. Buzzer Box asks the user to keep the green ball in the blue dome center while an invisible force pushes it away. Shape Match asks the user to select the shape shown on the left from the selection on the right. Box Sorting asks the user to sort the box into the correctly colored bin while occasionally giving incorrect feedback for correct sorts. Guided Tai-Chi asks the user to follow the colored spheres with their tracked hands while providing positive feedback. E shows a participant performing Tai-Chi.

the 2D valence-arousal model. This model creates four quadrants in which all emotions are described as having low/high valence and low/high arousal.

All tasks are performed in a similar virtual environment: an open space with a large black floor and a skybox with colors picked to match the intended emotion [22]. Affective tasks have a score, with correct interactions gaining points to encourage engagement with the task. Users are told there is a score threshold that must be reached to "win", though tasks are always ended after a certain amount of time to avoid unwanted frustration due to poor performance and to help standardize responses. Tasks are pictured in Figure 1.

As previously stated, some amount of physiological reaction may be due to the motion required by the task, and the motion itself may elicit an emotional reaction [44]. To consider this, we also create a neutral "motion-only" variant of each task that attempts to have the user replicate the motions of the original without any other affective influences; scores are removed, the skybox is changed to gray, and any non-essential positive/negative feedback is removed. Their inclusion is intended to allow us to better understand the relationship between motion and emotion, and how much emotional reaction is triggered by other affective influences in the tasks.

All tasks were originally verified in a pilot study [45], each being shown to have potential to elicit their intended emotion. Each has since been slightly modified based on user feedback.

Frustration: Buzzer Box Frustration is considered to have high arousal with low valence. We consider that difficult tasks [5] and frequent unpleasant noises [11] elicit stress, a similar emotion, and thus center task design around these traits. Buzzer Box aims to elicit frustration by presenting a deceptively simple task with a difficult goal and high amounts of negative feedback, based loosely on an analogous 2D task [27].

Buzzer Box presents a box, including four pegs, that follows the user's hand position and orientation. At the start of the task, a basic physics ball drops into the box and users are told to use the hand to balance the box to keep the ball within a marked center zone. The user gains points while the ball is in the zone and loses points otherwise. However, strong simulated forces that the user is not told about push the ball away from the box's center and pegs and the user must frantically rotate the hand to counteract force effects. A loud buzzer is played when the ball hits a peg or the box's walls, and is heard frequently due to the task difficulty.

In the neutral alternative, the user is given the same box and ball with a similar task to keep the ball in a marked zone. The zone is initially placed in a corner and moves to another corner when the ball reaches it. This similarly requires the user to frequently rotate the hand. The extra forces and loud buzzer sound are removed.

Confusion: Box Sorting Confusion is among the lesser-studied emotions, and while it was not always recognized as such [16], it has recently been deemed important. Confusion is generally considered to have moderately high arousal and low valence, placing it in a different position in the same low valence / high arousal quadrant as frustration. Confusion can be elicited when the user becomes unsure of something after having previously thought it to be understood [2, 9]. Box Sorting aims to elicit confusion by presenting a simple task but occasionally breaking its rules without explanation.

In Box Sorting, the user is stationed in front of a conveyor belt and four colored bins. Once the task starts, boxes colored to match bins appear on the conveyor belt at a rate of one per second. The user is asked to sort the boxes into bins by color (grabbing by reaching out and squeezing the trigger, throwing by releasing the trigger during a throw motion). If the user performs an incorrect sort, a loud buzzer noise is played. Once every five to ten correct sorts, the thrown box will change color after release, thus creating an incorrect sort and confusing the user. To successfully perform the task, the user will make many tossing motions.

In its neutral alternative, all boxes are colored gray and there is only a single, wider bin. Boxes no longer change color when thrown. The results in a similar frequent tossing motion.

Boredom: Shape Match Boredom is considered to have low arousal and valence. We consider a boring active task to be one that is easy but with frequent interaction [5]. Shape Match aims to elicit boredom by presenting a trivial task for a long period of time.

In Shape Match, the user is shown a shape image on a 2D panel and must select the matching shape from a set on a separate panel. Upon a correct match, points are awarded and the user is given a two second break before the next target shape is shown. Upon an incorrect match, points are deducted and the loud buzzer is played. Standard wand selection techniques are used to avoid interaction methods that may produce entertainment. The Box Sorting task was originally planned as the boring task, but prototypes showed that the act of throwing the boxes made the task moderately enjoyable to pilot participants. Motion in Shape Match is minimal, requiring only slight wrist rotations.

Due to the minimal motion involved, the neutral alternative is largely similar. Shapes are presented on a single panel, the target shape is highlighted with an orange border, and the user must select it with the same wand-based technique. The time between targets is reduced to half a second, and the overall time for the task is reduced.

Pleasure/Excitation: Guided Tai-Chi Pleasure is typically considered to have high valence and neutral-to-high arousal. Tai-Chi aims to elicit pleasure by mimicking basic low-intensity tai-chi movements often done for relaxation or pleasure [41], combined with game elements meant to enhance satisfaction and excitement.

In Tai-Chi, two differently-colored spheres move along a pre-recorded path, indicating desired hand positions for a basic tai-chi routine. The user is asked to place their matching-colored hands in the spheres to follow their path, with leading trails showing two seconds of their future path to minimize confusion. The score increases every second that the hands maintain contact with the spheres. A score multiplier increases while contact is maintained to encourage consistent performance, and is reset upon error. The controllers gently vibrate while contact is maintained to add a pleasing effect. Motivational background music is included, as well as basic animations and other positive sound effects for performing well. To successfully perform the task, the user will make a wide range of slow controlled hand translations (rotation is not enforced).

