# Mental Workload Wars: Audio-based XR Awakens

Emma Jane Pretty
emma.pretty@tuni.fi
Gamification Group, Tampere University
Tampere, Finland

Renan Guarese
guarese@kth.se
Digital Futures, KTH Royal Institute of Technology
Stockholm, Sweden

## Abstract

Considering the need for valid and thorough metrics for mental workload in different modalities of extended reality (XR), we partially replicate a previous validation study under a different context: audio-based XR. Using two available datasets, we assess whether the cognitive subscale of the Video Game Demand Scale (VGDS) provides a valid measure of mental workload by comparing it with the mental demand item of the NASA Task Load Index in two distinct XR experiences: a rhythm-based virtual reality exergame, and an augmented reality sonification guidance task. Our results show a strong and moderately stable correlation between the two metrics under different difficulty conditions in each task, further validating the VGDS in multimodal XR contexts.

## Keywords

AR, VR, Cognitive Load, Workload, Audio Interfaces, Sonification

**ACM Reference Format:**
Emma Jane Pretty and Renan Guarese. 2025. Mental Workload Wars: Audio-based XR Awakens. In *Proceedings of ACM international joint conference on Pervasive and Ubiquitous Computing (UbiComp '25).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3714394.3756229

## 1 Introduction

The accurate measurement of cognitive load is critical for understanding user experience, accessibility, and the design of interactive digital systems [1]. This is especially important as digital technologies increasingly permeate everyday life, including education, healthcare, training, and entertainment. In such contexts, cognitive load measures can inform how systems support users' mental effort, reduce unnecessary strain, and promote effective interaction. As extended reality (XR) technologies continue to advance and become more widely adopted, the needs for these validated measurement has increased. Even by refining the scope into audio-based XR, its current applications are being used not only for entertainment but also for serious purposes such as rehabilitation, skill development, and immersive learning [4, 17]. These environments pose unique challenges for cognitive load measurement due to their high interactivity and multisensory nature.


*UbiComp '25, Espoo, Finland*

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/10.1145/3714394.3756229

Despite the growing importance of cognitive load measurement in XR, most studies continue to rely on post-task self-report instruments developed for general high-workload settings, as opposed to live physiological sensors [13]. The NASA Task Load Index (NASA-TLX) [10], is still the most commonly used tool for mental workload [1], even when focusing on audio-based XR [4]. Originally designed to assess workload in aviation and other general-purpose tasks, the NASA-TLX has been applied widely in interactive system research, including video games and XR applications [1]. However, its use in these contexts is not without limitations. First, the NASA-TLX was not specifically developed for interactive or immersive environments. Its items may not fully capture the nuances of mental workload in contexts where users engage with dynamic, multimodal stimuli and adaptive system responses. Second, as XR technologies introduce novel forms of sensory and cognitive engagement, there is growing concern about whether the NASA-TLX retains its validity in these settings [1]. Recent works have questioned its continued appropriateness for evaluating cognitive load in gaming and XR environments, noting that it may not distinguish well between different types of cognitive effort or account for the specific characteristics of these systems [1, 14].

A further conceptual concern arises from the distinction between workload and *cognitive* load, which are related but not interchangeable constructs. Workload refers to the total demand placed on an individual during task performance, such as cognitive, physical, temporal, and emotional components [15]. Cognitive load, by contrast, focuses specifically on the mental resources required for task-relevant information processing [13]. Although the NASA-TLX includes a mental demand subscale, many studies conflate these constructs by reporting overall NASA-TLX workload scores as measures of cognitive load [1]. This practice risks mischaracterising the cognitive demands of XR and gaming systems, where physical and sensory demands may inflate workload without necessarily increasing cognitive processing demands.

