Enhanced cognitive workload evaluation in 3D immersive environments with TOPSIS model — Source link ↗

Yuyang Wang, Jean-Rémy Chardonnet, Frederic Merienne

Related papers:

- Using Structural Similarity to Predict Future Workload Behavior in the Cloud

- Self directed learning based workload forecasting model for cloud resource management

- Development of the step complexity measure for emergency operating procedures using entropy concepts

- On the accuracy of objective image and video quality models: New methodology for performance evaluation

- A Rank-Based Method for the Usability Comparison of Competing Products

Enhanced cognitive workload evaluation in 3D immersive environments with TOPSIS model

Yuyang Wang, Jean-Rémy Chardonnet, Frédéric Merienne

Enhanced cognitive workload evaluation in 3D immersive environments with TOPSIS model[☆]

Yuyang Wang [1,*], Jean-Rémy Chardonnet [2], Frédéric Merienne [3]

Arts et Métiers Institute of Technology, LISPEN, HESAM Université, 2 Rue Thomas Dumorey, 71100 Chalon-sur-Saône, France

A R T I C L E   I N F O

A B S T R A C T

Research puts forward perception-based cognitive workload evaluation methods to help VR developers and users measuring their workload when playing with a VR application. Approaches to measure workload based on biosensors have progressed significantly, while evaluation based on subjective methods still rely on standard questionnaires such as the NASA-TLX table, the Subjective Workload Assessment Technique and the Modified Cooper Harper scale. The pre-defined questions enable operators to carry out experiments and analyse the data more easily than with biofeedback. However, the subjective evaluation process can bias the results because of unperceived internal changes and unknown factors among users. It is therefore necessary to have a method to handle and analyse this uncertainty. We propose to use the Technique for Order Performance by Similarity to Ideal Solution (TOPSIS) model to analyse the NASA-TLX table for measuring the overall user workload instead of using the classical weighted sum method. To show the advantage of the TOPSIS approach, we performed a user experiment to validate the approach and its application to VR, considering factors including the VR platform and the scenario density. Three different weighting methods, including the fuzzy Analytic Hierarchy Process (AHP) from fuzzy logic, the classical weighting based on pairwise comparison and the uniform weighting method, were tested to see the applicability of the TOPSIS model. The results from TOPSIS were consistent with those from other evaluation methods; a significant reduction in the coefficient of variation (CV) was observed when using the TOPSIS model to analyse the NASA-TLX scores, indicating an enhanced precision of the workload evaluation by the TOPSIS method. Our work has a potential application for VR designers and experimenters to compare cognitive workload among conditions and to optimize the settings.

1. Introduction

Virtual reality (VR) has become popular due to the fast development of affordable head-mounted devices (HMDs) with various available applications. Current VR HMDs mainly include high spatial and temporal resolution with dual displays (for example 1440×1600 pixels per eye and 90Hz to 120Hz refresh frame rate frequencies), achieving high-fidelity stereoscopic vision and image rendering, and low latency body tracking experience. Some international companies such as Google, Facebook, Microsoft, HTC, Samsung and Apple are participating in the development of new VR hardware with their own advantages (Chang and Chen, 2017), bringing a total market size estimated to more than $100 billion in the next decades (Bellini et al., 2016; Framingham, 2016; Merel, 2017).

Despite this growth and the capabilities of the latest devices, if a VR-based task is not designed with an appropriate level of cognitive workload to match a user's expertise, the task completion performance may be restrained (Zhang et al., 2017). Cognitive workload in this case is a term that refers to the cost of completing a task (Hart, 2006). It can be defined as the amount of cognitive resources used per unit time to reach the performance required by the task (Wickens et al., 2015). For pragmatic purposes, Blackwood (1900) proposes a simple definition that the workload is the ratio of time required to time available (time required/time available). When the time required to complete a task is longer than

the time available, it is cognitive overload, and vice versa. For example, when a user navigates in a virtual environment as a primary task, he/she may have to interact with internal objects as a secondary task, and because of this, the non-skilled user would keep overloading for longer time. More introduction to cognitive workload can also be found in Paas et al. (2003).

Although workload is defined at a qualitative level, researchers are still trying to find measurable criteria on a multi-dimensional basis to quantify this phenomenon. The approaches used in the literature to measure cognitive workload include subjective measurements via questionnaires, physiological measurements via biosensors and performance measures (Paas et al., 2003; Zhang et al., 2017). Physiological means for measuring cognitive workload have progressed due to the development of new types of biosensors, while subjective measurement methods are still based on long-living existing questionnaires such as the Subjective Workload Assessment Technique (SWAT) (Reid and Nygren, 1988), the Modified Cooper Harper scale (MCH) (Kilmer et al., 1988), or the NASA-TLX table (Hart, 2006; Hart and Staveland, 1988).

Subjective measurements require users to provide feedback of their experience through questionnaires. Among well-known questionnaires, the MCH method was developed by Harper and Cooper (1986), which is still being used nowadays as a reliable measurement of aircraft performance. The user evaluates the performance of a specific task in terms of controllability, workload, and attainable performance goals, on a 1-10 scale; then, the assessment is analysed through a statistical study. On the other hand, the SWAT method consists of three criteria to measure workload: time load, effort load and psychological stress load (Reid and Nygren, 1988). The user is asked to sort these factors, then to rate each of them in a 3-scale table, and the final evaluation is derived through a conversion table that provides the relationship between the 1-3 scale table and the final score. NASA-TLX (Hart, 2006; Hart and Staveland, 1988) is another widely accepted subjective workload evaluation method applied in the computer-human interface field (Cannavò et al., 2020; Ma and Kaber, 2006). Based on a multi-dimensional rating procedure, NASA-TLX obtains an overall workload score according to a weighted average of ratings on six criteria: *Mental Demand* (MD), *Physical demand* (PD), *Temporal Demand* (TD), *Performance* (Pe), *Effort* (Ef) and *Frustration* (Fr). Despite the simplicity of implementing these different methods, subjective evaluation cannot generally be performed in real time (while performing the tasks), which can lead to biased results because of unaware internal changes, e.g., the psychosocial environment (Casner and Gore, 2010), thus resulting in high uncertainty.