In its neutral alternative, the score, multipliers, music, and vibrations are removed, and users are asked to follow the same pre-recorded path. Removal of all features resulted in users not being able to tell when their hands were in the guiding spheres, so small animations that denoted successful contact were preserved.

Neutral A neutral room was designed to be placed in between tasks so that users could recover from the emotion elicited from the previous task, as is commonly recommended with film datasets [10]. The room contains a sky-blue skybox with basic rain sounds and a timer showing time remaining until the next task.

4 Experiment Methods

There were two main user study goals. First, to validate our proposed affective tasks as a useful emotion elicitation tool, and second, to compare these active stimuli to passive stimuli (a set of 360° videos). For this, subjects were exposed to each task and its neutral alternative and to a set of videos selected from a previous study [21]. During each stimulus, physiological readings are taken from sensors described below. After each stimulus, subjects performed a self-assessment noting the emotion they felt most strongly and their ratings on the virtual SAM.

4.1 Participants and Apparatus

54 subjects (age: mean=21.5, sd=2.8, Gender: 20 female, 34 male) used an HTC Vive Pro Eye headset and controllers to view and interact with the environment. Heart rate information was streamed to the system from a Polar H10 chest strap using the ANT+ protocol. A Polar Verity Sense wristband was used instead if the chest strap did not fit the subject. Electrodermal activity (EDA) and skin temperature were streamed to the system from an EmotiBit sensor strapped to the wrist over a dedicated WiFi connection to the EmotiBit Oscillosope, and from there to the system using the OSC protocol. Eye tracking and positional data were captured natively from the headset.

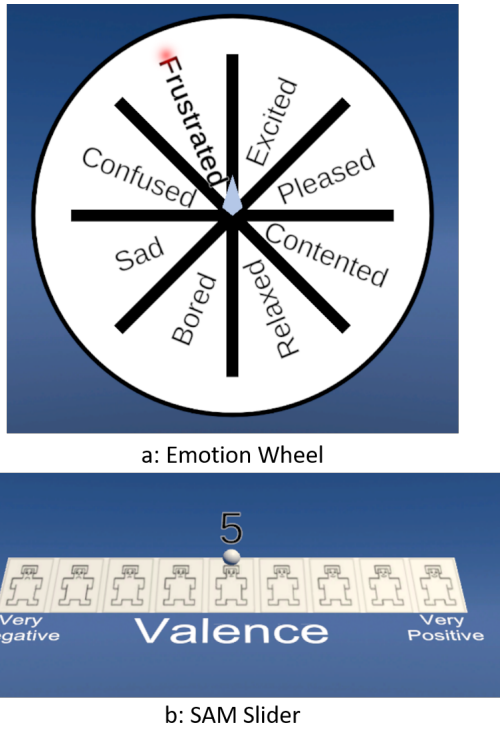


Figure 2: Interface used for self-assessment after each task.

Tasks and experiment software were implemented using Unity 2019.4.11. Experiments used an Alienware Aurora R13 with an Intel i7-12700KF processor and a GeForce RTX 3080 graphics card. The study was approved by the university Institutional Review Board under approval number SP21-79 CACS.

4.2 Experiment Design and Metrics

The experiment had a within-subjects design. The independent variable was the stimulus experienced. Subjective dependent variables include the self-reported valence, arousal, emotion label, and emotion strength given after each stimulus.

4.3 Procedure

Subjects first filled out a consent form and background questionnaire, noting any familiarity with VR and filling a custom subscale of the Immersive Tendencies Questionnaire (ITQ) [36, 43] (items 1, 3, 6, 7, 14). After an initial discussion of the experiment, subjects donned the headset and sensors. Subjects were told that they would be exposed to a series of basic tasks and videos, and that they should be attentive to their emotional state during the each experience as they would complete a self-assessment after each. They were then given a brief tutorial defining valence and arousal and explaining how to use the end-of-task assessment tool. The experiment proctor confirmed their understanding and answered any questions on the concepts. The main part of the experiment was divided into three phases, consisting of two task phases and one video phase. The order of the phases was randomized. After the final phase was complete, subjects were interviewed about their responses.

4.3.1 Task Phase

Subjects completed four tasks in each task phase, with one version of each of the four tasks in each phase, with randomized order. Two tasks were randomly selected to be presented as the neutral versions in the first phase, with the other two presented as the neutral version in the second phase. The distribution was balanced so that, combined, the two phases included both versions of each task.

Each task had a duration based on preliminary testing and pilot results. Buzzer Box lasted 90 seconds, as testing indicated that the difficult task would get more frustrating over time, but a user may give up if it lasts too long. Its neutral counterpart lasted 60 seconds so as not to elicit boredom. Box Sorting lasted 45 seconds, so the broken rule would happen a small number of times and the subject would not catch on to the system cheating. Its neutral version also lasted 45 seconds. Shape Match lasted 180 seconds, as testing showed that the simple task would get more boring over time. Its neutral version lasted 60 seconds to reduce eliciting boredom. Tai-Chi lasted 120 seconds, as testing suggested this gave users enough time to enjoy the experience before the novelty wore off. Its neutral version had the same duration, as it used the same recording. Subjects rested in the neutral room for 30 seconds between tasks.

4.3.2 Video Phase

A selection of seven 360° videos from [21] (listed in Table 1) were shown to subjects. A subset was used to avoid unnecessarily lengthening the experiment. Specific videos were chosen because they were among those which elicited the strongest emotions (i.e., had an arousal/valence value furthest from the neutral 5). Two were chosen from each quadrant, with the exception of the high arousal / low valence quadrant, due to a lack of such videos in the dataset.