Newer instruments have been developed to better capture the overall workload and task demand of specific contexts, such as the Video Game Demand Scale (VGDS) [3]. The VGDS was designed to assess workload in video games and gamified experiences in a way that reflects the unique characteristics of interactive play. It includes subscales that target different types of demand, including cognitive, physical, and emotional components. While the VGDS was not developed to resolve the broader conceptual conflation of workload and cognitive load, it offers a more context-sensitive way to measure different facets of demand in gaming environments [15]. In the present work, we focus specifically on the VGDS cognitive subscale, which targets mental workload in video game play. The VGDS has shown promise in initial validation studies involving conventional screen-based video games [3], recently including a

few works addressing immersive platforms as well [2, 12]. However, its applicability to audio-based XR contexts remains largely unexplored. Critically, the VGDS has not yet been systematically validated for audio-based experiences [14], nor has its convergent (or other types of) validity been tested when administered immediately after gameplay in immersive environments [13], as opposed to recalling on players' most recent playing sessions [2, 3]. These represent significant gaps, as audio interfaces are increasingly being used in XR for purposes where visual channels are already overloaded or not available [4, 7, 17], and precise mental workload measurement should be essential.

In this work, we examine the relationship between VGDS cognitive subscale scores and NASA-TLX mental demand scores within two audio-based XR datasets [6, 7], testing the stability of this relationship across difficulty levels, and evaluating whether the VGDS provides consistent convergent validity in both contexts. By comparing the VGDS to the NASA-TLX, we aim to extend the empirical foundation for the VGDS and provide guidance for researchers seeking to measure mental workload in XR applications. Our contributions are as follows:

- We provide the first empirical comparison of VGDS and NASA-TLX scores in audio-based XR contexts across two distinct sound modalities: musical and sonification.
- We assess the convergent validity of the VGDS in situ, immediately after XR task completion.
- We examine the stability of the VGDS–NASA-TLX relationship across difficulty levels, contributing evidence for the use of the VGDS in XR, for mental workload measurement.

## 2 Related Works

Through a recent literature review of two decades of academic works, Babaei et al. [1] analysed the theoretical and methodological inconsistencies found in the use of the NASA-TLX. The authors point out severe drawbacks of the measure itself, suggesting the use of more suitable tools for better convergent validity within interaction tasks [1]. When exploring alternatives, Pretty et al. [13] found several physiological metrics related to cognitive load (including heart rate variability, tonic electrodermal activity, and electromyography amplitude) to be significant predictors of the cognitive subscale of the VGDS. In another study comparing these results against NASA-TLX responses [15], some degree of test-retest reliability was found between the two subjective tools, as well as strong correlations between their totals at two time points: immediately after gameplay, and a month later, in recollection. These findings suggest VGDS to be a strong contender to replace the NASA-TLX as the gold standard for measuring workload, at least within gamified scenarios.

Using different XR experiences as stimuli, a couple of works have measured and compared VGDS and NASA-TLX results. By surveying an XR-themed internet forum, Bowman et al. [2] asked players to recall on their most recent XR playing sessions, performing convergent and predictive validity analyses between VGDS and NASA-TLX metrics. Throughout the different results, authors found multiple significant correlations between both instruments, including total scores and analogous subscales [2]. Kryston et al. [12] performed a correlation analysis between the cognitive demand subscale of VGDS and the raw results for the NASA-TLX, both collected immediately after an immersive public speaking training session. The authors found a positive correlation between the two metrics [12].

Although SIM-TLX, an alternative subjective workload instrument which has been developed and validated specifically for XR experiences Harris et al. [9], the NASA-TLX remains as the most used metric in audio-based XR experiences [4]. Despite its established use, SIM-TLX falls out of the scope of our current analysis, which focuses on VGDS instead, considering its extensive validity on gamified experiences. On that note, Tammy Lin et al. [16] applied the VGDS to participants playing Beat Saber, while testing for the effect of different playable angle conditions on the psychological and physical activities of participants. However, the authors did not report on the collection of any alternative workload instruments, which hinders the use of their dataset for comparison purposes.

To address this gap, we present a study that compares datasets including the VGDS and NASA-TLX in two audio-based XR modalities: a VR musical exergame (Beat Saber) [5, 6] and an AR sonification guidance task, in a lack of visibility scenario [7, 8]. These two experiences represent distinct forms of non-speech audio interfaces within XR: musical and sonification [11], allowing us to assess the generalisability of the VGDS across these modalities.

