*Journal of the Brazilian Computer Society, 2026, 32:1,* doi: 10.5753/jbcs.2026.8215


# Effects of task difficulty and musical training in Virtual Reality: Observations of cognitive load and task accuracy in a VR rhythm exergame

**Kyla Ellahiyoun** [ **RMIT University** | ***kyla.ellahiyoun@hotmail.com*** ]
**Emma Jane Pretty** [ **Tampere University** | ***emma.pretty@tuni.fi*** ]
**Renan Guarese** [ **Maynooth University** | ***guarese@acm.org*** ]
**Marcel Takac** [ **RMIT University** | ***marcel.takac@rmit.edu.au*** ]
**Haytham Fayek** [ **RMIT University** | ***haytham.fayek@rmit.edu.au*** ]
**Fabio Zambetta** [ **RMIT University** | ***fabio.zambetta@rmit.edu.au*** ]

*Research Centre of Gameful Realities, Tampere University, Kalevantie 4, Tampere, 33100, Finland.*

**Received:** 13 May 2026 • **Accepted:** 05 August 2026 • **Published:** 14 August 2026

**Abstract** This study explores the relationship between musical training, cognitive load (CL), and task accuracy within the virtual reality (VR) rhythm exergame Beat Saber across increasing task difficulty levels. Participants ($N = 32$) had their physiological metric data measured while playing the VR rhythm exergame under three task difficulty levels, completing a series of questionnaires after each task. Using regression analyses, we found that task difficulty and gaming experience significantly predicted perceived mental demand, whereas musical training did not. Musical training, along with lower perceived mental demand, increased gaming experience, and greater physiological arousal, significantly predicted higher task accuracy. These results suggest that musical training enhances task-specific performance but does not directly reduce perceived mental demand. Future research should consider alternative methods of grouping musical training and the additional predictability of flow and self-efficacy.

**Keywords:** Virtual Reality, Exergames, Biofeedback, Physiological Sensors, Cognitive Load, Human-Computer Interaction

# 1 Introduction

Music psychology has long been investigating the effect of music on cognitive functions. This has resulted in the realisation that music has positive outcomes on executive functioning [Chee *et al*., 2022].Recently, this interest in cognitive outcomes is extended to observing cognitive load (CL) during music-oriented tasks. CL accounts for limitations of working memory and considers that the delivery mode of novel information should follow the scientific principle of parsimony [Sweller *et al*., 1998], being presented with the simplest explanation. This is important when presenting new information or tasks, as comparing the expertise between individuals or groups of people can be explained through theories of optimal performance [Yerkes and Dodson, 1908; Csikszentmihalyi, 1975]. In relation to CL and expertise, there is evidence of a negative relationship between a higher level of skill and acquired cognitive load during challenging tasks [Haith and Krakauer, 2018]. CL is prominently researched with regards to instructional design [Sweller *et al*., 1998], where it is established that the inherent task difficulty and its delivery method could significantly affect the accumulation of CL [Sweller *et al*., 1998]. To address this effect, some researchers have opted for commercial games that are recognisable to the public [Huang, 2020]. A notable example being Beat Saber[1], a popular VR rhythm exergame [Mueller and Isbister, 2014] that combines upbeat music and physical movement. There have been few attempts to modify games such as Beat Saber or create new rhythm games for research participants to engage with [Lemmens and von Münchhausen, 2023], however, they are susceptible to limited generalisation as they can lack elements which make games inherently enjoyable [Cutting *et al*., 2023].

[1]`https://www.beatsaber.com/`

## 1.1 Music Sophistication & Training

An abundance of research has defined musical abilities by the duration of formal musical training (MT), ignoring other factors that do not overtly contribute to the skillset [Levitin, 2012; Guarese *et al*., 2022]. A widely adopted measure of musical ability in the general population is the Goldsmiths Music Sophistication Index (Gold-MSI) [Müllensiefen *et al*., 2014]. It observes music sophistication as a psychometric construct, encompassing independent dimensions comprising skills and achievements, musical perception and music-making, amount of practice, emotional and functional usage of music, and creativity [Müllensiefen *et al*., 2014]. Despite this, its MT subscale alone is frequently used to group participants, likely due to its perceived objectivity and alignment with more traditional conceptions of musical ability. For example, studies have chosen certain items from the MT subscale to categorise low, intermediate, and high music sophistication groups [Matziorinis *et al*., 2023; Guarese *et al*., 2024], or only considered MT to inform their definition of expertise [Chaddock-Heyman *et al*., 2021; Guarese *et al*., 2022]. As a result, inconsistencies in how musical training is operationalised across studies complicate direct comparison of findings, leading to an overlook of the broader cognitive or experience factors that contribute to task performance.

### 1.2 Cognitive Load Theory

Cognitive Load Theory was conceived because of competing learning and memory frameworks [Atkinson and Shiffrin, 1968; Baddeley and Hitch, 1974], demonstrating its adaptability for instructional design [Sweller *et al*., 1998]. Furthermore, by accounting for limitations of working memory, the delivery mode of novel information should be parsimonious, following the simplest explanations [Sweller *et al*., 1998]. This notion is supported by widely accepted assumptions within CL research, such as element interactivity [Sweller, 1994, 2010] and schema acquisition [Chi *et al*., 1982]. Firstly, CL is considered a multidimensional construct, which describes the workload that performing a given task imposes on a learner's cognition [Paas and Van Merriënboer, 1994; Sweller *et al*., 1998]. Continuing from this, element interactivity refers to the number of elements being processed simultaneously in working memory for schema construction and their interactions to occur [Sweller, 1994], essentially describing the inherent task difficulty. Furthermore, schema acquisition relies on long-term memory to guide behaviour from a repository of stored information [Kalyuga and Singh, 2016]. As both of these processes are shaped by prior knowledge and experience, researchers exploring the relationship between CL and task performance increasingly acknowledge the role of domain-specific expertise.

### 1.3 Expertise in Context

Expertise is characterised as a measurable degree of skill, whereby a necessary level of skill or knowledge is objectively referred to as competency [Greenfield *et al*., 1994]. In acknowledgement of this, Kirschner and Williams [2013] argued that expertise should reframe the expert versus novice dichotomy by positioning players as experts in their own understanding. It competes with the notion of invalidating the experiences of beginners by focusing on the agency of 'experts' [Kirschner and Williams, 2013]. Following their example, Toft-Nielsen [2016] argued that to acquire expertise in games, the amount of knowledge required to become an expert should be considered. Accordingly, research measuring expertise in virtual reality (VR) exergames should appraise the contextual parameters used to define expertise. In the present study, participants are grouped using the MT subscale of the Gold-MSI, but our focus extends beyond music-specific skill. In addition to musical training, experience levels in digital games, VR, and Beat Saber are included as indicators of broader task-relevant expertise. Taken together, these varying forms of experience may help explain differences in both perceived mental demand and task accuracy, particularly in interactive, high-paced environments.

In this study, we examine how musical training, gaming experience, and task difficulty interact to influence CL and performance in a VR rhythm exergame. Using Beat Saber as a testbed, participants completed gameplay sessions of increasing task difficulty levels while physiological metrics (electrodermal activity; EDA) and subjective measures of CL (cognitive subscale of the video game demand scale; VGDS [Bowman *et al*., 2018]) were collected. This study adds to the growing body of work on individual differences in immersive gameplay experiences, making the following contributions:

- A user study and analysis on how domain-specific experience and task difficulty impact both perceived mental demand and task performance;
- Empirical evidence that previous musical training leads to higher task accuracy in a VR rhythm exergame environment;
- A discussion highlighting the utility of EDA as a physiological metric predictor of performance in VR rhythm exergame contexts.

We expect these contributions to aid in the discussion of how we can augment performance prediction with physiological metric data in complex and highly interactive environments, as well as the adaption of games to prior expertise. The remainder of this paper is structured as follows: Section 2 reviews prior work on CL and expertise in games, exergames, and VR. Section 3 defines the research questions guiding this study, while Section 4 outlines the experimental design, task conditions, and metrics used. Section 5 presents our key findings, followed by our interpretation of them and their limitations in Section 6, and finally concluding in Section 7.

## 2 Related Works

As the present study seeks to understand expertise within cognitive engagement, the discussion changes course to consolidate motivational theories into practice [Moreno and Mayer, 2007; Huang *et al*., 2021].

### 2.1 Cognitive Load in games

A popular use of CL is adapting game or character behaviour in real-time using physiological metrics [Yannakakis *et al*., 2016; Pretty *et al*., 2023a]. Examples include measuring heart rate, stress, and cognitive effort, which can be integrated into different game elements to modify the player experience [Mandryk *et al*., 2006; Nacke *et al*., 2011; Robinson *et al*., 2020]. Besides real-time adaption, studies have tried to understand the relationship between subjective [Bowman *et al*., 2018; Pretty *et al*., 2024b] and objective metrics [Nacke *et al*., 2011; Pretty *et al*., 2024a] of CL, i.e., whether the user's self-assessed perception of their mental state correlates to its physiological metric. Pretty *et al*. collected multiple physiological metrics during gameplay (including electroencephalography, electromyography, heart rate, heart rate variability, EDA, and eye blink rate), analysing them against one another [Pretty *et al*., 2024a], and also against subjective measures: the Task Load Index [Hart and Staveland, 1988] (NASA-TLX) and VGDS [Bowman *et al*., 2018] [Pretty *et al*., 2024b]. Their comparisons indicated strong correlations in some of these pairings, one of which includes EDA, which we aim to reproduce in our study, although under our VR rhythm exergame scenario.