4.3.3 End-Of-Task Assessment

After each task and video, users completed a self-assessments using a simple in-world interface (pictured in Figure 2). To obtain a descriptive emotion label, a wheel showing 8 emotions (frustrated, confused, sad, bored, excited, pleased, contented, relaxed) was displayed. Two emotions were included from each quadrant of a circumplex model, to display a variety of options while keeping the interface reasonably simple. Four labels were those we desired to elicit with tasks. The other four were from common responses in early user testing, and filled needed spots in the quadrants. Once a label was selected, users were given a slider tool to rate how intensely they felt that emotion on a scale from 1 to 7. Finally, they used a similar slider tool to rate their valence and arousal on a SAM-like scale [7] from 1 to 9. Sliders were adjusted by touching either side of a controller touchpad to move a selection icon (sphere) on the slider image. The currently selected number was displayed above the slider to clearly indicate the current value.

After the participant completed all three phases, they were interviewed for any feedback about any tasks in which they gave what we considered an unusual response or had unusual behavior while performing, and were asked if they felt more emotionally affected by the videos or the tasks.

5 Results

The primary purpose of this work is to validate the affective tasks as emotional stimuli. If we can demonstrate their validity, it provides grounds to explore more active affective stimuli and produce emotion recognition models from them. Again, we consider a good library of stimuli to precisely elicit a diverse range of emotions (i.e. a different emotional response from each stimulus) to avoid unreliable or redundant stimuli. Our validation approach therefore considers the frequency of reported emotion labels per stimulus and the difference in reported valence / arousal ratings between stimuli. We avoid direct comparison between task and video pairs, instead comparing overall data about their groups, as we consider this more relevant to the general comparison of active and passive stimuli.

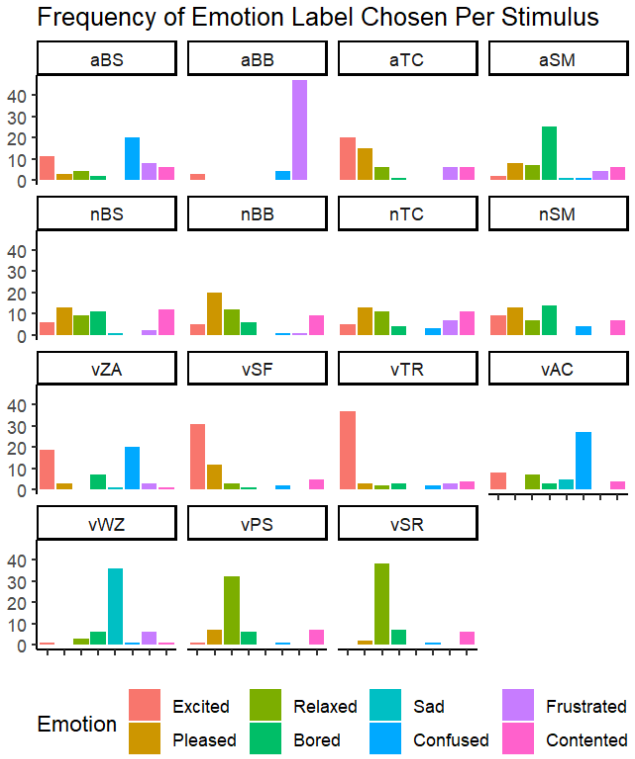


Figure 3: Frequency of reported emotion label per stimulus. Recommended to view in color.

Because measures are non-normally distributed, we primarily analyze differences using Friedman tests and followup pairwise Wilcoxon Signed-Rank tests with Holm correction to address family-wise error. While we primarily focus on validating the affective tasks, we also aim to compare them to their neutral counterparts to check that the added affective components did change the emotional response. Additionally, we report findings from video stimuli and compare the consistency of reporting between them and tasks. Table 1 lists general summary statistics on all stimuli.

5.1 Reported Emotion Labels

Figure 3 shows the frequency with which each stimulus was given each emotion label, in part demonstrating the consistency of eliciting named emotions. Among affective tasks, Buzzer Box and Shape Match appear the most consistent, primarily eliciting frustration (47) and boredom (25), respectively. Tai-Chi primarily elicited excitation (20) and pleasure (15), both being high valence emotions but differing in arousal. Box Sorting was the least consistent, primarily eliciting confusion (20) and excitation (11), both being high arousal and differing in valence, but with a wider distribution among others.

Neutral tasks were less consistent, showing a generally wider distribution of ratings. The most commonly reported emotions among them are pleased and relaxed, showing a general trend towards positive low arousal emotions. Videos showed high consistency, with each video having a stand-out, most common emotion label, aside from Zombie Apocalypse, which was split between confused (20) and excited (19).

Table 1: Descriptive results for stimuli used in this study. (A/N/V) denotes if it is an Affective task, its Neutral variant, or a Video. Code denotes what each stimulus is labeled as in figures. Target indicates the emotion targeted by a task's design or the mean valence / arousal values of the videos from their source experiment [21]. Emotion Label shows the most frequently chosen label of the eight given emotions and the number of subjects who picked it. Target Quadrant indicates if the stimulus is meant to elicit High or Low Arousal or Valence.