Many researchers have investigated the possibility of measuring cognitive workload through the physiological response of users, which does not require a direct response from the user. Contrary to subjective measures, results from physiological evaluation are representative of the actual task workload (Miller, 2001). Some of the frequently applied physiological methods include cardiac activity, eye gaze and electro-encephalography (EEG) (Gerry et al., 2018; Zhang et al., 2017). Cardiac activity is the most common approach for measuring workload in driving and flight simulators and is measured through blood pressure, heart rate and heart rate variability (HRV) (Hoover et al., 2012). Heartbeats can be analysed in the time and frequency domains to determine the HRV. Some early studies found that heart rate and HRV correlate with workload: the higher the mental workload, the higher the heart rate and the lower the HRV (Metalis, 1991; Mulder, 1986). Pupil dilation has recently been used as an indicator for mental workload in the HCI community since eye gaze signals indicate a user's cognitive state (Kosch et al., 2018). EEG signals are hypersensitive and credible for regular memory load evaluation, especially the alpha and theta wave-bands of EEG that are reflective of the task difficulty (Gevins et al., 1998). Physiological measurements alleviate the complexity of the experiment in that users do not have to conduct a second test or to be asked for feedback. However, these measures are subject to several sources of error: the accuracy of physiological data heavily relies on the performance and precision of the sensors; such signals are sensitive to disturbances.

Performance-based techniques are the third type of workload measurements. Paas et al. (2003) reports performance variables (e.g., reaction time, accuracy, and error rate) from a secondary task could reflect the cognitive workload imposed by a primary task. These criteria have been found to be correlated with cognitive workload (Son and Park, 2011). However, the insensitivity of some of the measurements while using performance-based methods is one drawback (Shakouri et al., 2018). For instance, a task with a low demanding workload can enable an excellent performance during the beginning stage of the task but then performance will degrade as the user becomes fatigued or distracted (Casner and Gore, 2010), leading to mixed and noisy results. Accordingly, using this method along with other workload evaluation methods may improve the quality of the measure.

Because the complexity of a task in a virtual environment can affect information processing, performance and user attention (Ma and Kaber, 2006), a virtual environment can act on users at both the cognitive and perceptual levels (Milleville-Pennel and Charron, 2015). To design and gain a better experience of VR technologies, advanced evaluation methods have to be developed to qualify and compare the cognitive workload induced by different VR applications. Because of the multi-faceted and multi-dimensional nature of cognitive workload, it is hard to define quantitative criteria that heavily rely on the competences and efforts of users in a specific application (Longo, 2014). Since subjective evaluation can be easily conducted and interpreted on one-dimensional or multi-dimensional scales, it is the most preferred approach (Eraslan et al., 2016). However, this evaluation method generally fails to control dispersion effects, noise and uncertainty during subjective investigation (Katicic et al., 2015), posing a threat to what we refer to the measuring precision. Considering that most current methods often fail to take into account inherent uncertainties during subjective judgment and comparison processes (Zhou and Chan, 2017), we propose to use the *Technique for Order Performance by Similarity to Ideal Solution* (TOPSIS) (Yoon and Hwang, 1995) model to enhance the precision of the subjective evaluation of workload. TOPSIS was developed for decision-makers to measure and compare the relative performance among different alternatives; it has been widely used in many evaluation processes such as supply chain management, manufacturing system design, business and marketing management and health management (Behzadian et al., 2012). The TOPSIS method is adopted in this work as a novel mean to compute an overall cognitive workload in VR applications. We supposed that the weighting methods will affect the weighted sum approach or from the TOPSIS. In this case, we introduced the fuzzy analytic hierarchy process (AHP) as an alternative to the existing weighting methods, which was detailed in subsection 2.2. The TOPSIS method was applied on top of three different weighting techniques to get a workload score from the NASA-TLX table.

The purpose of this study is not to develop a new approach to deduce mathematical operations and equations for measuring the cognitive workload, but rather, to improve the precision and quality of current evaluation approaches with existing models. The improvement of precision would enable VR designers and experimenters to better discriminate differences among settings and to optimize their applications easily. The VR domain requires many evaluation methods to obtain user's feedback regarding the interaction design, but the application of appropriate methods ensuring quality of the feedback in this field is rather scarce, which motivated us to involve TOPSIS and fuzzy AHP to improve the measuring quality of cognitive workload in VR. It is worth noting that our proposed method can be applied not only to VR, but also to any human-computer interaction application for which cognitive load is important to consider.

The proposed model was validated by a user study in section 4: two factors known to influence cognitive workload during a navigation task in a virtual environment are supposed to demonstrate the effectiveness of the TOPSIS method. These factors include the VR platform (HMD and CAVE) and the scenario density. For each of them, the overall workload

score was computed by the classical weighted sum method (Hart, 2006) and our TOPSIS model. The overall workload measured from TOPSIS was firstly validated by ensuring consistency of the results with the literature. In addition, compared to the weighted sum method, the TOPSIS model significantly reduced the data dispersion in terms of the coefficient of variation (CV), showing more accurate workload scores.