Concerning Beat Saber specifically, Tammy Lin *et al.* [2023] tested the effect of different playable angle conditions on the psychological and physical activities of participants. Although they also used the VGDS to measure demand, they did not include results from physiological metric sensors, other than an accelerometer measuring spatial displacement to indicate physical activity. Another study investigated hand-eye coordination in a Beat Saber context proposed that VR rhythm exergame training might lead to faster music instrument mastery [Rutkowski *et al.*, 2021]. However, that effect is outside the scope of this study, as our experiment aims to combine different research areas in the context of VR.

For a comprehensive review of multi-dimensional CL measurements, across a variety of tasks and inclusive of all categories of CL measurement, see [Chen *et al.*, 2016].

### 2.2 Exergames in VR

VR offers pedagogical means for users to learn, engage, and enhance their knowledge and skills within a simulated immersive environment. These digital simulations provide a psychological experience dependent on interplays between technology and perceptual processes [Steuer, 1992; Takac *et al.*, 2023; Renata *et al.*, 2024]. They can contribute to a reduction in confounding variables that are typical of laboratory and clinical settings [Renata *et al.*, 2024]. As attributable to music, VR is a primary medium within entertainment and video games that seeks to provide an engaging experience to the extent of distracting a user from external stimuli [Schmidt *et al.*, 2018]. Extending from this, previous studies have found that situating exergames within a VR context is not only a promising intervention for encouraging physical exercise [Huang, 2020; Hedlund *et al.*, 2023a; Yoo *et al.*, 2020] but also provides a more immersive experience and feeling of agency than traditional exergames [Campo-Prieto *et al.*, 2021].

Several works have explored the exertion aspects of specific sports in an immersive context [Gradl *et al.*, 2016; Sawan *et al.*, 2021], such as jogging [Hedlund *et al.*, 2023b], climbing [Kajastila *et al.*, 2016], rowing [van Delden *et al.*, 2020; Arndt *et al.*, 2018], skiing [Miura *et al.*, 2023], and cycling [Wintersberger *et al.*, 2022]. A significant concern about exertion in VR is attributable to its physically intensive nature, which could lead to discomfort or severe after-effects from prolonged use, such as cybersickness [LaViola, 2000]. It is defined as symptoms of eye strain, headache, sweating, disorientation, vertigo, nausea, or vomiting and can occur strictly from visual perception alone [LaViola, 2000]. A standardised measure for cybersickness in VR research is the Simulator Sickness Questionnaire (SSQ) [Kennedy *et al.*, 1993], originally developed to assess simulator sickness in aviators, which will be applied in the current experiment, being in line with current VR experimental protocols for movement-based experiences [Hedlund *et al.*, 2023a,b; Wintersberger *et al.*, 2022].

Beyond cybersickness, Arndt *et al.* [Arndt *et al.*, 2018] used a breathing sensor to gather physiological metric data during their rowing simulation. In their VR condition, participants showed significantly better results in breaths per minute and rhythm–which are crucial in rowing performance–, when compared to a non-VR baseline. Given that rhythm is highly relevant in our VR rhythm exergame scenario, we will also collect physiological metric data, which is highly interconnected to physical exertion [Elvitigala *et al.*, 2024]. Multimodal measurement of the target construct, CL, also has benefits for conducting a more holistic exploration of this phenomenon, and thus we also measure EDA, which has also been linked to CL [Pretty *et al.*, 2024a].

## 3 Research Aim and Questions

Although aspects of music sophistication and CL have been jointly researched [Benz *et al.*, 2016], there are limited studies that attempt to explore this interaction in a VR rhythm exergame [Tammy Lin *et al.*, 2023]. As the interaction between music-related activities and executive function activation is well-established in music psychology literature [Benz *et al.*, 2016], and working memory is a component of executive function [Schnotz and Kürschner, 2007], this could posit CL as a subjective measure of executive function. Furthermore, game difficulty in Beat Saber is referred to as task difficulty to align with terminology used in existing literature. Thus, the present study aims to investigate the interactions of music sophistication, CL, and task accuracy in a VR rhythm exergame as task difficulty increases, given the following research questions (RQ):

- **RQ1:** What are the effects of task difficulty and game/VR experience on cognitive load?
- **RQ2:** How does previous musical training affect...
  - **a)** performance, and
  - **b)** cognitive load

  ...in a rhythm exergame?

## 4 Methodology

### 4.1 Participants

Recruitment of participants was performed via word of mouth, social media websites, and advertising posters at the local university. Participants were screened relative to equipment safety guidelines[2] and demographic characteristics in VR studies [Slater and Sanchez-Vives, 2016; LaViola, 2000]. The eligibility criteria were as follows: (1) being 18 to 35 years old; (2) being a near-native level proficient English speaker; (3) not being currently pregnant; (4) having no vision impairments that cannot be corrected by glasses or contact lenses; (5) having not recently undergone any medical procedure (including cosmetic); (6) not having a pacemaker; (7) not suffering from a heart or serious medical condition; (8) having no physical disabilities that impair balance, motor action, or standing with assistance; and (9) having no history of significant motion sickness, active nausea, and vomiting or epilepsy. Out of the people who applied and met the screening criteria above, 32 participants participated in the present study. They were given the opportunity to enter a draw for one of two gift

[2] `https://www.meta.com/quest/safety-center/quest-pro/`

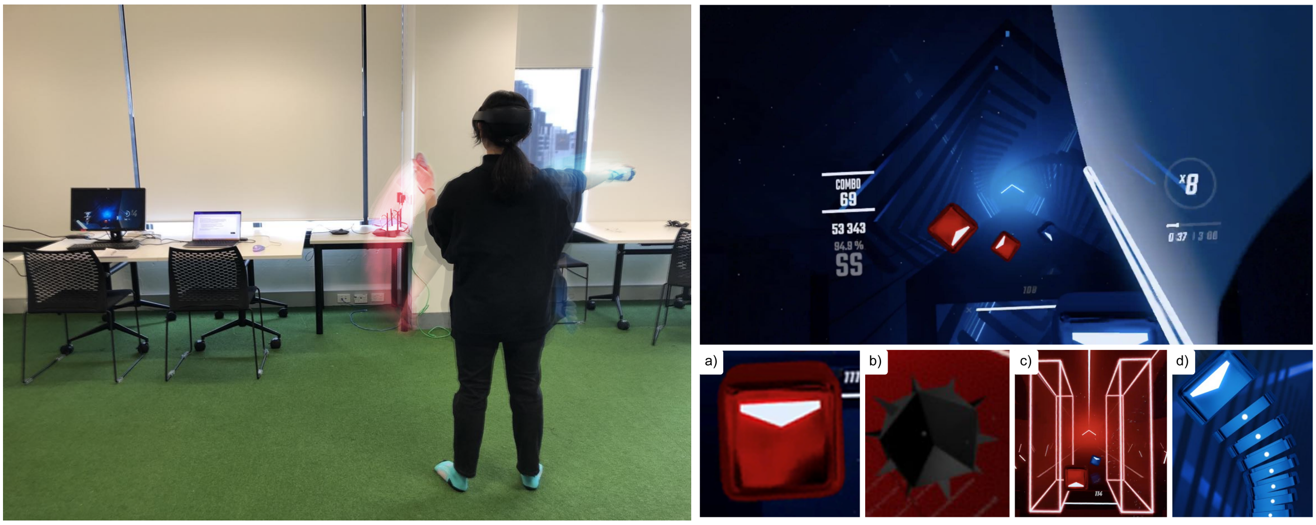


**Figure 1.** Left – Superimposed photo of the experimental layout: The screen on the left displays the casting of Beat Saber from the MetaQuest Pro, which is being recorded on OBS. The laptop is used for all surveys, and the participant stands in an open space unobstructed by objects during gameplay. Right – Screenshots from Beat Saber and its main interactable Objects: a) blocks which represent notes of a song, b) mines which players should avoid, c) walls (an obstacle) that players' head should avoid, and d) chain notes which represent sounds that are continuous.

cards at the completion of the study. The sample consisted of 11 females (34.4%), 18 males (56.3%), and 3 non-binary (9.4%) people aged 18 to 35 years ($M = 23.63, SD = 4.58$). However, the EDA data of five participants had too much noise and the features could not be extracted, making the final number of participants 27. In addition, the MT group split was as such: 13 participants in the low group, and 14 participants in the high group.

**Table 1.** Frequency Statistics of Participants' (N=32) Experience with Digital Games, VR, and Beat Saber

| Measure | Score | $f$ | Relative $f$ | $cf$ | Perc. |
|---|---|---|---|---|---|
| | 7 | 17 | 53.13% | 32 | 100 |
| | 6 | 8 | 25.00% | 15 | 46.88 |
| | 5 | 5 | 15.63% | 7 | 21.88 |
| **Dig. Games exp.** | 4 | 1 | 3.13% | 2 | 6.25 |
| | 3 | 1 | 3.13% | 1 | 3.13 |
| | 2 | 0 | 0.00% | 0 | 0.00 |
| | 1 | 0 | 0.00% | 0 | 0.00 |
| | 7 | 1 | 3.13% | 32 | 100 |
| | 6 | 3 | 9.38% | 31 | 96.88 |
| | 5 | 3 | 9.38% | 28 | 87.50 |
| **VR exp.** | 4 | 11 | 34.38% | 25 | 78.13 |
| | 3 | 9 | 28.13% | 14 | 43.75 |
| | 2 | 2 | 6.25% | 5 | 15.63 |
| | 1 | 3 | 9.38% | 3 | 9.38 |
| | 7 | 0 | 0.00% | 32 | 100 |
| | 6 | 2 | 6.25% | 32 | 100 |
| | 5 | 1 | 3.13% | 30 | 93.75 |
| **Beat Saber exp.** | 4 | 12 | 37.50% | 29 | 90.63 |
| | 3 | 5 | 15.63% | 17 | 53.13 |
| | 2 | 1 | 3.13% | 12 | 37.50 |
| | 1 | 11 | 34.38% | 11 | 34.38 |

1 = "Never played"; 2 = "Not played in the last 5 years"; 3 = "Played at least once in the last 5 years";
4 = "Play at minimum once a year"; 5 = "Play at minimum once a month"; 6 = "Play at minimum once a week"; 7 = "Play daily".