Stimulus	Code	Time (s)	Target	Emotion Label	Target Quadrant	Mean (V,A)	Median (V,A)
(A) Box Sorting	aBS	45	Confusion	Confused (20)	HALV	(5.18, 5.35)	(6, 6)
(A) Buzzer Box	aBB	90	Frustration	Frustration (47)	HALV	(4.09, 6.62)	(4, 6.5)
(A) Tai-Chi	aTC	120	Pleased / Excited	Excited (20)	HAHV	(6.44, 6.05)	(7, 7)
(A) Shape Match	aSM	180	Bored	Bored (25)	LALV	(5.28, 3.8)	(5, 3)
(N) Box Sorting	nBS	45	Neutral	Pleased (13)	Neutral	(5.34, 3.43)	(6, 3)
(N) Buzzer Box	nBB	60	Neutral	Pleased (20)	Neutral	(6.25, 4.09)	(7, 4)
(N) Tai-Chi	nTC	120	Neutral	Pleased (13)	Neutral	(5.62, 4.39)	(6, 4)
(N) Shape Match	nSM	60	Neutral	Bored (14)	Neutral	(5.49, 3.64)	(6, 3)
(V) Zombie Apocalypse	vZA	265	(3.2, 5.6)	Confused (20)	HALV	(4.93, 5.57)	(5, 6)
(V) Speed Flying	vSF	154	(6.75, 7.42)	Excited (31)	HAHV	(6.77, 5.67)	(7, 6)
(V) Tight Rope	vTR	151	(6.46, 6.91)	Excited(37)	HAHV	(5.92, 6.07)	(6, 6)
(V) Abandoned City	vAC	50	(4.39, 2.77)	Confused (27)	LALV	(5.3, 4.59)	(5, 5)
(V) War Zone	vWZ	183	(2.53, 3.82)	Sad (36)	LALV	(3.93, 4.39)	(4, 4)
(V) Pacific Sunset	vPC	134	(6.19, 1.81)	Relaxed (32)	LAHV	(7.15, 3.54)	(7, 3)
(V) Sunrise	vSR	120	(6.57, 1.57)	Relaxed (38)	LAHV	(6.98, 2.98)	(7, 2)

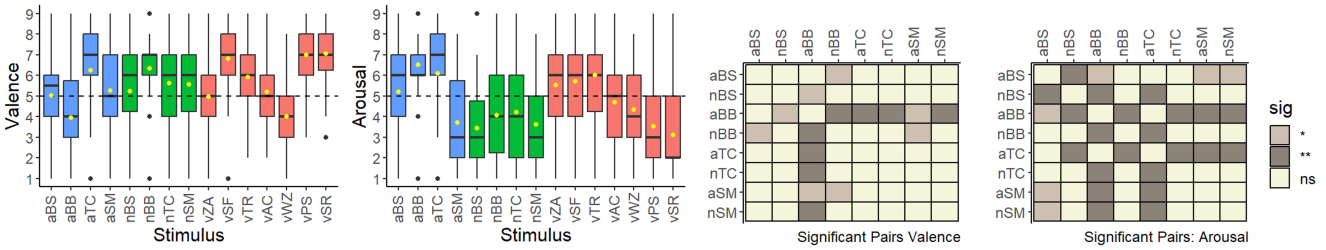


Figure 4: Results from end-of-task SAM assessments. Left: Boxplots summarizing reported valence and arousal for each stimulus. Interior black lines indicate medians, interior yellow dots indicate means. Right: Results matrix of pairwise wilcoxon signed rank tests after Holm correction. Darkest cells indicate $p < .001$ for the row and column stimulus pair. Lighter cells indicate $.001 < p < .05$ for that pair. Lightest cells indicate no significant difference found.

5.2 SAM Ratings

End-of-task SAM reports are summarized in Figure 4. Median valence and arousal for affective Buzzer Box (4, 6.5) and Tai-Chi (7, 7) fall in expected ranges. For Box Sorting, median valence (6) and arousal (6) were both above neutral; while this was expected for arousal, we intended for valence to be below neutral in response to the "confusing" stimuli. For Shape Match, median arousal (3) was below arousal, as expected, but median valence (5) was at neutral. We note a general bias towards high valence across the stimuli.

The neutral variants of tasks show a trend of high valence and low arousal, indicating a generally pleasant and calming task in line with subjects frequently reporting "pleased" as their emotion label.

Valence and arousal medians for videos generally followed expected trends with a few notable exceptions in videos meant to elicit a low valence. For example, the mean[2] valence among our subjects for War Zone was 3.93 (SD = 1.83), which is shown as significantly higher than the original mean of 2.53 in [21] ($t = 5.894, p < .001$). Similarly, our mean valence for Zombie Apocalypse was 4.93 (SD = 1.992) and is significantly higher than the original mean of 3.2 ($t = 3.208, p = .002$). Some similar cases are seen with arousal as well, such as our mean arousal for Speed Flying (5.67, SD = 2.02) being significantly lower than in the original (7.42, $t = -6.258, p < .001$). This generally points to a slight difference in stimulus perception between our studies' subject pools, possibly due to difference in balance of gender or cultural background [8].

Figure 4 also notes statistically significant comparisons between tasks after Holm correction, with a darker square indicating a significant difference in the row / column pair. We consider it important that the stimuli differ significantly in valence or arousal as it suggests they are actually eliciting a wider range of emotions. A Friedman test shows that the type of task experienced has an effect on both valence [$\chi^2(7) = 60.84, p < .001, W = .16$] and arousal [$\chi^2(7) = 112.1375, p < .001, W = .297$].

For reported arousal, we note first a significant difference between affective and neutral variants for Box Sorting ($r = .563$), Buzzer Box ($r = .782$), and Tai-Chi ($r = .66$; $p < .001$ for all pairs), all with large effect sizes showing the affective version elicited higher arousal. This is expected, as these three were intended to elicit above-neutral arousal. Shape Match did not differ from its neutral variant ($p = 1$, corrected), so both may have been similarly boring.