2. Methods for computing the cognitive workload scores

Throughout this paper, we will focus on the NASA-TLX table to demonstrate our methodology, since the criteria are well defined and widely accepted for the analysis of workload in virtual reality.

2.1. Weighted sum method

The calculation of the overall workload from the NASA-TLX table consists of two steps. First, participants have to perform a pairwise comparison of the criteria provided in the NASA-TLX table, based on the task they conducted and experienced; for example, if the user thinks that *Mental demand* is more important than *Physical demand*, then the weight of *Mental demand* is incremented by one while that of *Physical demand* remains unchanged. After a total of 15 comparisons, the weighting coefficients for the six criteria are obtained by normalisation. These comparisons are used to determine the weighting coefficient for each criterion. In the rest of the paper, this weighting coefficient calculation method will be named *Hart*. Second, participants are given another questionnaire to quantify the score for each criterion. The overall workload score is computed as a sum of the scores for each criterion weighted by their respective weighting coefficient (Eq. 1) (Hart, 2006; Hart and Staveland, 1988). Subjective workload evaluation can be completed directly by the participants without any requirement of sophisticated devices, which is nearly straightforward and inexpensive to perform.

$$TLX_{Hart} = W_{MD} * MD + W_{PD} * PD + W_{TD} * TD \\ + W_{Pe} * Pe + W_{Ef} Ef + W_{Fr} * Fr \quad (1)$$

TLX_{Hart} is the overall cognitive workload measured by the weighted sum method, W with a subscript is the corresponding weighting coefficient for each criterion.

Past research in many areas reports another approach to process the NASA-TLX table and obtain the cognitive workload score in a straightforward manner: a uniform weight is applied to each criterion (Kamaraj et al., 2016; Tubbs-Cooley et al., 2018), which means W_{MD}, W_{PD}, W_{TD}, W_{Pe}, W_{Ef} and W_{Fr} are equal to 16.67%. This approach is a particular case of the *Hart* weighting method. However, this approach is not adapted to different situations where the importance of each criterion needs to be differentiated.

2.2. Fuzzy Analytic Hierarchy Process

In the *Hart* weighting approach, participants have to compare the criteria in pairs to determine which one is more important than the other one. However in practice, the process of comparison and decision making is associated with the strong vagueness of human thinking: decision-makers generally give some or all pairwise comparison values with an uncertainty degree instead of precise ratings, and such uncertainty degree is represented from an appropriate semantic scale (Singh et al., 2013; Yu, 2002). As they usually are unable to explicit about their preference due to the fuzzy nature of the comparison process, decision-makers (VR users in the current context) usually tend to give interval judgments with semantic scale than fixed value (Gumus, 2009).

To cope with it, mathematicians proposed the fuzzy method in real practice in which an uncertain pair-to-pair comparison exists. Mouzé-Amady et al. (2013) propose a fuzzy integral approach based on the Sugeno integrals to determine the weighting coefficients. The weighting coefficients are determined with at least one global measure (either a subjective rating or an objective cue, e.g., HRV, reaction time) to serve as an aggregation criterion. Then data-driven models (e.g., minimum specificity principle, simulated annealing technique) are used to find the optimal weighting coefficients and to fit the global criterion. In this approach, the weighting coefficients are no longer set based on the participant's subjective evaluation. To focus on evaluation with only subjective ratings, we will rather use the fuzzy AHP proposed by Chang (1996) to compute the weighting coefficients, as an alternative to the *Hart* weighting approach. Here we briefly describe the basic concepts behind the fuzzy AHP. More theoretical details can be found in the literature (Chang, 1996).

2.2.1. Definition of triangular fuzzy numbers

Each fuzzy number can be represented by a membership function. A fuzzy number is called a triangular fuzzy number (TFN) when it can be described by the following function $\mu_M(x)$: $\mathbb{R} \rightarrow [0, 1]$,

$$\mu_M(x) = \begin{cases} \dfrac{x-l}{m-l} & x \in [l, m] \\ \dfrac{x-u}{m-u} & x \in [m, u] \\ 0 & otherwise \end{cases} \quad (2)$$

where $l \leq m \leq u$, l, m, u are the lower, modal and upper values of the TFN. This function is called the membership function.

The TFN has some pre-defined arithmetic operators. If $M_1 = (l_1, m_1, u_1)$ and $M_2 = (l_2, m_2, u_2)$ are two TFNs, the operation rules between the two TFNs are given as,

$$\begin{cases} M_1 \oplus M_2 = (l_1 + l_2, m_1 + m_2, u_1 + u_2) & (3a) \\ M_1 \otimes M_2 = (l_1 l_2, m_1 m_2, u_1 u_2) & (3b) \\ \lambda M_1 = (\lambda l_1, \lambda m_1, \lambda u_1) & (3c) \\ M_1^{-1} = (l_1, m_1, u_1)^{-1} \approx \left(\dfrac{1}{u_1}, \dfrac{1}{m_1}, \dfrac{1}{l_1} \right) & (3d) \end{cases}$$

where $\oplus$ denotes the extended summation of two TFNs and $\otimes$ denotes the extended multiplication.

2.2.2. Formulation of a fuzzy synthetic extent analysis

Assuming a triangular fuzzy comparison matrix $\tilde{A} = (a_{ij})_{n \times n}$, the extent analysis first sums up each row of this matrix, then normalizes the row sums with respect to the ith row,

$$S_i = \sum_{j=1}^{n} a_{ij} \otimes \left[\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \right]^{-1} \quad (4)$$

where $a_{ij} = (l_{ij}, m_{ij}, u_{ij})$ is a triangular fuzzy number. According to the operations rules, S_i is also a triangular fuzzy number.