## 4.2 Materials

### 4.2.1 Demographic measures

The pre-experimental survey included questions pertaining to participants' previous experience with digital games, VR, and Beat Saber on a 7-point Likert scale, from "Never played" to "Play daily" (more details on Table 1). The questions were as follows: "What best describes your experience with (1) digital games; (2) Virtual Reality; and (3) Beat Saber?".

**Gold-MSI** The Gold-MSI [Müllensiefen *et al*., 2014] is a 38-item self-report measure of music sophistication comprising five subscales (active engagement, perceptual abilities, MT, singing abilities, and emotions) that contribute to a general music sophistication factor. Out of those, 31 items are scored on a Likert scale ranging from 1 (completely disagree) to 7 (completely agree), with the remaining 7 items being scored under categorical responses (e.g., "I have had formal training in music theory for 0/.5/1/2/3/4-6/7-10/11+ years."). The primary Gold-MSI, derived from 18 items, demonstrated good internal reliability $\alpha = .88$, with its subscales exhibiting acceptable to good internal reliability between $\alpha = .71$ and $\alpha = .85$ [Chee *et al*., 2022], where a higher score signifies higher music sophistication. Responses were scored using the Gold-MSI scoring template[3] for further analyses. Due to time constraints in this study, the Gold-MSI listening tasks were excluded, and only the derived general and MT scores were analysed. It is acknowledged that less conventional musical training exists that cannot be captured by the Gold-MSI alone, requiring the development of an extensive questionnaire. For this reason, we also captured experience with Beat Saber specifically, as mentioned above.

**Video Game Demand Scale** The VGDS is a multidimensional instrument designed to assess perceived demand across five domains: cognitive, emotional, controller, exertional,

[3] `https://shiny.gold-msi.org/gmsiscorer/`

and social [Bowman *et al.*, 2018]. It has demonstrated strong internal consistency (e.g., cognitive: $\alpha = .90$) and good convergent validity with the mental demand subscale in NASA-TLX ($\beta = .927$). The VGDS also shows predictive validity, with each subscale significantly linked to corresponding self-reported effort (e.g., Paas' mental effort item predicting cognitive load, $r = .817$). Although we administered the full scale (excluding the social subscale), this study reports on the cognitive subscale as our subjective measure of CL.

**Simulator Sickness Questionnaire** Research in VR warrants the monitoring for simulator sickness or cybersickness symptoms as a necessary duty of care. The SSQ [Kennedy *et al.*, 1993] is a 16-item questionnaire with three subscales (nausea, oculomotor, and disorientation) scored on a 4-point Likert scale from 0 to 3 (0 = none; 1 = slight; 2 = moderate; 3 = severe), and it demonstrates a high reliability of $\alpha = .91$ [Xu *et al.*, 2020]. The SSQ was administered after each VR play session for the researcher to determine the participants' ability to continue with the study.

#### 4.2.2 Apparatus

**Physiological metric sensor** **EDA is a measure of sympathetic arousal, and has strongly been linked with** cognitive load in video game and non-video game studies, making it appropriate for exploration in a VR rhythm exergame context [Setz *et al.*, 2010; Vanneste *et al.*, 2021; Pretty *et al.*, 2024b]. The Emotibit[4] has 2 Ag/AgCL EDA electrodes for collecting a number of different physiological metrics. It uses an Adafruit Feather M0 WiFi board powered by a 400mAh Lithium ion battery, a microSD card to record the data, and it comes with straps for different placements on the body. In our experiment, it was attached to the participants' right ankle, with them being instructed to keep their feet still as much as possible. Typically, EDA sensors are placed on the finger or wrist however this would introduce large movement artefacts due to the nature of the VR rhythm exergame and risked the Emotibit data becoming completely unusable and thus the least stationary and most appropriate location for the sensor was chosen. The Emotibit Oscilloscope (v1.5.10) was used to stream and record the data, whereas the Emotibit DataParser (v1.11.4) was used to parse the datastream for analysis.

**Computer hardware** The MetaQuest Pro[5] was used as a VR head-mounted display due to it being untethered, with its own central processing unit. As shown in Figure 1-left, the surveys and the recording of Beat Saber version 1.31.0 gameplay were completed on separate computers to reduce distraction and ensure the clear separation between task and survey completion, with the recordings being completed using the Open Broadcaster Software[6] (OBS version 29.1.3) saved in MP4 format.

**Game environment** Beat Saber was the VR rhythm exergame environment used in this experiment and **it was chosen for a number of reasons. Its core mechanic requires timing directed movements to the beat of the music, so rhythm perception and audio-motor synchronisation developed through musical training are plausibly relevant to performance. The VR rhythm exergame produces a precise score at the end of each song based on a number of factors mentioned below, which we normalise into a task accuracy percentage comparable across songs and task difficulty levels. Finally, it is a widely played commercial title, thus increasing the ecological validity of the study, which purpose-built research games often lack [Cutting *et al.*, 2023; Pretty *et al.*, 2023b], and an established platform in exergames research, thus attempting to contributing to a multidimensional understanding of this VR rhythm exergame from a research perspective [Huang, 2020; Tammy Lin *et al.*, 2023].** The VR rhythm exergame requires players to react to visual cues of approaching coloured blocks by performing slicing movements to the rhythm or beat of a piece of music. Each block represents a note and it needs to be cut in the direction (horizontally, vertically, or diagonally) indicated by an arrow on each block. Figure 1-right demonstrates the virtual environment and objects players can interact with. The scoring system depends on the swinging angle movement when cutting a note and how precise a cut is through the centre of a note. Since each song has a variable number of notes, and the task difficulty being played highly impacts on the amount of points each cut grants the player, we will use a ratio to normalise the scoring as a metric of accuracy. Hence, we define task accuracy as the percentage of the score obtained by the player divided by the maximum possible score in that specific song and task difficulty. The ranking system in the VR rhythm exergame is based on this same task accuracy, divided in percentiles; however, this is only accessible during gameplay without modifying the VR rhythm exergame.

Previous studies did not have congruent song choices to reference from. By means of consulting online game forums[7] and extensive playthroughs by the authors, the songs selected needed to be distinguishable in task difficulty (see Table 2). Similar song durations for the three experimental conditions (easy, normal, and hard) were chosen to control for differences in exhaustion beyond task difficulty. For consistent gameplay, the following changes were made to the VR rhythm exergame settings:

- Switched on: "no fail", "automatic player height", fixed note (block) colours (red = right, blue = left), and the VR rhythm exergame environment was set to "The First".
- Switched off: "arc haptics", "arc visibility", "environment effects" and "flickering".

**Besides the settings listed above, we also controlled song selection, task difficulty order, song durations, tutorial exposure, hardware, and sensor placement. The note charts, scoring algorithm, audio, and visual environment are authored by the VR rhythm exergame's developers and were used as shipped. We report the VR rhythm exergame version, songs, chart statistics (Table 2), and all modified settings so the task can be reconstructed exactly.**

[4] `https://www.emotibit.com`
[5] `https://www.meta.com/quest/quest-pro/`
[6] `https://www.obsproject.com/`
[7] `https://reddit.com/r/beatsaber/comments/rpb8i7/beat_saber_music_packs_from_best_to_worst/`

**Table 2.** Beat Saber Song Details for Experimental Conditions[a] Tutorial, Easy, Normal and Hard

| | **Tutorial** | **Easy** | **Normal** | **Hard** |
|---|---|---|---|---|
| **Song Choice** | Balearic Pumping | Rum n Bass | POP/STARS | Natural |
| **Song Artist** | Jaroslav Beck | Boom Kitty | K/DA | Imagine Dragons |
| **Duration (minutes)** | 02:14 | 03:09 | 03:09 | 03:06 |
| **Beats Per Minute** | 111 | 132 | 170 | 100 |
| **No. of Notes**[b] | 170 | 155 | 341 | 553 (526)[c] |
| **Notes per second** | $\sim$1.26 | $\sim$0.82 | $\sim$1.80 | $\sim$2.83 |
| **No. of Walls** | 9 | 23 | 41 | 0 |
| **No. of Mines** | 6 | 2 | 2 | 0 |

a) 'Experimental Conditions' align with the task difficulty levels of the songs played, the only exception being the Tutorial, as participants played a song on 'Easy' task difficulty level to warm up. b) 'Notes' refers to the coloured blocks players interact with in the VR rhythm exergame. c) Although there are 526 notes in this song, the maximum combo participants could achieve is 553 due to chain notes, making that a more accurate representation for the 'number of notes' in this song.