Between tasks, we note that Shape Match elicited significantly less arousal than Box Sorting ($p = .002, r = .514$), Buzzer Box ($p < .001, r = .745$), and Tai-Chi ($p < .001, r = .664$), with large effect sizes for all comparisons. This is expected, as Shape match was intended to have below-neutral arousal and the others to have above-neutral arousal. Between the three high-arousal tasks, Buzzer Box was shown to elicit significantly higher arousal than Box Sorting ($p = .002, r = .56$), establishing some difference within the high-

[2]While we generally consider median data and non-parametric tests for our analysis, we compare means here as only means were reported in the original work.

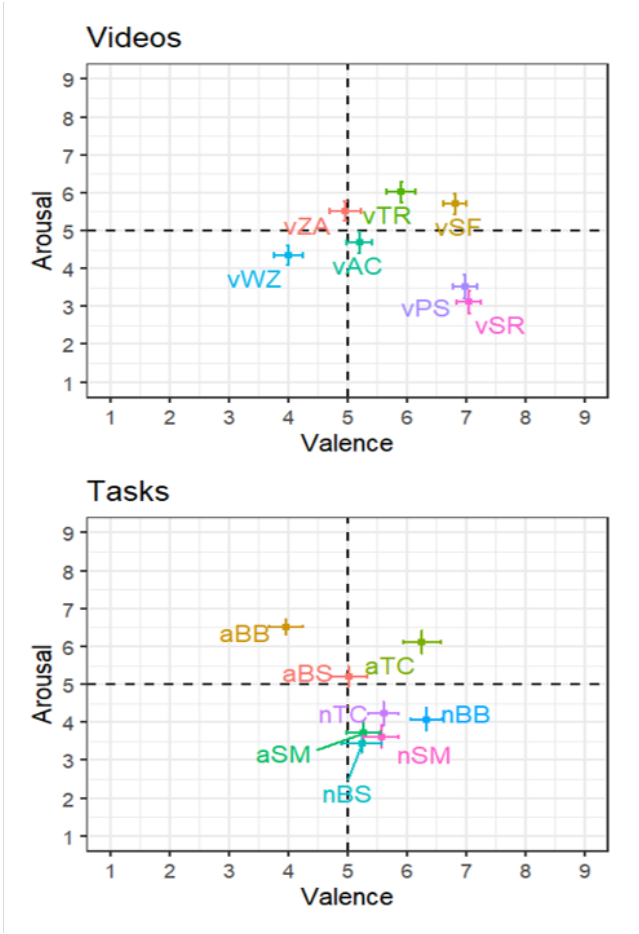


Figure 5: Mean self-reported valence and arousal per stimulus mapped as 2D coordinates. Bars show standard error of the means in both dimensions.

arousal group.

For reported valence, we notice fewer significant differences between stimuli. Buzzer box was the only task to elicit a difference between affective and neutral variants ($p < .001, r = .669$), while Box Sorting ($p = 1$ after correction), Tai-Chi ($p = .827$) and Shape Match ($p = 1$) did not.

Between the tasks, Tai-Chi was intended to elicit high valence, while the others were intended to elicit low valence. In line with this, we see that Buzzer Box elicited lower valence than Tai-Chi ($p < .001, r = .648$) and, to a lesser extent, Shape Match ($p = .005, r = .506$). However, we do not see significant differences between other pairs. This is in line with the general trend of higher-than-expected arousal across most stimuli discussed above, with Box Sorting and even Shape Match being perceived more positively than intended.

Of note, when combining differences in both valence and arousal, we see a difference in at least one factor for all task pairs except Box Sorting and Tai-Chi.

Figure 5 visualizes mean valence and arousal in a grid to more clearly distinguish high and low quadrants, allowing us to see comparisons between tasks and videos more clearly. Visually, we note a similar spread between the two groups in both dimensions, with a slightly wider spread in valence for videos. We also see video stimuli reaching further into the high valence, low arousal quadrant with Pacific Sunset and Sunrise, while tasks appear to reach further into the low valence, high arousal quadrant with affective Buzzer Box. This echoes trends from other work stating that few videos elicited high arousal and low valence [21].

Table 2: Intraclass Correlation Coefficients with 95% confidence intervals showing interrater reliability.

	ICC - Arousal	ICC - Valence
Affective Tasks	0.963 [0.883, 0.997]	0.919 [0.74, 0.994]
Neutral Tasks	0.514 [-0.555, 0.965]	0.66 [-0.086, 0.976]
Videos	0.948 [0.873, 0.989]	0.968 [0.921, 0.993]

Finally, following analysis from [21], we compare the ability for groups of stimuli to elicit varying emotions by comparing their overall standard deviation. Affective tasks varied widely on arousal ($M = 5.39, SD = 2.28$) with very similar variance shown by videos ($M = 4.7, SD = 2.29$). Affective tasks also varied on valence ($M = 5.13, SD = 2.33$) with slightly less variance shown by videos ($M = 5.85, SD = 1.98$), though this may be explained by a slight over-representation of high valence videos in our chosen subset.

5.3 Reporting Consistency

In addition to eliciting significantly different emotions in users, good elicitation stimuli should elicit those emotions consistently between users [34]. Higher consistency, we expect, would result in stronger physiological datasets. Following analysis from [46] and guidelines from [19], we perform a two-way random effects, consistency, multiple raters ICC using study subjects as raters and stimuli as subjects to get a measure of interrater reliability for groups of stimuli. We use consistency, rather than absolute agreement, as a measure because we consider it more important that rating patterns are consistent between users (e.g. most users report Buzzer Box as lower valence than Tai-Chi) than stimuli always receiving the same rating (e.g. Buzzer Box is given the same valence rating across most users). Results are listed in Table 2.