Given two triangular fuzzy numbers, $S_1 = (l_1, m_1, u_1)$ and $S_2 = (l_2, m_2, u_2)$, the degree of possibility that $S_2 \geq S_1$ is defined as

$$V(S_2 \geq S_1) = \begin{cases} 1 & m_2 \geq m_1 \\ 0 & l_1 \geq u_2 \\ \dfrac{l_1 - u_2}{(m_2 - u_2) - (m_1 - l_1)} & otherwise \end{cases} \quad (5)$$

To compare S_1 and S_2, both values of $V(S_1 \geq S_2)$ and $V(S_2 \geq S_1)$ must be computed. Further, the degree of possibility for a convex fuzzy number S to be larger than k convex fuzzy numbers S_i, $i = 1, 2, ...k$ can be defined by

$$\begin{aligned} & V(S \geq S_1, S_2, ..., S_k) \\ & = V[(S \geq S_1) \text{ and } (S \geq S_2) \text{ and} \\ & ... \text{ and } (S \geq S_k)] = \min V(S \geq S_i) \end{aligned} \quad (6)$$
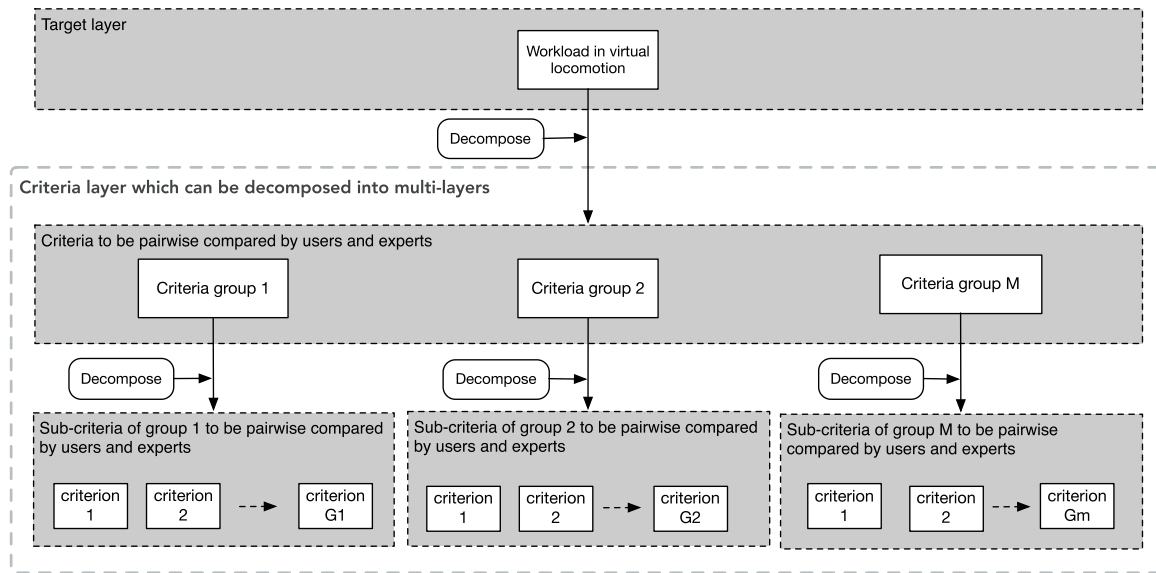
Fig. 1. General hierarchical structure of workload evaluation; this structure can include related indices from different groups or sources.
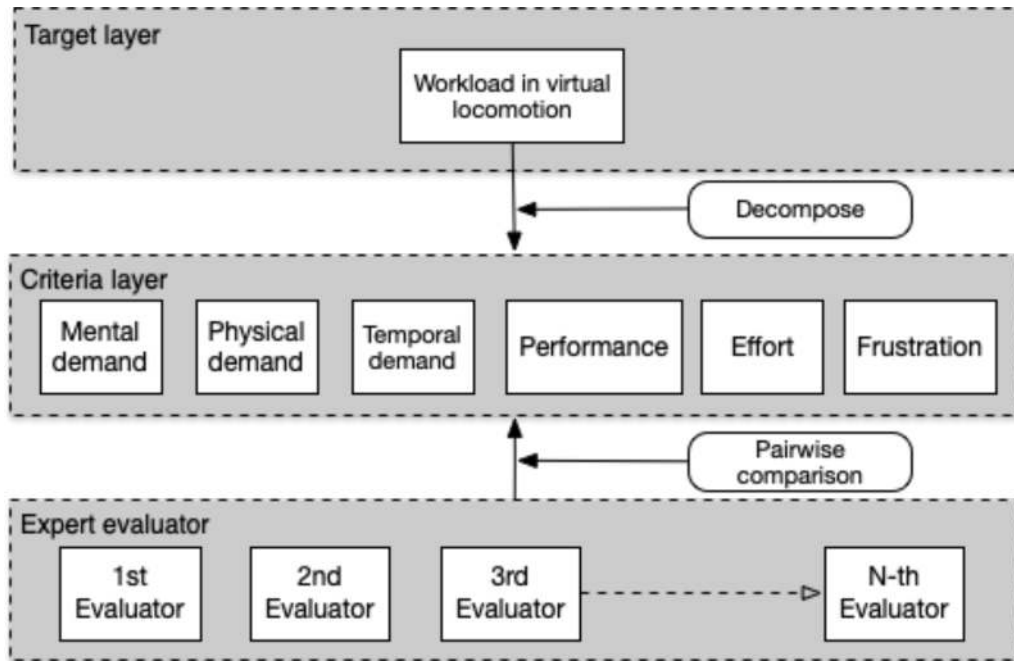

Fig. 2. A specific case of hierarchical structure to measure workload with the NASA-TLX criteria.