## 4.3 Procedure

This study was approved by the Human Research Ethics Committee at RMIT University (Project Number: 26489). Eligible participants provided written informed consent upon arrival to the lab. They completed a survey battery prior to experiencing the VR conditions in the following order (1) demographic data about their previous experience with digital games, VR, and Beat Saber, and (2) the Gold-MSI. Participants were provided a tutorial on how to use the MetaQuest Pro and how to play Beat Saber, which consisted of a tutorial accompanied by an Easy level song to reduce novelty effects for consecutive conditions through gameplay exposure. For each condition, participants were instructed to press 'play' after a count of 5 when the researcher said 'Go'. All participants played the same songs in the same order of task difficulty levels: from easiest to hardest (see Table 2). **Although a learning effect is expected, we believe counterbalancing the task difficulty order would hinder the experience of novice players, in an effort to minimize the difference with players who already had previous experience**. At the completion of each song, participants reported their scores verbally before removing the headset. These scores were also double-checked by the researchers in the OBS recording after the experiment, and then converted to a percentage over the maximum possible score for that song and task difficulty, regarded as task accuracy. Participants were then asked to respond to the SSQ and VGDS. After all conditions were done, they would complete a post-experiment survey, which asked participants to rank their preference of task difficulty order by level of enjoyment, and the reasoning for their choice.

## 4.4 Design

The present study employed a within-subjects experimental design to evaluate changes to dependent variables across the main independent variable of task difficulty.

## 4.5 Data Analysis

### 4.5.1 EDA Pre-Processing

EDA was recorded at 15 Hz and pre-processed using the *neurokit2* Python package. The raw signal was first smoothed with a fourth-order low-pass Butterworth filter (cutoff: 5 Hz) and then decomposed into tonic and phasic components using the *eda_process* function, which applies continuous deconvolution. Skin conductance responses (SCRs) were identified using a peak detection algorithm with adaptive thresholding and a minimum amplitude criterion of 0.01 $\mu$S. From this, the following measures were derived for each trial: average tonic EDA, average phasic EDA, number of SCR Peaks, and average SCR amplitude. All EDA variables were standardised ($M = 0, SD = 1$) to enable direct comparison across each other.

### 4.5.2 Statistical Analysis

All statistical analyses were conducted in R (v4.4.1). Data were screened for missing values, and observations with incomplete data for the variables of interest were excluded, as mentioned above, this was present in five participants' EDA data. Multivariate normality of the EDA variables was assessed separately for each task difficulty condition using Mardia's test. Results showed that assumptions of multivariate normality were not fully met, with only the SCR Peaks variable conforming to normality across conditions and thus SCR Peaks was the only EDA variable kept for the final regression models.

Multiple ANOVA and MANOVA tests were initially conducted to ascertain which variables were suitable for the two linear regression models that aimed to see which objective and subjective data could predict task accuracy and perceived mental demand. Variables were deemed suitable if they could display significant (or near-significant) differences across the 3 task difficulty levels as specificity of the measurement is a prerequisite for our analysis. A one-way MANOVA was conducted with task difficulty (easy, normal, and hard) as the independent variable, and four standardised EDA indices: tonic EDA, phasic EDA, SCR Peaks, and SCR amplitude as dependent variables. A follow-up ANOVA and post hoc tests were conducted for variables showing significant effects. Perceived mental demand, as measured by the cognitive subscale of the VGDS, was also compared across task difficulty conditions using a one-way ANOVA, with follow-up post hoc tests. This was also done for comparing task accuracy across task difficulty levels. Additional regression analyses were performed to examine the predictive value of prior experience variables (Beat Saber, VR, and digital game experience) on task accuracy, perceived mental demand, and SCR Peaks.

Finally, to address the research questions of this study, two multimodal linear regression models were conducted to evaluate the predictive value of task difficulty, Beat Saber experience, VR experience, digital games experience, MT group, and SCR Peaks on perceived mental demand and task accuracy (with the addition of perceived mental demand to the latter model). Interaction effects were tested for the particular variables of interest (MT group & CL, and MT group & task difficulty), however none were significant, and thus the final models only included main effects. **This has added benefit of reducing multicollinearity, as well as simplifying the models. Adjusted $R^2$ values were used as effect sizes to demonstrate the amount of variance explained by the models.**

# 5 Results

## 5.1 Demographics

Frequencies of responses to pre-experimental survey questions can be found on Table 1, which describes players' experience with digital games, VR, and Beat Saber on a 7-point Likert scale. On average, 78.1% participants played digital games once a week or more ($M = 6.22, SD = 1.04$), 56.27% have played in VR at least once a year ($M = 3.69, SD = 1.45$), and 62.5% have played Beat Saber at least once in the last 5 years ($M = 2.91, SD = 1.59$). An independent samples t-test found no statistically significant difference between male ($M = 72.78, SD = 4.48$) and female ($M = 69.18, SD = 4.02$) participants for general musical sophistication scores ($p = .588$). Participants' mean SSQ scores were within the expected range across all conditions: easy ($M = 8.32, SD = 2.51$), normal ($M = 12.06, SD = 3.24$) and hard ($M = 16.29, SD = 4.57$). Two participants scored above the standard threshold and were monitored to determine their fitness to continue the study.

## 5.2 Qualitative Feedback

Following the main experiment, participants ranked the task difficulty levels by which they found the most enjoyable accompanied by reasoning. Of interest was their highest ranked task difficulty, which was hard ($56\%, n = 18$), normal ($31\%, n = 10$), and easy ($13\%, n = 4$). Feedback from the hard condition was clustered thematically: the need to concentrate ("the harder it was, the more I had to actually focus"), high stimulation ("it was a fun and novel challenge") and flow ("appropriate task difficulty for my skill level"; "…helped me to get into the zone"). The normal task difficulty was described as "the perfect balance for my current expertise in the game", in addition to it being the "right amount of task difficulty for the game". Participants who enjoyed the easy task difficulty the most emphasised on performance, stating it was "easier to slash the cubes, resulting in [a] higher score and better ranking".

## 5.3 Multivariate Analysis of Variance of EDA by Task Difficulty

A MANOVA with task difficulty as the predictor revealed a non-significant multivariate effect on the combined EDA indices (Pillai's Trace $= .137, F(8, 138) = 1.27, p = .267$). Follow-up univariate ANOVAs indicated that only SCR Peaks differed significantly by task difficulty ($F(2, 71) = 3.54, p = .034$), with an almost marginal effect also found for SCR Amplitude ($F(2, 71) = 2.95, p = .059$). Post hoc comparisons (Tukey HSD) for SCR Peaks indicated a significant difference between easy and hard tasks ($p = .026$), with more SCR Peaks during hard tasks (see Table 3-top). No significant differences were found between normal and either easy or hard conditions. Based on this and the previous normality results, SCR Peaks was chosen for both of the final regression models.

## 5.4 Effects of Task Difficulty on Task Accuracy and Perceived Mental Demand

A one-way ANOVA revealed a significant effect of task difficulty on perceived mental demand ($F(2, 93) = 18.62, p < .001$). Post hoc Tukey HSD tests confirmed that participants reported significantly higher CL in hard tasks compared to both normal ($p = .002$) and easy tasks ($p < .001$), and also reported significantly higher CL in normal compared to easy tasks ($p = .028$), as seen in Table 3-bottom. Assumptions of normality (Shapiro-Wilk $p = .077$) and homogeneity of variance (Levene's $F(2, 93) = 2.11, p = .128$) were met. As these findings indicate, the songs were different in difficulty, and thus the task difficulty variable was included in the final regressions.

A one-way ANOVA revealed a significant effect of task difficulty on task accuracy ($F(2, 93) = 10.88, p < .001$). Post hoc Tukey HSD tests indicated that task accuracy was significantly lower in hard tasks compared to both easy ($p < .001$) and normal tasks ($p = .003$), while no significant difference was observed between easy and normal tasks ($p = .499$). The assumption of normality was met (Shapiro-Wilk $p = .099$), but Levene's test indicated a violation of the homogeneity of variance assumption ($F(2, 93) = 4.78, p = .011$). A non-parametric Kruskal-Wallis test confirmed the effect of task difficulty on task accuracy ($\chi^2(2) = 14.51, p < .001$). Pairwise Wilcoxon tests further supported the finding that task accuracy was significantly lower in hard compared to easy ($p < .001$) and normal tasks ($p = .025$), with no significant difference between easy and normal tasks ($p = .857$), as seen in Table 4.

## 5.5 Experience Factors as predictors of task accuracy, perceived mental demand, and SCR Peaks

A linear regression assessed whether Digital Games, VR, and Beat Saber experience predicted perceived mental demand. The model was significant ($F(3, 92) = 4.81, p = .004$), explaining 10.7% of the variance ($R^2_{adj} = .107$). Both Digital Games ($\beta = -35.09, p = .013$) and Beat Saber experience ($\beta = -32.91, p = .016$) were significant negative predictors,

**Table 3.** Post hoc comparisons for SCR Peaks and Cognitive Load (Tukey HSD)

| Comparison | Mean Difference | 95% CI | $p$ |
|---|---|---|---|
| **SCR Peaks** | | | |
| Hard vs Easy | 0.73 | [0.07, 1.38] | .026* |
| Normal vs Easy | 0.33 | [-0.33, 0.99] | .460 |
| Normal vs Hard | -0.40 | [-1.06, 0.26] | .329 |
| **Cognitive Load** | | | |
| Hard vs Easy | 190.10 | [115.67, 264.54] | <.001*** |
| Normal vs Easy | 81.77 | [7.34, 156.20] | .028* |
| Normal vs Hard | -108.33 | [-182.77, -33.90] | .002** |

**Table 4.** Post hoc comparisons for Task Accuracy (Wilcoxon pairwise tests)

| Comparison | Med. Diff. | Group Medians (IQR) | $p$ |
|---|---|---|---|
| Hard vs Easy | -22.0 | 53.8 (41.5) vs 75.8 (20.4) | <.001*** |
| Hard vs Normal | -24.6 | 53.8 (41.5) vs 78.4 (22.6) | .025* |
| Normal vs Easy | 2.6 | 78.4 (22.6) vs 75.8 (20.4) | .857 |

while VR experience was not ($p = .22$). Multicollinearity was low (all $VIFs < 3$), and visual inspection of histograms and Q-Q plots displayed normal residuals.