Thus, we ask the following research questions:

- How does the VGDS relate to NASA-TLX mental demand scores across two different audio-based XR modalities?
- Is the strength of this relationship stable across task difficulty levels within each modality?

## 3 Methodology

We partially replicate the methodology of a study conducted by Pretty et al. [15], applying their data analysis methods onto two mental workload datasets, each collected on distinct sound-based immersive experiences, as detailed below:

- Dataset 1: **Beat Saber** task, by Ellahiyoun et al. [5, 6]
  - Task: playing specific songs on the commercially available exergame Beat Saber[1]
  - Sample size: 27 (aged 18 to 35; $M = 23.63, SD = 4.58$)
  - Conditions:
    - **Easy**: the *easy* difficulty level of Beat Saber, lower number of notes to strike, slow-paced.
    - **Hard**: the *hard* difficulty level of Beat Saber, higher number of notes to strike, fast-paced.
- Dataset 2: **Sonification** task, by Guarese et al. [7]
  - Task: locating targets in a large cupboard, being guided by audio XR feedback while blindfolded.
  - Sample size: 18 (aged 22 to 50; $M = 30.84, SD = 6.92$)
  - Conditions:
    - **Easy**: pitch changes according to user's hand-to-target distance, with the bottom of the pitch scale placed at the target position, originally named *Target-based Dynamic pitch* [7], considered *easy* here for its best performance and usability [8] results in the original experiment.
    - **Hard**: pitch changes according to user's hand-to-target distance, with the center of the pitch scale placed at

[1] https://beatsaber.com/

the target position, originally named *Target-centered Dynamic pitch* [7], considered *hard* here for its worst performance and usability [8] results in the original experiment.

- Metrics: both subjective quantitative measures for mental load were applied via questionnaires directly after each condition was experienced by the respective participants of each experiment.
  - NASA TLX [10]: a subjective metric for workload experience, measured under six subscales: mental demand. physical demand, temporal demand, performance, effort, and frustration.
  - VGDS [3]: a subjective metric for demand in videogames, measured under four different factors, each with multiple subscales: cognitive, emotional, physical, and social.

# 4 Results

## 4.1 Task Difficulty Comparison

To confirm that the task conditions represented distinct levels of difficulty, we compared the averages for the NASA-TLX and the VGDS cognitive composite scores across the two levels for each XR experience. Normality was assessed using Shapiro-Wilk tests. For Beat Saber, VGDS composite scores did not significantly deviate from normality (easy: $W = 0.96$, $p = .32$; hard: $W = 0.94$, $p = .055$). However, NASA-TLX scores showed evidence of non-normality (easy: $W = 0.86$, $p < .001$; hard: $W = 0.91$, $p = .01$). For sonification, VGDS composite scores were consistent with normality (easy: $W = 0.93$, $p = .25$; hard: $W = 0.99$, $p = .99$), while NASA-TLX scores were marginally not normal (easy: $W = 0.89$, $p = .04$) or normal (hard: $W = 0.92$, $p = .11$).

Based on these results, a combination of parametric and non-parametric tests was applied. For Beat Saber, significantly higher NASA-TLX mental demand scores for the hard level ($M = 4.91, SD = 1.69$) compared to the easy level ($M = 2.31, SD = 1.26$) were found through a Wilcoxon rank-sum test ($W = 125$, $p < .001$). VGDS cognitive composite scores were significantly higher in the hard condition ($M = 4.66, SD = 1.94$) than the easy condition ($M = 3.03, SD = 1.79$), as confirmed by a Welch Two Sample $t$-test ($t(57) = -6.25$, $p < .001$, 95% $CI = [-2.15, -1.11]$).