2.2.3. Procedures to perform a fuzzy AHP analysis

Step 1: Problem analysis and hierarchical structure formulation Following the general procedures used in the AHP method, it is necessary to define factors that affect the goal. For example, a complex problem like workload can be decomposed into many criteria, and each criterion can be decomposed into many sub-criteria which can also be further decomposed (Fig. 1), as in the NASA-TLX table. It is worth mentioning that the proposed approach can be applied to other workload measurement methods during a simulation experience, and not only to the NASA-TLX, as long as the measuring criteria from multigroups can be validated. For example, Harris et al. (2019) propose a simulation task load index (SIM-TLX) where the criteria are introduced by integrating the NASA-TLX, the SURG-TLX and external indexes (Wilson et al., 2011). In this case, there are three groups of criteria, and each group has its own sub-criteria that should be placed according to

the structure presented in Fig. 1. And thanks to the hierarchical structure, the total number of comparisons would exponentially reduce despite increasing measuring criteria. In our case, we applied this structure to the NASA-TLX and for a navigation task in a virtual environment; the target or the goal was: which criterion is more important to reduce workload during virtual locomotion? Pairwise comparisons are performed for each criterion based on the announced target. The corresponding hierarchical structure is shown in Fig. 2.

Step 2: Determining the fuzzy linguistic scale

When requested to compare or evaluate objects, individuals generally use linguistic expressions such as "very important", "little important", "good" and "bad". These linguistic expressions contain uncertain and fuzzy information that needs to be processed. To best fit subjective evaluations, we adapted pre-defined linguistic expressions from previous studies (Hong et al., 2018; Novák and Perfilieva, 1999) that relied on

Table 1
Linguistic rating scales and corresponding fuzzy numbers

Linguistic scales	TFNs
Equally important (EI)	(1, 1, 1)
Weakly more important (WI)	(1, 1, 3/2)
Strongly more important (SI)	(3/2, 2, 5/2)
Very strongly more important (VI)	(5/2, 3, 7/2)
Absolute important (AI)	(7/2, 4, 9/2)

the original work of Saaty (1987). Comparative judgment generally uses expressions from a linguistic set $\mathcal{F}=\{$Equally important, Weakly more important, Strongly more important, Very strongly more important, Absolutely important$\}$ that maps a corresponding fuzzy number, as presented in Table 1.

Step 3: Fuzzy comparison using the fuzzy linguistic scale

Experienced evaluators are invited to fill a comparison table where they perform pairwise comparisons. Considering NASA-TLX, questions can be "what is the relative importance of criterion $C_i, i = 1, 2, .., p$ compared to $C_j, j = 1, 2, .., p, i \neq j$ to measure cognitive workload during a virtual locomotion". The filled comparison table forms a fuzzy comparison matrix, denoted by $\widetilde{C}$,

$$
\widetilde{C} = \begin{matrix} & \begin{matrix} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 \end{matrix} \\ \begin{matrix} C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \\ C_6 \end{matrix} & \begin{pmatrix} 1 & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{21} & 1 & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{31} & c_{32} & 1 & c_{34} & c_{35} & c_{36} \\ c_{41} & c_{42} & c_{43} & 1 & c_{45} & c_{46} \\ c_{51} & c_{52} & c_{53} & c_{54} & 1 & c_{56} \\ c_{61} & c_{62} & c_{63} & c_{64} & c_{65} & 1 \end{pmatrix} \end{matrix}
\tag{7}
$$

where $c_{ij} = \frac{1}{c_{ji}}$, and

$$
\begin{cases} c_{ij} = \left(l_{ij}, m_{ij}, u_{ij} \right) \\ \dfrac{1}{c_{ji}} = \left(\dfrac{1}{u_{ij}}, \dfrac{1}{m_{ij}}, \dfrac{1}{l_{ij}} \right) \end{cases}
\tag{8}
$$

An example of matrix is provided in Annex Appendix A.

Each evaluator provides a fuzzy comparison matrix. The final evaluation matrix $\widetilde{C}^h$ aggregates the responses from all evaluators. Assuming we have H evaluators, we here take the average of all fuzzy comparison matrices,

$$
c_{ij} = \left(\frac{1}{H} \sum_{h=1}^{H} l\left(c_{ij}^h \right), \frac{1}{H} \sum_{h=1}^{H} m\left(c_{ij}^h \right), \frac{1}{H} \sum_{h=1}^{H} u\left(c_{ij}^h \right) \right)
\tag{9}
$$

where $l(\cdot), m(\cdot)$ and $u(\cdot)$ are functions to find the lower, modal and upper values of the TFN.

Step 4: Weighting vector determination using the extent analysis

With the extent analysis method formulated in Eq. 4, we are able to describe each criterion with a triangular fuzzy number. In order to determine the weighting vector of the criteria, the principle of comparison of fuzzy numbers must be used.