Additionally, a linear regression assessed whether Digital Games, VR, and Beat Saber experience predicted task accuracy. The model was significant ($F(3, 92) = 9.91, p < .001$), explaining 21.9% of the variance ($R^2_{adj} = .220$). Digital Games ($\beta = 5.42, p = .006$) and Beat Saber experience ($\beta = 4.08, p = .030$) were significant positive predictors of task accuracy, while VR experience was not ($p = .39$). Multicollinearity was low (all $VIFs < 3$), and visual inspection of histograms and Q-Q plots displayed normal residuals. As such, both Digital Games and Beat Saber experience were significant in these regressions, thus they were included in the final regression models.

A separate regression predicting SCR Peaks from the same experience variables, as well as MT group, was not significant ($F(3, 70) = 0.63, p = .595, R^2_{adj} = -0.015$). The individual experience factors did not meaningfully predict EDA response, however MT group did ($\beta = -0.54, p = .028$). Multicollinearity was low (all $VIFs < 3$), and visual inspection of histograms and Q-Q plots displayed normal residuals.

**Table 5.** Regression coefficients for predictors of Cognitive Load, SCR Peaks, and Task Accuracy

| Predictor | $\beta$ | SE | $p$ |
|---|---|---|---|
| **Final Combined Model Predicting Cognitive Load** | | | |
| Task Difficulty (Hard) | 226.66 | 32.03 | <.001*** |
| Task Difficulty (Normal) | 96.80 | 31.07 | .002** |
| Dig. Games experience | -28.56 | 11.59 | .016* |
| Beat Saber experience | -25.42 | 8.10 | .003** |
| SCR Peaks | -20.17 | 13.30 | .133 |
| MT Group | 21.27 | 26.18 | .420 |
| **Final Combined Model Predicting Task Accuracy** | | | |
| Perceived Mental Demand | -0.053 | 0.01 | <.001*** |
| MT Group | 10.51 | 3.32 | .002** |
| Task Difficulty (Hard) | -12.27 | 5.10 | .018* |
| Task Difficulty (Normal) | -3.81 | 4.88 | .438 |
| Dig. Games experience | 3.72 | 1.48 | .014* |
| Beat Saber experience | 4.26 | 1.07 | <.001*** |
| SCR Peaks | 711.38 | 346.72 | .043* |

## 5.6 Combined Predictive Models of Task Accuracy and Perceived Mental Demand

The final regression model to predict perceived mental demand including task difficulty, Digital Games experience, Beat Saber experience, MT group, and SCR Peaks (see Table 5-top) was also significant, $F(6, 67) = 12.43, p < .001$, explaining 48% of the variance ($R^2_{adj} = .48$). Significant predictors included task difficulty (both normal ($\beta = 96.80, p = .002$) and hard ($\beta = 226.66, p < .001$) vs. easy), Beat Saber experience ($\beta = -25.42, p = .003$), Digital Games experience ($\beta = -28.56, p = .016$), but not SCR Peaks ($\beta = -20.17, p = .133$) or MT group ($\beta = 21.27, p = .42$). Normality testing of residuals showed normal histogram and Q-Q plots, as well as low multicollinearity (all $VIFs < 2$), although the Shapiro-Wilk test indicated non-normal residuals ($p = .024$). For visual representation of this model, see Figure 2.

A final regression model was conducted to predict task accuracy from perceived mental demand, musical training group, task difficulty, Digital Games experience, Beat Saber experience, and SCR Peaks, as seen in Table 5-bottom. The overall model was significant, $F(7, 66) = 20.50, p < .001$, explaining 65.2% of the variance in task accuracy ($R^2_{adj} = .652$). Perceived mental demand was a significant negative predictor ($\beta = -0.053, p < .001$), indicating that higher perceived CL was associated with lower performance. Musical training group significantly predicted task accuracy ($\beta = 10.51, p = .002$), with participants in the high musical training group achieving higher scores overall. Task difficulty also influenced task accuracy, with significantly lower performance on hard tasks compared to easy ($\beta = -12.27, p = .018$), but no significant difference between normal and easy conditions. Both Digital Games experience ($\beta = 3.72, p = .014$) and Beat Saber experience ($\beta = 4.26, p < .001$) were significant positive predictors of task accuracy, and SCR Peaks also contributed modestly to the model ($\beta = 3.39, p = .043$). Residuals were approximately normal based on visual inspection of Q-Q plots and histograms, and multicollinearity was low (all $VIFs < 2$). For a visualisation of the predictors of task accuracy, see Figure 3.

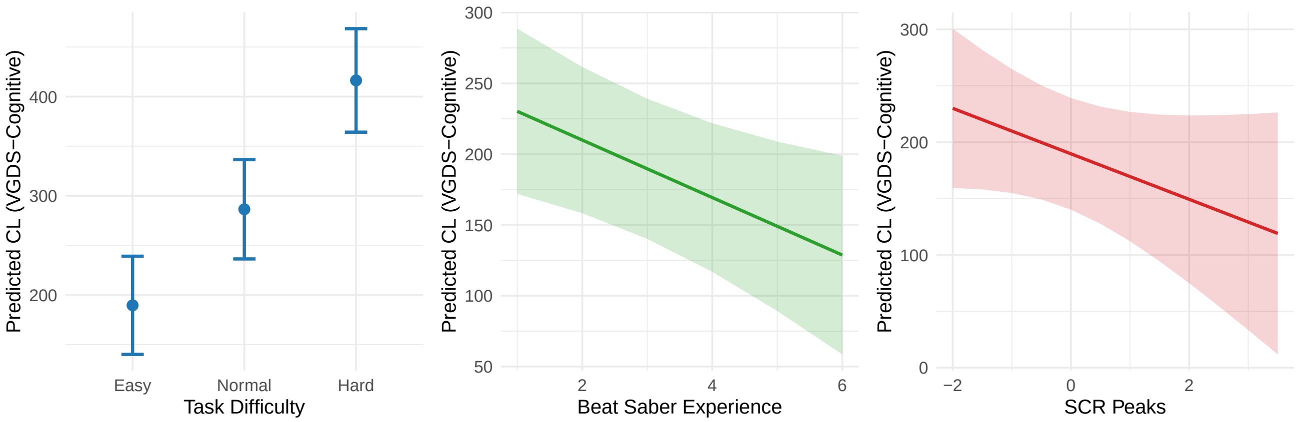


**Figure 2.** Predicting Perceived Mental Demand from Task Difficulty, Beat Saber experience, and SCR Peaks


**Figure 3.** Predicting Task Accuracy from MT Group, Beat Saber experience, and SCR Peaks

# 6 Discussion

The present study investigated whether musical training could explain variations in CL and task accuracy in the VR rhythm game Beat Saber as task difficulty increased. Through regression analyses, we determined that while task difficulty and gaming experience significantly predicted perceived mental demand, musical training alone did not significantly contribute to subjective perceptions of cognitive effort. However, musical training significantly predicted greater task accuracy in gameplay. These results imply that musical training contributes specifically to improved performance, potentially via enhanced motor coordination or visuospatial processing, rather than lowering perceived mental demand.

**Importantly, the exploratory nature of this study implies that these results should be noted with caution as observations within the specific experimental conditions tested and considering a young adult population**.

## 6.1 Task Difficulty, Experience, and Cognitive Load

**Considering RQ1, our regression analyses revealed that task difficulty significantly impacted perceived mental demand, with the hard condition eliciting the highest perceived CL, with high effect sizes being found, followed by normal and easy task difficulty levels, in line with the findings of Tammy Lin *et al.* [2023]. This pattern is consistent with the chart statistics presented in Table 2, in which notes per second more than tripled from the easy to the hard condition (0.82 to 2.83), requiring participants to process an increasing number of elements simultaneously and thereby satisfying the assumption of element interactivity [Sweller, 2010]. Although the number of walls and mines is different across the conditions, we believe that the songs still increase in task difficulty in a way that aligns with cognitive load theory, namely the quantity of information presented Sweller [2011], represented through notes per second and total number of notes. Importantly, specific Beat Saber experience and broader Digital Games experience significantly predicted lower perceived mental demand, suggesting that familiarity with the VR rhythm exergame's mechanics and digital interaction environments effectively reduced perceived mental demand. These effects should be interpreted in light of the experience distributions reported in Table 1, whereby the influence of Digital Games experience was estimated within a predominantly experienced sample, while Beat Saber experience spanned the full range from never-players to weekly players. In contrast, musical training did not significantly predict perceived mental demand, highlighting that task-specific and domain-specific experiences rather than broader musical training influenced participants' subjective experience of CL.**

These findings underscore the critical role of domain-specific familiarity in managing CL within immersive and interactive digital environments. **Furthermore, higher SCR Peaks were associated with better task accuracy. EDA measures sympathetic arousal, and in an active VR rhythm exergame this arousal can come from physical exertion, stress, or excitement as well as from cognitive processing [Elvitigala *et al.*, 2024]. In our models, SCR Peaks predicted task accuracy and failed to predict perceived mental demand, therefore EDA in this task may instead reflect arousal or engagement rather than CL specifically. Regardless, this is still a useful result in that SCR Peaks added predictive value for performance beyond task difficulty and experience with a strong effect being noted, which supports EDA as a performance indicator in rhythm-based VR even if the direct relationship with CL is less clear.** Our subjective data support CL theory's claim that familiarity and expertise shape perceived task difficulty in complex, multimodal tasks such as a VR rhythm exergame.