Following standard interpretations [19], we see that affective tasks have excellent consistency for arousal ($ICC = .963, p < .001$) and good consistency for valence ($ICC = .919, p < .001$). Neutral tasks, on the other hand, show poor and unreliable consistency for both arousal ($ICC = .514, p = .108$) and valence ($ICC = .66, p = .035$). This is in line with affective tasks eliciting more tightly targeted emotions while neutral tasks were more spread out. Videos showed excellent consistency for both arousal ($ICC = .948, p < .001$) and valence ($ICC = .968, p < .001$).

5.4 Other Mediating Variables

During the last 38 experiment runs (n=38), each subject was asked if they felt they were more emotionally affected by the tasks or videos as a whole. From this, 23 chose tasks, 14 chose videos, and 1 chose no difference.

To determine if this preference moderated the relationship between stimulus and reported valence / arousal, we conducted two two-way ANOVAs (preference and stimulus as independent variables) with the 37 subjects who noted a preference [3]. Results suggested preference had no interaction with stimulus type for valence [$F(9.03, 315.96) = .9, p = .525$] or arousal [$F(8.58, 300.21) = 1.56, p = .130$]. However, considering a majority stated that they felt more affected by tasks, we believe this warrants further investigation.

We also consider that a subject's personal immersive tendencies might reflect how they experience these stimuli. To evaluate this effect, we summed scores from the subjects' ITQ items to create a single ITQ score (maximum possible 35, $Mean = 24.47, SD = 4.36$ for all subjects) and calculate Spearman's correlation with SAM reportings. We find a small positive correlation for both valence ($\rho = .14[.06, .23], p < .001$) and arousal ($\rho = .17[.09, .26], p < .001$), indicating that subjects with stronger immersive tendencies tend to

report more arousal and pleasure across these stimuli.

6 Discussion

Looking across our analyses, we see a strong trend that suggests affective tasks are capable of eliciting consistent (Sections 5.1 and 5.3) and diverse (Section 5.2) emotions. The Buzzer Box and Tai-Chi tasks elicited their targeted emotions both in terms of most common label selected and expected valence and arousal. Box Sorting and Shape Match elicited the most commonly expected emotion label and expected Arousal, but elicited higher valence than expected.

For Box Sorting, we expect those who noticed the "cheating" system were confused, but those who did not were just excited by the game. For Shape Match, the high valence plus high reports of boredom suggest that it may have been boring but not actually felt very negatively. This is in line with generally higher than expected valence across most stimuli, possibly indicative of bias from the novelty of the VR experience, and may warrant further investigation into the nature of low-valence emotions in a novel game-space.

Again, all affective tasks show a difference in at least one of valence or arousal between all pairs except Box Sorting and Tai-Chi. For most pairs, this is another indication of diverse elicitation. For Box Sorting and Tai-Chi, one possibility is that both were perceived as similarly exciting, and that additional steps should be taken to ensure that Box Sorting comes across as confusing.

Results also suggest that affective tasks elicited significantly different emotions from their neutral counterparts. All neutral tasks elicited a similarly wide distribution of reported labels, low arousal, and high valence, indicating the elicitation of a general calm pleasure. This distinction suggests that it is not just the motions performed during the tasks that elicited emotions, and the added affective stimuli were effective. In theory, it then may be possible to apply these same added stimuli (unseen physics, loud sounds, elongated wait times, score systems, or cheat mechanics) to other tasks with different motions to elicit similar emotions.

Knowing that the tasks elicit different emotions, we compare their general variance and interrater reliability to videos and see that they elicit a similarly wide range of emotions with only a small drop in consistency for valence ratings. Thus we posit that affective tasks have a similar capability to elicit emotion. However, we consider there may be other strengths that make tasks more desirable in certain cases.

Videos and other passive elicitation media are often driven by narrative elements [17, 40] which may not resonate with people in different demographics [34]. 360° video also uniquely allows each viewer to have a different experience, as certain interactions are not enforced. This limitation can be overcome with tasks, as they forego narrative involving other characters in favor of focusing on the user, and require all users to interact with them in the same way. In fact, subjects who stated they felt more emotionally affected by the tasks said it was because they felt the stimulus was actually happening to them instead of being something they had to witness.

Additionally, low valence / high arousal emotions are often difficult for passive stimuli to elicit [1, 21, 40], often having to use horror or gore imagery [20] to be effective. Our Buzzer Box task instead effectively targets this quadrant through frustration. We consider this easier to elicit through an interactive task and a potentially better choice for subjects who may be sensitive to fear stimuli.

Finally, due to the difference in how emotions are elicited in narrative videos vs. egocentric tasks, we must address that there may be a more fundamental difference to their experience. Narrative experiences may more intuitively align with our common concept of "emotions" while one may intuit cognitive tasks to elicit some other form of affective feeling. Our results clearly show that something was elicited through the tasks, but, for example, while the Tight Rope video and Box Sorting task share the same median valence and arousal, they also have very different label reporting patterns, indicating some difference in feeling. One explanation is that the two valence and arousal dimensions do not, on their own, explain all variance in emotion, but another possibility is that there may be a deeper experiential difference in what is felt during each stimulus. More research into this experiential difference between active and passive stimuli could be instrumental in furthering our understanding of emotion in general.

All considered, the strengths shown for our affective tasks and others [17] advocate for the construction of a general library aiming to elicit a wide range of possible emotions, akin to those available for videos [15, 21]. With the converging standardization of VR interfaces, such a library could eventually be as easy to import to new projects as current video libraries. We consider our work a first step to identifying promising task designs and building this library.