For sonification, no significant differences were observed between easy ($M = 3.78, SD = 2.07$) and hard ($M = 4.89, SD = 1.64$) conditions for NASA-TLX mental demand scores (Wilcoxon rank-sum test: $W = 110$, $p = .10$) or VGDS cognitive scores (Welch Two Sample t-test: $t(31) = -0.82$, $p = .42$, 95% $CI = [-0.98, 0.42]$), also comparing its easy ($M = 3.80, SD = 1.79$) and hard ($M = 4.08, SD = 1.88$) results. Considering the lack of significant difference in these comparisons–likely due to the small sample size–, we proceed by relying only on the performance [7] and usability [8] metrics from the original study, which were significantly better for the easy condition, in comparison to its hard counterpart.

## 4.2 Convergent Validity Analyses

We combined correlational analyses and linear modelling to evaluate the convergent validity of the VGDS cognitive subscale. Correlations assessed the strength of association between VGDS and NASA-TLX mental demand scores within each task and difficulty level. Linear models allowed us to test whether VGDS scores predicted NASA-TLX scores across conditions and whether this relationship was influenced by task difficulty. Fisher's $z$ tests were used to determine the sensitivity of the measures, by identifying whether the strength of these correlations across difficulties were significantly different.

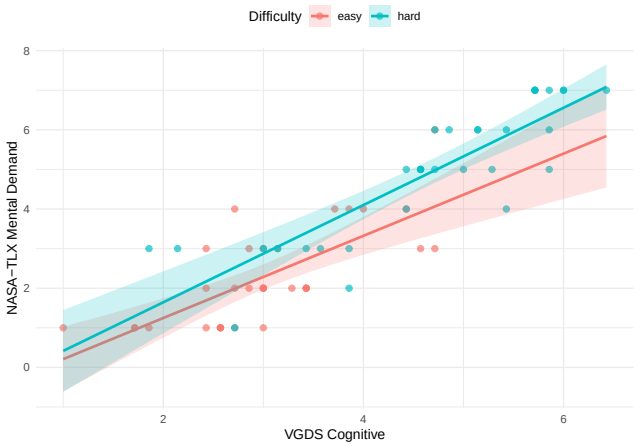


**Figure 1: Correlation between NASA-TLX and VGDS Cognitive Scores on Easy and Hard tasks in Beat Saber**

In Beat Saber, Spearman's correlations indicated a strong positive association between VGDS composite and NASA-TLX mental demand in both easy ($\rho = 0.72$, $p < .001$) and hard ($\rho = 0.87$, $p < .001$) conditions. Linear modelling revealed a significant main effect of VGDS on NASA-TLX ($\beta = 1.04$, $SE = 0.18$, $t(60) = 5.76$, $p < .001$), but no significant VGDS × Difficulty interaction ($\beta = 0.20$, $SE = 0.22$, $t(60) = 0.85$, $p = .40$). Fisher's $z$ test comparing correlation strength across difficulty levels found no significant difference ($z = -1.48$, $p = .14$).

In the sonification task, Spearman's correlation between VGDS and NASA-TLX was moderate and significant in the easy condition ($\rho = 0.53$, $p = .035$), and Pearson's correlation was strong and significant in the hard condition ($r = 0.77$, $p < .001$). Linear modelling indicated a positive and significant main effect of VGDS on NASA-TLX ($\beta = 1.19$, $SE = 0.42$, $t(29) = 2.84$, $p = .01$) and no significant VGDS × Difficulty interaction ($\beta = 0.05$, $SE = 0.56$, $t(29) = 0.09$, $p = .93$). Fisher's $z$ test indicated no significant difference in correlation strength between conditions ($z = 1.10$, $p = .27$).

# 5 Discussion

Our findings suggest that the VGDS cognitive subscale has promise as a practical measure of mental workload immediately after task completion in audio-based XR tasks that combine visual, physical, and cognitive engagement, as seen in the rhythm-based XR exergame. The strong and consistent relationship between VGDS and NASA-TLX mental demand scores in this setting indicates that users' self-reports on these two scales are closely aligned when interacting with a fast-paced, visually rich, and physically active XR environment. This supports the idea that the VGDS can capture relevant aspects of mental demand in contexts that share features