Although these results underline CL as a key differentiator across task conditions, potential confounding effects from VR novelty should be acknowledged. Considering our participant sample consisted of limited previous VR experience, and despite instructions to report CL based on the task alone, CL results can be conflated by way of exposure to a virtual environment [Anglin *et al.*, 2017; Frederiksen *et al.*, 2020] in the sense that participants may have felt overwhelemed by the VR experience if they were inexperienced, and though this would be a consistent conflation within a singular participant, it may not be true of all participants.

## 6.2 Musical Training & Cognitive Load

As for **RQ2a**, MT group was not a significant predictor of perceived mental demand in the regression model. In contrast, task difficulty, Beat Saber, and Digital Games experience significantly predicted perceived mental demand with **large effect sizes**, with participants reporting higher perceived mental demand in higher task difficulty levels and lower perceived mental demand if they had greater familiarity with the VR rhythm exergame and games generally. This suggests that perceived mental effort was more strongly shaped by the nature of the task and prior experiences than by MT (see Figure 2). At least within the context of our study, MT did not generalise to perceived mental demand, however, one regression found that expertise did predict objective CL. Thus, their inclusion helped account for individual variation, which may interact with task perception in more complex or indirect ways. This incongruence of findings highlights the influence of domain-specific expertise, such as regular exposure to game mechanics, on perceived mental demand. That is, rather than being modulated by broader cognitive skills associated with MT, perceived mental demand in this task appeared more tightly linked to situational familiarity, whereas objective CL was linked more specifically to task-specific difficulty [Sweller *et al.*, 2011].

These findings also did not replicate previous results reported by Pretty *et al.* [2024a], who found stronger associations between subjective and physiological metrics of CL. The discrepancy may be attributed to differences in task complexity, participant familiarity with VR, the distinct physiological metrics used, or the nature of cognitive load posed by Beat Saber relative to the tasks employed in previous studies. Further research should aim to clarify these methodological variations and examine under what conditions different CL measurements converge. **Finally, our findings apply to Beat Saber specifically, and need replication in other rhythm games prior to ensuring generalisation. Beat Saber tests whether players can time their movements to the notes of a song, but it does not test pitch, melody, or the fine motor skills used to play an instrument. The increase in task accuracy we found for the high MT group shows that rhythm skills carry over to this VR rhythm exergame, and rhythm games with different demands may show different results. As musical training is multidimensional, VR rhythm games that require different musical skills could have caused previous music experience to play a significant effect on CL.**

## 6.3 Musical Training & Task Accuracy

In the final regression model predicting task accuracy, related to **RQ2b**, MT group emerged as a significant predictor, alongside task difficulty, SCR Peaks, Beat Saber and Digital Games experience. Participants with higher MT performed more accurately overall, even when accounting for other factors. This aligns with existing research suggesting that MT supports visuospatial processing [Strong and Mast, 2019], a skill essential in Beat Saber's fast-paced virtual environment [Tian *et al.*, 2023].

Task difficulty and game experience also significantly influenced task accuracy–harder tasks reduced task accuracy, while both specific VR rhythm exergame and general game familiarity improved it. Although no significant interaction or moderating factor was found between expertise and perceived mental demand on task performance, the result of musical training and CL predicting task accuracy does corroborate the findings in CL literature, despite no interactions being found and **a medium effect size** [Sweller *et al.*, 2011]. **Additionally, higher SCR Peak values were associated with higher task accuracy (see Figure 3), and due to the nature of EDA measurement, this result open to several interpretations: arousal may facilitate performance in time-sensitive gameplay, better-performing players may exert more physical effort, or both variables may respond to a third factor such as engagement. Because EDA in a VR rhythm exergame may reflect a mixture of cognitive load, and affective and physical demands, our results show that we cannot interpret SCR Peaks as only reflecting CL.**

Visual inspection of task accuracy data (Figure 4) suggests that although the higher MT group exhibited task accuracy levels closer to each other, the low MT group had a wider spread of task accuracy across all task difficulty levels notwithstanding lower medians. This addresses a critical pitfall of the musicians versus non-musicians dichotomy within the literature, which Talamini *et al.* [2017] underscore to be from expertise acquired via informal training, as opposed to formal, and the ignorance of other traits that contribute towards musical abilities. A possible explanation for this observation in our results is the grouping method for low and high MT.

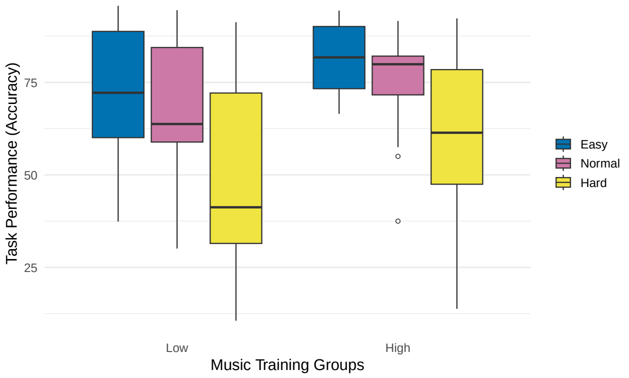


**Figure 4.** Changes in Task Accuracy between Musical Training groups as Task Difficulty increases.

These findings support the need for more holistic classifications of musical ability [Baker *et al.*, 2020; Dahary *et al.*, 2020; Matziorinis *et al.*, 2023], and contribute to the broader discussion around the relationship between MT and executive function [Bowmer *et al.*, 2018].

## 6.4 Limitations & Future Directions

**The exploratory nature of this study imposes inherent limitations for real-world applications beyond providing insight into interactions between CL, task difficulty, and expertise, which results should be noted with caution as observations within the specific experimental conditions tested**. Notably, the small sample size for music research and the method of categorising groups, the lack of a deeper qualitative analysis, and **the use of only a singular VR rhythm exergame have likely impacted the results. As musical training is multidimensional, other VR rhythm games may require different musical skills that Beat Saber does not, which for instance could have caused previous music experience to play a larger effect on CL.**

Firstly, **the sample may be biased due to recruiting on social media and within a university context, resulting in a younger sample than in previous studies. However, our results differ from the literature, which is known for observing CL in young children or older adult populations within educational research** [Yeung *et al.*, 1998; Berends and van Lieshout, 2009; Chaddock-Heyman *et al.*, 2021]). **Beyond age, the experience profile of the sample constrains generalisation. As shown above, 78.1% of participants played digital games at least weekly, while only 12.5% used VR at least monthly and 34.4% had never played Beat Saber. The sample therefore represents experienced gamers who were largely novices in VR and in this specific VR rhythm exergame. The experience variables were included as predictors in the final models, but a sample containing more VR-experienced users or Beat Saber experts would increase generalisation. In addition, the low number of participants and high number of predictors likely led to an underpowered and potentially overfitted regression model, although all predictors were tested independently before the final regression model, validating their final inclusion (of which all VIFs were low).**

Secondly, MT groups were categorised by a singular item, as opposed to multiple or an entire subscale, which is a deviation from previous research [Dahary *et al.*, 2020; Matziorinis *et al.*, 2023]. In supplement to this, sociodemographic factors such as occupation and income were not assessed despite eliciting a high impact on music sophistication collectively in the original Gold-MSI study [Müllensiefen *et al.*, 2014]. Furthermore, future research would benefit from including objective listening tasks, as participants may underestimate or overestimate their responses on the Gold-MSI survey battery, potentially skewing results. It is recommended for future work using the Gold-MSI to account for sociodemographic and objective listening task differences to better compare participant characteristics between studies. **While this approach follows previous works in the literature, we also note that converting a continuous measure into a categorical variable reduces information granularity and statistical power.**

**In addition, each task difficulty level used a different song, meaning task difficulty may be confounded with song selection. As Table 2 shows, the three experimental songs differed not only in note density (0.82, 1.80, and 2.83 notes per second) but also in tempo (132, 170, and 100 BPM), number of walls and mines, musical genre, and participant familiarity. Any of these characteristics could have influenced CL and task accuracy independently of the task difficulty level. We attempted to control fatigue, by selecting songs of near-identical duration (3:09, 3:09, and 3:06 minutes), but the remaining differences persist. Finally, the task difficulty ordering used (from easy to hard) introduces a clear learning and adaptation bias, which should be noted as it certainly influenced the results. This experimental design was chosen in an effort to minimize the difference between novices and experienced players, since the higher task difficulty levels in rhythm games are generally designed to be inaccessible for inexperienced people. Despite our efforts to minimize it, CL still presented an inverse correlation with previous Beat Saber experience (Figure 3-center), which indicates the learning effect was not high enough to negate the expected CL effect on novices.**

Thus, future work should consider song features beyond note count, such as angular variability, handedness alternation, or pattern complexity when assessing task difficulty. **In addition, other rhythm games could be included in future research to better support the generalisability of such findings, as other rhythm games may better capture other skills required to establish high levels of musical ability.** Although not assessed, additional insight could be derived from observing flow [Csikszentmihalyi, 1975], given that qualitative statements from participants infer different levels of challenge being appropriate for their current skillset. Furthermore, in observing perceived mental demand within a VR rhythm exergame task, there remains uncertainty in how much participants' performance influenced their reporting. Future research is encouraged to explore the relationship between knowledge of performance and subsequent perceived mental demand, in addition to examining the effects of each VGDS item and accounting for self-efficacy regarding task performance.