Assume that $d'(C_i) = \min V(S_i \geq S_j)$ for $j = 1, 2, ...p, i \neq j$, the weighting vector of p criteria is defined as,

$$
W' = \left[d'(C_1), d'(C_2), ..., d'(C_p) \right]^T
\tag{10}
$$

where $C_i, i = 1, 2, ..., p$ are the criteria. Applying a normalisation operation, the final normalised weighting vector is

$$
\begin{aligned} W &= \left[\frac{d'(C_1)}{\sum_{i=1}^{p} d'(C_i)}, \frac{d'(C_2)}{\sum_{i=1}^{p} d'(C_i)}, ..., \frac{d'(C_p)}{\sum_{i=1}^{p} d'(C_i)} \right]^T \\ &= \left[d(C_1), d(C_2), ..., d(C_p) \right]^T \end{aligned}
\tag{11}
$$

3. TOPSIS method

The TOPSIS method proposed by Hwang and Yoon (1981) is a comprehensive within-group evaluation method that can make full use of raw data and reflects the gap between the evaluated objects. The basic idea is developed based on normalised original data represented in the matrix form to find the optimal and the worst solutions within a finite set of alternatives. Then, the distance between each evaluation object and the optimal and the worst solutions is calculated separately, which gives the relative closeness. This value is used as the basis for evaluating the merits and demerits. The method does not strictly rely on the data distribution and sample content, and the calculation process is simple and easy.

3.1. Data homogenization

The TOPSIS method uses the distance scale to measure the difference among samples. To use the scale, it is necessary to normalise the index attributes in the same manner (e.g., for attribute A, the bigger the number, the better the result, while for attribute B, the smaller the number, the better the outcome. Such inconsistency can lead to inconvenience and confusion for the calculation in the next steps). Usually, a cost-type indicator is converted to a benefit-type indicator (that is, the higher the value, the better the result; in fact, almost all evaluation methods need this step to homogenise the raw data).

3.2. Construction of the normalised matrix

Let n be the number of objects to be evaluated, each object has m attributes; then the original data matrix is constructed as,

$$
X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix}
\tag{12}
$$

To perform dimensionless calculations, we need to construct a weighted canonical matrix in which the attributes are normalised vectors, that is, each column element is divided by the norm of the current column vector (using the cosine distance measure),

$$
z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}
\tag{13}
$$

The normalised non-dimensional matrix becomes,

$$
Z = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,m} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n,1} & z_{n,2} & \cdots & z_{n,m} \end{pmatrix}
\tag{14}
$$

3.3. Identification of the optimal and the worst solutions

There exist two idealised goals. One is the positive ideal goal or the optimal goal, and the other one is the negative ideal solution or the worst goal. The positive optimal solution Z^+ consists of the maximum value of each column element in Z:

$$
Z^+ = \begin{pmatrix} \max\left(z_{1,1}, z_{2,1}, \cdots, z_{n,1} \right) \\ \max\left(z_{1,2}, z_{2,2}, \cdots, z_{n,2} \right) \\ \vdots \\ \max\left(z_{1,m}, z_{2,m}, \cdots, z_{n,m} \right) \end{pmatrix}
\tag{15}
$$

The negative worst solution Z^- consists of the minimum value of each column element in Z:
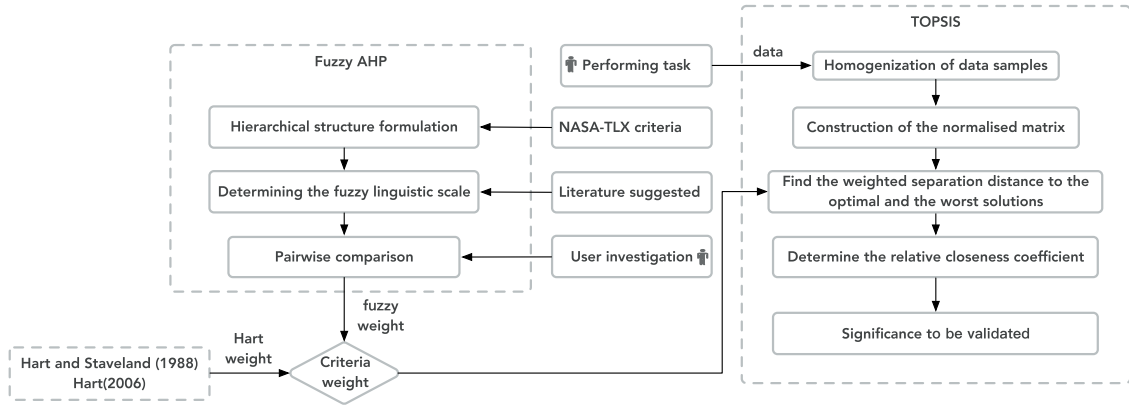
Fig. 3. Schematic process of the proposed TOPSIS evaluation procedure; the TOPSIS method can use different weighting approaches.

$$Z^- = \begin{pmatrix} \min(z_{1,1}, z_{2,1}, \cdots, z_{n,1}) \\ \min(z_{1,2}, z_{2,2}, \cdots, z_{n,2}) \\ \vdots \\ \min(z_{1,m}, z_{2,m}, \cdots, z_{n,m}) \end{pmatrix} \qquad (16)$$

3.4. Calculation of the separation distance

The TOPSIS algorithm can be easily adapted to process subjective evaluations by considering that the best criterion to evaluate should be the closest to the optimal solution, and the farthest to the worst one. The separation distance is used to measure the distance between the current sample and the optimal/worst solution, and can be defined as the following,

$$\begin{cases} D_i^+ = \sqrt{\sum_{j=1}^{m} w_j \left(Z_j^+ - z_{i,j} \right)^2} & (17a) \\ D_i^- = \sqrt{\sum_{j=1}^{m} w_j \left(Z_j^- - z_{i,j} \right)^2} & (17b) \end{cases}$$

where w_j is the weight of the jth attribute (importance) and is derived from the fuzzy AHP method explained in the previous section.