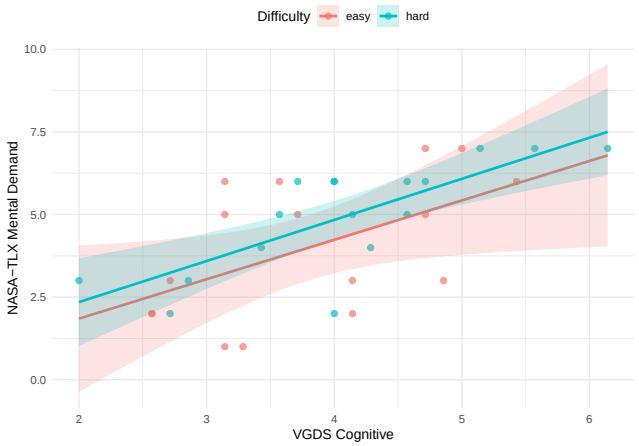


**Figure 2: Correlation between NASA-TLX and VGDS Cognitive Scores on Easy and Hard methods in the Sonification Task**

with conventional gaming, which, outside of XR, has been replicated before [3, 15], even with objective metrics related to cognitive load [13].

In contrast, although existent, a slightly more variable association between VGDS and NASA-TLX scores was evident in the sonification guidance task, which highlights important considerations for this measurement in XR. The VGDS may not yet fully account for how users experience mental workload in XR environments that rely predominately on auditory interaction or that lack the multimodal feedback typical of games, considering the experiment for this dataset was conducted under a no-visibility condition [7]. Alternatively, it may reflect the challenges participants face in reflecting on and articulating their cognitive effort in less familiar or less visually engaging XR experiences. Considering its smaller sample size, this finding invites further reflection and research on whether different types of XR tasks require tailored measurement tools, such as the SIM-TLX [9], or at least modified versions of existing tools that are sensitive to sensory modality and interaction style.

Importantly, our analyses showed that the strength of the relationship between the VGDS and NASA-TLX was stable across difficulty levels in both audio-based XR modalities. This suggests that the VGDS cognitive subscale can provide a consistent picture of mental workload regardless of how challenging a task is perceived to be. This is encouraging for researchers and technology designers seeking to use the VGDS in studies or evaluations where task difficulty varies, as it indicates that the tool has high levels of sensitivity under varying levels of challenge. Considering the broader XR scope, our work adds audio-based and AR experiences to the corpus of works that have been demonstrating the usefulness [16] and validity [2, 12] of the VGDS as an established mental workload metric.

However, we note limitations of the current work. The sample sizes for both XR datasets are relatively small, limiting our ability to generalise these findings. We also focused on only two audio-based XR tasks, which do not represent the full diversity of audio XR applications now in use across domains such as education, health, and industry [4, 17]. Notably, we also only analysed examples for two types of non-speech auditory interfaces: musical and sonification [11], without a speech interface being evaluated thus far, leaving another gap for future studies.

Future work should aim to explore the VGDS in a broader range of XR tasks, including those that differ in their sensory complexity, level of immersion, and interactivity. Studies combining the VGDS with objective indicators of cognitive load, such as physiological measures (e.g., heart rate variability, pupillometry) [13] or behavioural markers (e.g., error rates, response times) [7], can help build a stronger case for its validity.

## 6 Conclusion

In this study, we used available datasets to show that the cognitive subscale of the VGDS provides a valid measure of mental workload in audio-based XR contexts, by comparing it with NASA-TLX mental demand scores in two distinct XR experiences: a rhythm-based VR exergame and an AR sonification guidance task. Our aim was to evaluate whether a tool developed for gaming contexts could be applied effectively to immersive environments that differ in sensory modality (musical and sonification) and interaction style (VR and AR), providing further reflection is this relationship.

In light of our results, we believe researchers should consider whether new or adapted scales are needed to capture mental demand in XR experiences that fall outside traditional gaming paradigms, such as audio-based or exertional experiences. As XR systems continue to evolve and become more integrated into daily life, ensuring that we have accurate, context-sensitive tools for measuring mental workload will be vital for supporting user experience, learning, and performance.

## Acknowledgments

The authors wish to thank Tampere University, KTH, and Digital Futures (Grant Number KTH-RPROJ-0146472) for their support.