# 7 Conclusion

In summary, this study explored the assumption that musical training could explain the accumulation of CL and task accuracy in the VR rhythm exergame Beat Saber across different task difficulty levels. This investigation found that changes in task difficulty led to an increase in CL and a decrease in task accuracy. While MT did not prove to be a significant predictor of perceived mental demand in our study, it was a significant predictor of task accuracy. These findings suggest that although MT may not necessarily influence perceived mental effort, it can still contribute to more accurate gameplay, possibly due to enhanced visuospatial processing or motor coordination.

From a theoretical viewpoint, this study adds to CL literature by meeting assumptions of element interactivity through a simple yet demanding commercial VR rhythm exergame environment. Moreover, the findings support prior research suggesting that domain-specific experience, in our context digital games and Beat Saber, is linked to improved task accuracy and reduced perceived CL. Future research would benefit from measuring constructs such as flow to provide further insights, as qualitative responses pointed to their relevance in shaping user experience. Thus, the present study offers an initial basis for understanding how task difficulty, game experience, and individual differences contribute to variability in CL and task accuracy, suggesting that musical training effects may require more refined classification to fully capture their impact in VR rhythm exergames.

# Declarations

## Authors' Contributions

KE, EJP, MT, HF, and FZ contributed to the conception of this study. KE performed the experiments. KE, EJP, and RG are the main contributors and writers of this manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests with any of the devices, resources or intellectual property used in the study.

## Acknowledgements

The authors would like to thank the anonymous participants for their time. This research was partly funded by RMIT University.

## Availability of data and materials

Based on the ethics committee review and approval of the experimental protocol, participant data cannot be made publicly available due to its traceability.

# References

Anglin, J., Sugiyama, T., and Liew, S. (2017). Visuomotor adaptation in head-mounted virtual reality versus conventional training. *Scientific reports*, 7(1):45469. DOI: 10.1038/srep45469.

Arndt, S., Perkis, A., and Voigt-Antons, J.-N. (2018). Using virtual reality and head-mounted displays to increase performance in rowing workouts. In *Proceedings of the 1st International Workshop on Multimedia Content Analysis in Sports*, MMSports'18, page 45–50, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3265845.3265848.

Atkinson, R. and Shiffrin, R. (1968). Human memory: A proposed system and its control processes. volume 2 of *Psychology of Learning and Motivation*, pages 89–195. Academic Press. DOI: 10.1016/S0079-7421(08)60422-3.

Baddeley, A. D. and Hitch, G. (1974). Working memory. 8:47–89. DOI: 10.1016/S0079-7421(08)60452-1.

Baker, D. J., Ventura, J., Calamia, M., Shanahan, D., and Elliott, E. M. (2020). Examining musical sophistication: A replication and theoretical commentary on the Goldsmiths Musical Sophistication Index. *Musicae Scientiae*, 24(4):411–429. DOI: 10.1177/1029864918811879.

Benz, S., Sellaro, R., Hommel, B., and Colzato, L. S. (2016). Music makes the world go round: The impact of musical training on non-musical cognitive functions—A Review. *Frontiers in Psychology*, 6:2023–2023. DOI: 10.3389/FPSYG.2015.02023.

Berends, I. E. and van Lieshout, E. C. (2009). The effect of illustrations in arithmetic problem-solving: Effects of increased cognitive load. *Learning and Instruction*, 19(4):345–353. DOI: 10.1016/j.learninstruc.2008.06.012.

Bowman, N. D., Wasserman, J., and Banks, J. (2018). Development of the video game demand scale. In *Video Games: A Medium That Demands Our Attention*, pages 208–233. Routledge. DOI: 10.4324/9781351235266-13.

Bowmer, A., Mason, K., Knight, J., and Welch, G. (2018). Investigating the impact of a musical intervention on preschool children's executive function. *Frontiers in Psychology*, Volume 9 - 2018. DOI: 10.3389/fpsyg.2018.02389.

Campo-Prieto, P., Cancela, J. M., and Rodríguez-Fuentes, G. (2021). Immersive virtual reality as physical therapy in older adults: present or future (systematic review). *Virtual Reality*, 25(3):801–817. DOI: 10.1007/s10055-020-00495-x.

Chaddock-Heyman, L., Loui, P., Weng, T. B., Weisshappel, R., McAuley, E., and Kramer, A. F. (2021). Musical Training and Brain Volume in Older Adults. *Brain Sciences*, 11(1):50. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/brainsci11010050.

Chee, Z. J., Leung, Y., and de Vries, M. (2022). Musical sophistication and its relationship with executive functions, autistic traits, and quality of life. *Psychomusicology: Music, Mind and Brain*, 32(3-4):87–97. Num Pages: 87-97 Publisher: Educational Publishing Foundation (US). DOI: 10.1037/pmu0000294.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., and Conway, D. (2016). *Robust Multimodal Cognitive Load Measurement*. Springer Publishing Company, Incorporated, 1st edition. DOI: 10.5555/2978285.

Chi, M. T. H., Glaser, R., and Rees, E. (1982). Expertise in problem solving. In Sternberg, R. J., editor, *Advances in the psychology of human intelligence*, pages 7–75. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ. Book.

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety: the experience of play in work and games*. Jossey-Bass Inc. Publishers. Book.

Cutting, J., Deterding, S., Demediuk, S., and Sephton, N. (2023). Difficulty-skill balance does not affect engagement and enjoyment: a pre-registered study using artificial intelligence-controlled difficulty. *Royal Society Open Science*, 10(2):220274. DOI: 10.1098/rsos.220274.

Dahary, H., Palma Fernandes, T., and Quintin, E.-M. (2020). The relationship between musical training and musical empathizing and systemizing traits. *Musicae Scientiae*, 24(1):113–129. DOI: 10.1177/1029864918779636.

Elvitigala, D. S., Karahanoğlu, A., Matviienko, A., Turmo Vidal, L., Postma, D., Jones, M. D., Montoya, M. F., Harrison, D., Elbæk, L., Daiber, F., Burr, L. A., Patibanda, R., Buono, P., Hämäläinen, P., Van Delden, R., Bernhaupt, R., Ren, X., Van Rheden, V., Zambetta, F., Van Den Hoven, E., Lallemand, C., Reidsma, D., and Mueller, F. F. (2024). Grand challenges in sportshci. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3613904.3642050.

Frederiksen, J. G., Sørensen, S. M. D., Konge, L., Svendsen, M. B. S., Nobel-Jørgensen, M., Bjerrum, F., and Andersen, S. A. W. (2020). Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: a randomized trial. *Surgical endoscopy*, 34:1244–1252. DOI: 10.1007/s00464-019-06887-8.

Gradl, S., Eskofier, B. M., Eskofier, D., Mutschler, C., and Otto, S. (2016). Virtual and augmented reality in sports: an overview and acceptance study. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, page 885–888, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2968219.2968572.

Greenfield, P. M., Brannon, C., and Lohr, D. (1994). Two-dimensional representation of movement through three-dimensional space: The role of video game expertise. *Journal of Applied Developmental Psychology*, 15(1):87–103. DOI: 10.1016/0193-3973(94)90007-8.

Guarese, R., Pretty, E., Renata, A., Polson, D., and Zambetta, F. (2024). Exploring audio interfaces for vertical guidance in augmented reality via hand-based feedback. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2818–2828. DOI: 10.1109/TVCG.2024.3372040.

Guarese, R., van Schyndel, R., and Zambetta, F. (2022). Evaluating micro-guidance sonification methods in manual tasks for blind and visually impaired people. In *OzCHI '22: Proceedings of the 34th Australian Conference on Human-Computer Interaction*. DOI: 10.1145/3572921.3572929.

Haith, A. M. and Krakauer, J. W. (2018). The multiple effects of practice: skill, habit and reduced cognitive load. *Current Opinion in Behavioral Sciences*, 20:196–201. DOI: 10.1016/j.cobeha.2018.01.015.

Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Hancock, P. A. and Meshkati, N., editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland. DOI: 10.1016/S0166-4115(08)62386-9.

Hedlund, M., Jonsson, A., Bogdan, C., Meixner, G., Ekblom Bak, E., and Matviienko, A. (2023a). BlocklyVR: exploring block-based programming in virtual reality. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*, MUM '23, page 257–269, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3626705.3627779.

Hedlund, M., Lundström, A., Bogdan, C., and Matviienko, A. (2023b). Jogging-in-place: Exploring body-steering methods for jogging in virtual environments. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*, MUM '23, page 377–385, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3626705.3627778.

Huang, K.-T. (2020). Exergaming Executive Functions: An Immersive Virtual Reality-Based Cognitive Training for Adults Aged 50 and Older. *Cyberpsychology, Behavior, and Social Networking*, 23(3):143–149. DOI: 10.1089/cyber.2019.0269.

Huang, W., Roscoe, R. D., Johnson-Glenberg, M. C., and Craig, S. D. (2021). Motivation, engagement, and performance across multiple virtual reality sessions and levels of immersion. *Journal of Computer Assisted Learning*, 37(3):745–758. DOI: 10.1111/jcal.12520.

Kajastila, R., Holsti, L., and Hämäläinen, P. (2016). The augmented climbing wall: High-exertion proximity interaction on a wall-sized interactive surface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 758–769, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2858036.2858450.