3.5. Determination of the relative closeness coefficient

Since the separation distance has two independent values (D_i^+ and D_i^-), to get a unified measurement in one dimension, the relative closeness coefficient (RCC) of each entry is defined as,

$$RCC_i = \frac{D_i^-}{D_i^+ + D_i^-}, \quad RCC_i \in [0, 1] \qquad (18)$$

When $RCC_i \rightarrow 1$, D_i^+ is small and D_i^- is large, it means that the measured object is close to having the best performance. However, depending on how the index attributes are formulated, it may be preferable to have a value of RCC_i close to 0, e.g., if the attributes are homogenised as a cost-type problem.

The performance of each object can be defined by the values of the RCC, ranking from large RCC to small RCC or in reverse order. Applying this principle to our case, a larger RCC value suggests a higher workload. In addition to observing the performance of each object individually, the RCC can be grouped according to the factors tested in the experiment and the difference of RCC can be compared among groups (e.g., HMD vs CAVE).

3.6. Composed multiple criteria evaluation system

Fig. 3 presents the whole diagram of the proposed method to evaluate cognitive workload using TOPSIS from the fuzzy AHP and the *Hart* weighting approaches. The framework requires two steps. The first step is to compute the weighting coefficients from either the fuzzy AHP approach, the pairwise comparisons suggested by Hart and Staveland (1988) and Hart (2006), or the uniform weighting approach. The second step is to use TOPSIS to compute the relative closeness coefficient using the weighting information obtained from the first step.

4. User study

We designed an experiment to investigate the effectiveness of the proposed method to enhance the precision of the cognitive workload evaluation. We applied our tool in a locomotion case in a VR application. In this experiment, we considered two factors that encounters in navigation experience in terms of cognitive load: the VR platform type (Sevinc and Berkman, 2020) and the scenario density (Parsons et al., 2009). Based on this principle, we conducted the experiment to study the effect of these two factors on cognitive workload.

If the proposed methods can measure cognitive workload with the expected performance after a VR task, we shall obtain the following results:

R1 The *RCC* can measure the cognitive workload arising from different VR platforms and scenario densities, with consistent results with the literature

R2 The *RCC* provides more precise information on cognitive workload than the classical weighting sum method.

R1 helps identifying the effectiveness of the proposed method as it could lead to results similar to the other methods and measure cognitive workload successfully. R2 further shows the advantage of this method over the others as it could reduce the data dispersion, represented by the CV.

4.1. Experiment equipment

The experiment was conducted in both a CAVE system and an HTC Vive head-mounted display to compare the effect of different displays on workload. The CAVE system consists of five walls with 4096 × 2160 px projectors that display stereoscopic images at a refresh frame rate of 120Hz. An in-house interface written in C++ was developed to connect all the devices of the system. On the other hand, the HTC Vive has a display resolution of 2160 × 1200 px with a refresh rate of 90Hz.

Navigation in virtual environments (forward/backward and left/right movements) was achieved using the joystick of an ART Flystick controller in the CAVE and the touchpad of the Vive controller with the HMD.
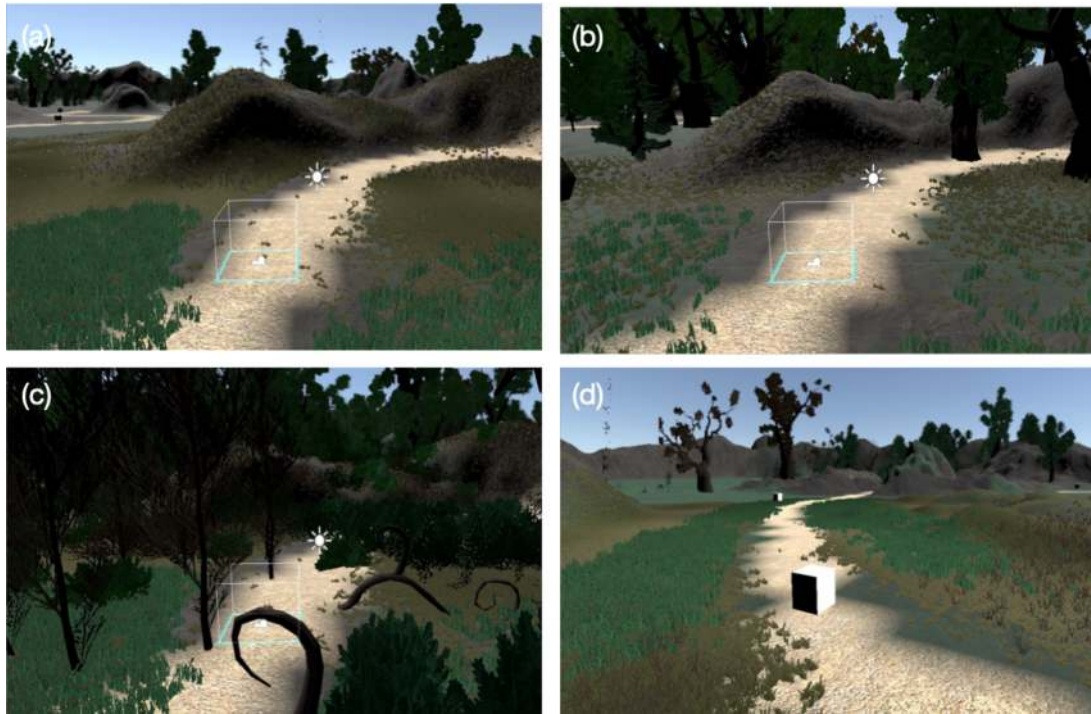
Fig. 4. Scenarios with (a) low, (b) middle and (c) high densities; (d): scenario with checkpoints where the user has to stop and interact with.
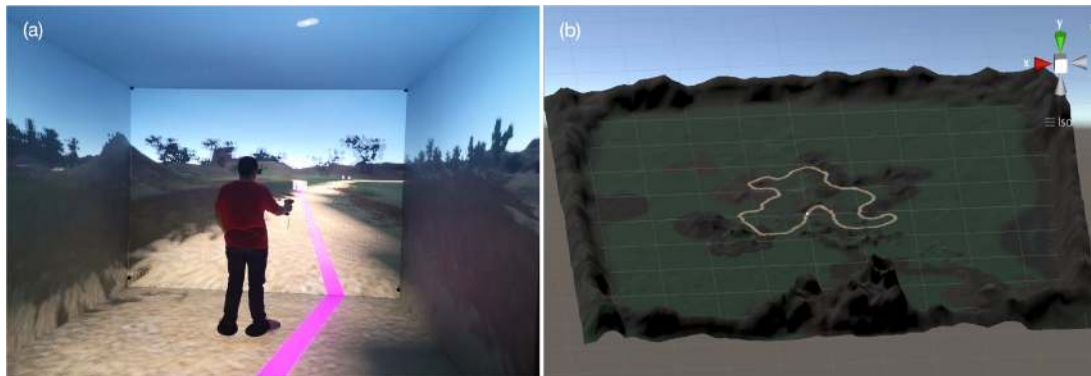