7 Conclusion

We explored the creation of a generalized library of affective tasks to elicit targeted emotions. Four tasks were introduced, each adding an affective element atop a required movement. Neutral variants of each task were created to give insight into the effects of these added elements versus the movement alone. An elicitation validation study had subjects perform each task variant and watch seven 360° videos, self-reporting their experienced emotion after each.

Results indicate that the affective tasks elicited a diverse range of emotions, with differing common emotion labels attached and valence / arousal ratings falling in different quadrants. Ratings also differ significantly between affective and neutral tasks, suggesting the task motion alone was not responsible for elicitation. The spreads of elicited emotion between tasks and videos were similar, with emotions eliciting a slightly wider range of arousal and video eliciting a wider range of valence, suggesting rouhgly similar elicitation capabilities. Interrater reliability suggested strong consistency between user ratings for affective tasks and videos.

One obvious limitation of this work is a relatively small number of tasks compared to the many videos typically available in video stimulus libraries. This investigation limited the number of tasks to allow experiment time for comparison to neutral variants and videos. An obvious next step, then, is to expand the library with new tasks to elicit other targeted emotions. Future work can also expand by crossing the affective elements added to tasks onto tasks with different motion requirements to see if the effects transfer over, or adding such elements to a single task at different times to see if the same type of task motion can elicit many emotions.

Future works must also evaluate the capabilities of emotion recognition models trained on acquired physiological data. Specifically, it is of interest to see how models trained on data elicited by tasks compare to those trained from video data. Based on prior research, we expect a model trained on task data would more accurately recognize emotion during active VR experiences such as work or training.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1815976.

References

[1] M. A. Anwary and S. Sigwebela. *Emotion Elicitation in the Laboratory: Virtual reality environments can make you feel fear*. PhD thesis, University of Cape Town, 2018.

[2] A. Arguel, L. Lockyer, G. Kennedy, J. M. Lodge, and M. Pachman. Seeking optimal confusion: a review on epistemic emotion management in interactive digital learning environments. *Interactive Learning Environments*, 27(2):200–210, 2019. doi: 10.1080/10494820.2018.1457544

[3] R. M. Baron and D. A. Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173–1182, 1986. doi: 10.1037/0022-3514.51.6.1173

[4] C. Bassano, G. Ballestin, E. Ceccaldi, F. I. Larradet, M. Mancini, E. Volta, and R. Niewiadomski. A vr game-based system for multimodal emotion data collection. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, MIG '19, 2019. doi: 10.1145/3359566.3364695

[5] F. Bevilacqua, H. Engström, and P. Backlund. Game-calibrated and user-tailored remote detection of stress and boredom in games. *Sensors (Switzerland)*, 19(13):1–43, 2019. doi: 10.3390/s19132877

[6] G. H. Bower, A. Sahgal, and D. A. Routh. Affect and Cognition. *Philosophical Transactions of the Royal Society; Series B*, 302(1110):387–402, 1983.

[7] M. Bradley and P. J. Lang. Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(I):49–59, 1994.

[8] M. M. Bradley, M. Codispoti, D. Sabatinelli, and P. J. Lang. Emotion and motivation ii: sex differences in picture processing. *Emotion*, 1(3):300, 2001.

[9] M. Chaouachi, I. Jraidi, S. P. Lajoie, and C. Frasson. Enhancing the learning experience using real-time cognitive evaluation. *International Journal of Information and Education Technology*, 9(10):678–688, 2019. doi: 10.18178/ijiet.2019.9.10.1287

[10] J. A. Coan and J. J. Allen. *Handbook of emotion elicitation and assessment*. Oxford university press, 2007.

[11] S. Cohen, G. Evans, D. Stokols, and D. Krantz. *Behavior, Health, and Environmental Stress*. Springer US, 2013.

[12] A. Costa, J. A. Rincon, C. Carrascosa, V. Julian, and P. Novais. Emotions detection on an ambient intelligent system using wearable devices. *Future Generation Computer Systems*, 92:479–489, 2019. doi: 10.1016/j.future.2018.03.038

[13] H. D. Fernández, M. Koji, and K. Kondo. Adaptable game experience based on player's performance and EEG. In *Proceedings of the 2017 NICOGRAPH International Conference*, pp. 1–8, 2017. doi: 10.1109/NICOInt.2017.11

[14] A. Fowler, K. Nesbitt, and A. Canossa. Identifying cognitive load in a computer game: An exploratory study of young children. *IEEE Conference on Computatonal Intelligence and Games, CIG*, 2019-Augus, 2019. doi: 10.1109/CIG.2019.8848064

[15] T. L. Gilman, R. Shaheen, K. M. Nylocks, D. Halachoff, J. Chapman, J. J. Flynn, L. M. Matt, and K. G. Coifman. A film set for the elicitation of emotion in research: A comprehensive catalog derived from four decades of investigation. *Behavior Research Methods*, 49(6):2061–2082, 2017. doi: 10.3758/s13428-016-0842-x

[16] U. Hess. Now you see it, now you don't–the confusing case of confusion as an emotion: Commentary on rozin and cohen (2003). *Emotion*, 3, 2003.

[17] K. Hidaka, H. Qin, and J. Kobayashi. Preliminary test of affective virtual reality scenes with head mount display for emotion elicitation experiment. In *2017 17th International Conference on Control, Automation and Systems (ICCAS)*, pp. 325–329, 2017. doi: 10.23919/ICCAS.2017.8204459

[18] C. Jacobé de Naurois, C. Bourdin, A. Stratulat, E. Diaz, and J. L. Vercher. Detection and prediction of driver drowsiness using artificial neural network models. *Accident Analysis and Prevention*, 126(December 2017):95–104, 2019. doi: 10.1016/j.aap.2017.11.038

[19] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012

[20] P. J. Lang, M. M. Bradley, B. N. Cuthbert, et al. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1:39–58, 1997.