Kalyuga, S. and Singh, A.-M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, 28(4):831–852. DOI: 10.1007/s10648-015-9352-0.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology*, 3(3):203–220. DOI: 10.1207/s15327108ijap0303_3.

Kirschner, D. and Williams, J. P. (2013). Experts and Novices or Expertise? Positioning Players through Gameplay Reviews. In *DiGRA Conference*, pages 1–18. DOI: 10.26503/dl.v2013i1.660.

LaViola, J. J. (2000). A discussion of cybersickness in virtual environments. *SIGCHI Bull.*, 32(1):47–56. DOI: 10.1145/333329.333344.

Lemmens, J. S. and von Münchhausen, C. F. (2023). Let the beat flow: How game difficulty in virtual reality affects flow. *Acta Psychologica*, 232:103812. DOI: 10.1016/j.actpsy.2022.103812.

Levitin, D. (2012). What does it mean to be musical? *Neuron*, 73(4):633–637. DOI: 10.1016/j.neuron.2012.01.017.

Mandryk, R. L., Inkpen, K. M., and Calvert, T. W.

(2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology*, 25(2):141–158. DOI: 10.1080/01449290500331156.

Matziorinis, A. M., Gaser, C., and Koelsch, S. (2023). Is musical engagement enough to keep the brain young? *Brain Structure and Function*, 228(2):577–588. DOI: 10.1007/s00429-022-02602-x.

Miura, Y., Wu, E., Kuribayashi, M., Koike, H., and Morishima, S. (2023). Exploration of sonification feedback for people with visual impairment to use ski simulator. In *Proceedings of the Augmented Humans International Conference 2023*, AHs '23, page 147–158, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3582700.3582702.

Moreno, R. and Mayer, R. (2007). Interactive Multimodal Learning Environments. *Educational Psychology Review*, 19(3):309–326. DOI: 10.1007/s10648-007-9047-2.

Mueller, F. and Isbister, K. (2014). Movement-based game guidelines. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2191–2200, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2556288.2557163.

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLOS ONE*, 9(2):e89642. DOI: 10.1371/journal.pone.0089642.

Nacke, L. E., Kalyn, M., Lough, C., and Mandryk, R. L. (2011). Biofeedback game design: using direct and indirect physiological control to enhance game interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 103–112, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/1978942.1978958.

Paas, F. G. and Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review*, 6(4):351–371. DOI: 10.1007/BF02213420.

Pretty, E. J., Fayek, H. M., and Zambetta, F. (2023a). A Case for Personalized Non-Player Character Companion Design. *International Journal of Human–Computer Interaction*, 0(0):1–20. DOI: 10.1080/10447318.2023.2181125.

Pretty, E. J., Guarese, R., Dziego, C. A., Fayek, H. M., and Zambetta, F. (2024a). Multimodal measurement of cognitive load in a video game context: A comparative study between subjective and objective metrics. *IEEE Transactions on Games*, 16(4):854–867. DOI: 10.1109/TG.2024.3406723.

Pretty, E. J., Guarese, R., Fayek, H. M., and Zambetta, F. (2023b). Replicability and transparency for the creation of public human user video game datasets. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 74–81. DOI: 10.1109/VRW58643.2023.00021.

Pretty, E. J., Guarese, R., Fayek, H. M., and Zambetta, F. (2024b). Comparing subjective measures of workload in video game play: Evaluating the test-retest reliability of the vgds and nasa-tlx. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, ICMI '24 Companion, page 56–60, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3686215.3690151.

Renata, A., Guarese, R., Takac, M., and Zambetta, F. (2024). Assessment of embodied visuospatial perspective taking in augmented reality: insights from a reaction time task. *Frontiers in Virtual Reality*, Volume 5 - 2024. DOI: 10.3389/frvir.2024.1422467.

Robinson, R., Wiley, K., Rezaeivahdati, A., Klarkowski, M., and Mandryk, R. L. (2020). "let's get physiological, physiological!": A systematic review of affective gaming. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '20, page 132–147, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3410404.3414227.

Rutkowski, S., Adamczyk, M., Pastuła, A., Gos, E., Luque-Moreno, C., and Rutkowska, A. (2021). Training using a commercial immersive virtual reality system on hand–eye coordination and reaction time in young musicians: A pilot study. *International journal of environmental research and public health*, 18(3):1297. DOI: 10.3390/ijerph18031297.

Sawan, N., Eltweri, A., De Lucia, C., Pio Leonardo Cavaliere, L., Faccia, A., and Roxana Moşteanu, N. (2021). Mixed and augmented reality applications in the sport industry. In *Proceedings of the 2020 2nd International Conference on E-Business and E-Commerce Engineering*, EBEE '20, page 55–59, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3446922.3446932.

Schmidt, S., Ehrenbrink, P., Weiss, B., Voigt-Antons, J.-N., Kojic, T., Johnston, A., and Möller, S. (2018). Impact of virtual environments on motivation and engagement during exergames. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. DOI: 10.1109/QoMEX.2018.8463389.

Schnotz, W. and Kürschner, C. (2007). A Reconsideration of Cognitive Load Theory. *Educational Psychology Review*, 19(4):469–508. DOI: 10.1007/s10648-007-9053-4.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., and Ehlert, U. (2010). Discriminating Stress From Cognitive Load Using a Wearable EDA Device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):410–417. DOI: 10.1109/TITB.2009.2036164.

Slater, M. and Sanchez-Vives, M. V. (2016). Enhancing Our Lives with Immersive Virtual Reality. *Frontiers in Robotics and AI*, 3. DOI: 10.3389/frobt.2016.00074.

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42(4):73–93. DOI: 10.1111/j.1460-2466.1992.tb00812.x.

Strong, J. V. and Mast, B. T. (2019). The cognitive functioning of older adult instrumental musicians and non-musicians. *Aging, Neuropsychology, and Cognition*, 26(3):367–386. DOI: 10.1080/13825585.2018.1448356.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4):295–312. DOI: 10.1016/0959-4752(94)90003-5.

Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review*, 22(2):123–138. DOI: 10.1007/s10648-

010-9128-5.

Sweller, J. (2011). CHAPTER TWO - Cognitive Load Theory. In Mestre, J. P. and Ross, B. H., editors, *Psychology of Learning and Motivation*, volume 55, pages 37–76. Academic Press. DOI: 10.1016/B978-0-12-387691-1.00002-8.

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory*. Springer New York, New York, NY. DOI: 10.1007/978-1-4419-8126-4.

Sweller, J., van Merrienboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3):251–296. DOI: 10.1023/A:1022193728205.

Takac, M., Collett, J., Conduit, R., and De Foe, A. (2023). A cognitive model for emotional regulation in virtual reality exposure. *Virtual Reality*, 27(1):159–172. DOI: 10.1007/s10055-021-00531-4.

Talamini, F., Altoè, G., Carretti, B., and Grassi, M. (2017). Musicians have better memory than nonmusicians: A meta-analysis. *PLOS ONE*, 12(10):e0186773. DOI: 10.1371/journal.pone.0186773.

Tammy Lin, J.-H., Wu, D.-Y., and Bowman, N. (2023). Beat Saber as Virtual Reality Exercising in 360 Degrees: A Moderated Mediation Model of VR Playable Angles on Physiological and Psychological Outcomes. *Media Psychology*, 26(4):414–435. DOI: 10.1080/15213269.2022.2154806.

Tian, M., Cai, Y., and Zhang, J. (2023). The Impact of Virtual Reality-Based Products on Mild Cognitive Impairment Senior Subjects: An Experimental Study Using Multiple Sources of Data. *Applied Sciences*, 13(4):2372. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/app13042372.

Toft-Nielsen, C. (2016). Gaming Expertise: Doing Gender and Maintaining Social Relationships in the Context of Gamers' Daily Lives. *Nordicom Review*, 37(s1):71–83. DOI: 10.1515/nor-2016-0024.

van Delden, R., Bergsma, S., Vogel, K., Postma, D., Klaassen, R., and Reidsma, D. (2020). VR4VRT: Virtual reality for virtual rowing training. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '20, page 388–392, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3383668.3419865.

Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., Depaepe, F., and Van den Noortgate, W. (2021). Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work*, 23:567–585. DOI: 10.1007/s10111-020-00641-0.

Wintersberger, P., Matviienko, A., Schweidler, A., and Michahelles, F. (2022). Development and evaluation of a motion-based VR bicycle simulator. *Proc. ACM Hum.-Comput. Interact.*, 6(MHCI). DOI: 10.1145/3546745.

Xu, W., Liang, H.-N., Zhang, Z., and Baghaei, N. (2020). Studying the Effect of Display Type and Viewing Perspective on User Experience in Virtual Reality Exergames. *Games for Health Journal*, 9(6):405–414. DOI: 10.1089/g4h.2019.0102.

Yannakakis, G. N., Martinez, H. P., and Garbarino, M. (2016). Psychophysiology in games. In Karpouzis, K. and Yannakakis, G. N., editors, *Emotion in Games: Theory and Praxis*, Socio-Affective Computing, pages 119–137. Springer International Publishing. DOI: 10.1007/978-3-319-41316-7_7.

Yerkes, R. M. and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5):459–482. DOI: 10.1002/cne.920180503.

Yeung, A. S., Jin, P., and Sweller, J. (1998). Cognitive load and learner expertise: Split-attention and redundancy effects in reading with explanatory notes. *Contemporary Educational Psychology*, 23(1):1–21. DOI: 10.1006/ceps.1997.0951.

Yoo, S., Gough, P., and Kay, J. (2020). Embedding a VR game studio in a sedentary workplace: Use, experience and exercise benefits. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3313831.3376371.