*VGDS was the chosen one! It was said that it would destroy the TLX, not join them! Bring balance to Mental Workload, not leave it in darkness!*

## References

[1] Ebrahim Babaei, Tilman Dingler, Benjamin Tag, and Eduardo Velloso. 2025. Should we use the NASA-TLX in HCI? A review of theoretical and methodological issues around Mental Workload Measurement. *International Journal of Human-Computer Studies* 201 (2025), 103515. doi:10.1016/j.ijhcs.2025.103515

[2] Nicholas David Bowman, Jih-Hsuan Tammy Lin, and Kevin Koban. 2021. Demanding on many dimensions: validating the interactivity-as-demand measurement model for VR-based video games. (2021).

[3] Nicholas David Bowman, Joseph Wasserman, and Jaime Banks. 2018. Development of the video game demand scale. In *Video games*. Routledge, 208–233.

[4] Isak de Villiers Bosman, Oğuz 'Oz' Buruk, Kristine Jørgensen, and Juho Hamari and. 2024. The effect of audio on the experience in virtual reality: a scoping review. *Behaviour & Information Technology* 43, 1 (2024), 165–199. doi:10.1080/0144929X.2022.2158371

[5] Kyla Ellahiyoun. 2023. *Effects of task difficulty and music expertise in virtual reality: an observation of cognitive load and task performance in Beat Saber.* Honours thesis. Royal Melbourne Institute of Technology (RMIT University), Melbourne Australia. doi:10.13140/RG.2.2.11581.78564

[6] Kyla Ellahiyoun, Emma Pretty, Renan Guarese, Marcel Takac, Haytham Fayek, and Fabio Zambetta. 2025. Effects of task difficulty and music expertise in VR: Observations of cognitive load and task accuracy in a rhythm exergame. In *Pre-print submitted for VRST'25.* doi:10.48550/arXiv.2507.06691

[7] Renan Guarese, Emma Pretty, Aidan Renata, Deb Polson, and Fabio Zambetta. 2024. Exploring audio interfaces for vertical guidance in augmented reality via hand-based feedback. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[8] Renan Guarese, Emma Pretty, and Fabio Zambetta. 2023. XR towards tele-guidance: mixing realities in assistive technologies for blind and visually impaired people. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 324–329. doi:10.1109/VRW58643.2023.00074

[9] David Harris, Mark Wilson, and Samuel Vine. 2020. Development and validation of a simulation workload measure: the simulation task load index (SIM-TLX). *Virtual Reality* 24, 4 (2020), 557–566.

[10] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[11] Thomas Hermann, Andy Hunt, John G Neuhoff, et al. 2011. *The sonification handbook*. Vol. 1. Logos Verlag Berlin.

[12] Kevin Kryston, Henry Goble, and Allison Eden. 2021. Incorporating virtual reality training in an introductory public speaking course. *Journal of Communication Pedagogy* 4 (2021), 133–151.

[13] Emma J. Pretty, Renan Guarese, Chloe A. Dziego, Haytham M. Fayek, and Fabio Zambetta. 2024. Multimodal Measurement of Cognitive Load in a Video Game Context: A Comparative Study Between Subjective and Objective Metrics. *IEEE Transactions on Games* 16, 4 (2024), 854–867. doi:10.1109/TG.2024.3406723

[14] Emma J. Pretty, Renan Guarese, Haytham M. Fayek, and Fabio Zambetta. 2023. 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)*. 74–81. doi:10.1109/VRW58643.2023.00021

[15] Emma Jane Pretty, Renan Luigi Martins Guarese, Haytham M Fayek, and Fabio Zambetta. 2024. 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*. 56–60.

[16] Jih-Hsuan Tammy Lin, Dai-Yun Wu, and Nicholas Bowman. 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 (2023), 414–435.

[17] Jing Yang, Amit Barde, and Mark Billinghurst. 2022. Audio augmented reality: A systematic review of technologies, applications, and future research directions. *journal of the audio engineering society* 70, 10 (2022), 788–809.

Received 10 July 2025; revised August 2025; accepted August 2025