[21] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams. A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Frontiers in psychology*, 8:2116, 2017.

[22] D. Liao, L. Shu, G. Liang, Y. Li, Y. Zhang, W. Zhang, and X. Xu. Design and Evaluation of Affective Virtual Reality System Based on Multimodal Physiological Signals and Self-Assessment Manikin. *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, 4(3):216–224, 2020. doi: 10.1109/JERM.2019.2948767

[23] N. Lipp, N. Dużmańska-Misiarczyk, A. Strojny, and P. Strojny. Evoking emotions in virtual reality: schema activation via a freeze-frame stimulus. *Virtual Reality*, 25:279–292, 2021.

[24] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, and G. Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8(1):1–15, 2018. doi: 10.1038/s41598-018-32063-4

[25] J. Marín-Morales, C. Llinares, J. Guixeres, and M. Alcañiz. Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors*, 20(18), 2020. doi: 10.3390/s20185163

[26] I. B. Mauss, L. McCarter, R. W. Levenson, F. H. Wilhelm, and J. J. Gross. The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, 5(2):175–190, 2005. doi: 10.1037/1528-3542.5.2.175

[27] D. J. McDuff, J. Hernandez, S. Gontarek, and R. W. Picard. COGCAM: Contact-free measurement of cognitive stress during computer tasks with a digital camera. *Conference on Human Factors in Computing Systems - Proceedings*, pp. 4000–4004, 2016. doi: 10.1145/2858036.2858247

[28] A. Mehrabian. Comparison of the pad and panas as models for describing emotions and for differentiating anxiety from depression. *Journal of psychopathology and behavioral assessment*, 19(4):331–357, 1997.

[29] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):1–16, 2020. doi: 10.1038/s41597-020-00630-y

[30] P.Ekman and W.V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[31] R. W. Picard. *Affective computing*. MIT press, 2000.

[32] R. Plutchik. The nature of emotions. *American Scientist*, 89(4):344–350, 2001.

[33] L. Reidy, D. Chan, C. Nduka, and H. Gunes. Facial electromyography-based adaptive virtual reality gaming for cognitive training. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, p. 174–183, 2020. doi: 10.1145/3382507.3418845

[34] J. Rottenberg, R. Ray, and J. Gross. Emotion elicitation using films. In J. A. Coan and J. J. Allen, eds., *The handbook of emotion elicitation and assessment*. New York: Oxford University Press, Inc, 2007.

[35] J. A. Russell. Evidence of convergent validity on the dimensions of affect. *Journal of Personality and Social Psychology*, 36(10):1152–1168, 1978. doi: 10.1037/0022-3514.36.10.1152

[36] S. Rózsa, R. Hargitai, A. Láng, A. Osváth, E. Hupuczi, I. Tamás, and J. Kállai. Measuring immersion, involvement, and attention focusing tendencies in the mediated environment: The applicability of the immersive tendencies questionnaire. *Frontiers in Psychology*, 13, 2022. doi: 10.3389/fpsyg.2022.931955

[37] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. A review of emotion recognition using physiological signals. *Sensors (Switzerland)*, 18(7), 2018. doi: 10.3390/s18072074

[38] I. Shumailov and H. Gunes. Computational analysis of valence and arousal in virtual reality gaming using lower arm electromyograms. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 164–169, 2017. doi: 10.1109/ACII.2017.8273595

[39] R. Somarathna, T. Bednarz, and G. Mohammadi. Virtual reality for emotion elicitation – a review. *IEEE Transactions on Affective Computing*, pp. 1–21, 2022. doi: 10.1109/TAFFC.2022.3181053

[40] N. S. Suhaimi, C. T. B. Yuan, J. Teo, and J. Mountstephens. Modeling the affective space of 360 virtual reality videos based on arousal and valence for wearable eeg-based vr emotion classification. In *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*, pp. 167–172, 2018. doi: 10.1109/CSPA.2018.8368706

[41] J. Sun and N. Buys. Health benefits of Tai Chi. *Canadian Family Physician*, 62(11):881–890, 2016.

[42] G. Wilson and M. McGill. Violent video games in virtual reality: Re-evaluating the impact and rating of interactive experiences. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '18, p. 535–548, 2018. doi: 10.1145/3242671.3242684

[43] B. G. Witmer and M. J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3):225–240, 06 1998. doi: 10.1162/105474698565686

[44] A. S. Won, B. Perone, M. Friend, and J. N. Bailenson. Identifying Anxiety Through Tracked Head Movements in a Virtual Classroom. *Cyberpsychology, Behavior, and Social Networking*, 19(6):380–387, 2016. doi: 10.1089/cyber.2015.0326

[45] J. W. Woodworth and C. W. Borst. Design and validation of a library of active affective tasks for emotion elicitation in vr. In *Proceedings of the 2nd Momentary Emotion Elicitation and Capture Workshop*, 2021.

[46] T. Xue, A. E. Ali, T. Zhang, G. Ding, and P. Cesar. Ceap-360vr: A continuous physiological and behavioral emotion annotation dataset for 360 vr videos. *IEEE Transactions on Multimedia*, 2021. doi: 10.1109/TMM.2021.3124080

[47] T. Yamakoshi, K. Yamakoshi, S. Tanaka, M. Nogawa, M. Shibata, Y. Sawada, P. Rolfe, and Y. Hirose. A preliminary study on driver's stress index using a new method based on differential skin temperature measurement. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 722–725, 2007. doi: 10.1109/IEMBS.2007.4352392