Fig. 5. (a) One user is doing the experiment inside the CAVE system, the pink line instructs the initial locomotion direction to prevent the user from getting lost once immersed; (b) overview of the 3D virtual environment with the trajectory to follow in beige colour.

4.2. Participants

We invited fifteen subjects including engineers and students (M_{age} = 23.1, SD_{age} = 1.82) from the university to participate voluntarily without compensation in the experiment. Each of them was given a brief introduction to the experiment and to the potential risks that could be encountered. They were allowed to terminate the experiment if they felt strong discomfort. Also, a pre-exposure questionnaire was filled by these participants to gain information about their health condition and background that would be utilised for data analysis if necessary. To minimise random noise, we performed a within-subject test. A consent form was also signed by participants.

4.3. Task design

Figure 5 shows how users conducted the task in the CAVE system and an overview of the 3D VE with the trajectory to follow highlighted in beige colour. The task performed in the HTC Vive followed the same settings as in the CAVE. The scenario consisted in walking in a forest area. By changing the density of objects present in the environment such as trees and flowers, we generated three different scenarios that we denoted by low (few objects), middle and high (many objects) densities (Fig. 4). According to the definition of cognitive workload in section 1, such a setting should impose to participants different levels of cognitive workload: the high-density scenario should be more cognitively demanding than the low-density scenario because the user needs more cognitive resources to interact with the different objects and perform tasks. In the low density scenario, the user should be rarely affected and distracted when performing tasks, thus implying less workload. The general experimental procedure was organised as follows:

- Participants filled a pre-exposure questionnaire. Each participant was given a training about how to use the CAVE or the HTC Vive, and what they had to do in the 3D VE.
- With the help of the experimenter, participants put on the devices and were exposed to the virtual environment. Then, participants started to navigate in the VE following a predefined path. To increase the task difficulty level, users had to touch each checkpoint
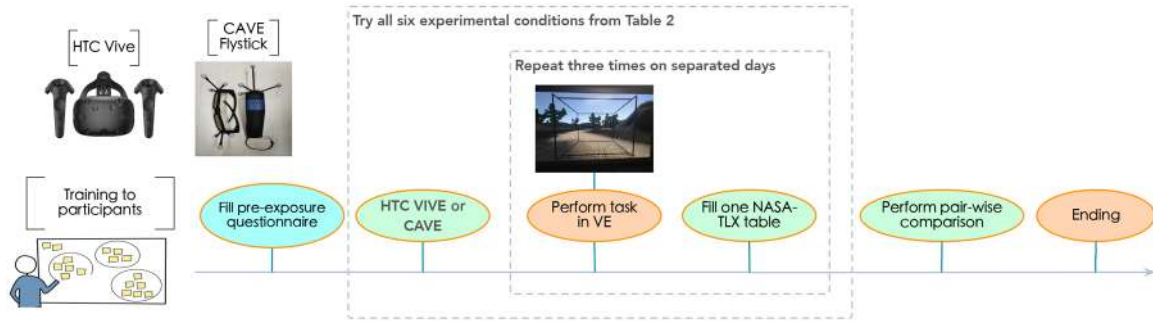
Fig. 6. Flow chart of the experiment.

Table 2
Experimental conditions

Experiment number	1	2	3	4	5	6
Platform type	HTC Vive	HTC Vive	HTC Vive	CAVE	CAVE	CAVE
Scenario density	Low	Middle	High	Low	Middle	High

represented by the cube and avoid touching obstacles (trees, flowers) in the scenario with the handed device (shown in Fig. 6, the HTC Vive controller in the HMD condition or the Flystick controller in the CAVE condition), then resume navigation (Fig. 4d).

- When participants reached the destination inside the VE, they were removed from the CAVE or the HTC Vive and were requested to fill a NASA-TLX table based on their experience and impression.
- The experiment was repeated three times under the same conditions on three separate days.

The experimental conditions are listed in Table 2. They were uniformly distributed and balanced, to minimise random errors and hybrid effects. Six different conditions had to be tested in random order, and each condition was repeated by one user three times on separate days. Totally, one user had to perform 18 (6 × 3) tests as we had six conditions. The flow chart of the experimental protocol is represented in Fig. 6. Participants were allowed to stop if they felt any discomfort.

4.4. Determination of weighting coefficients

We computed two different sets of weighting coefficients. The *Hart* weighting coefficient approach explained in subsection 2.1 was operated strictly following the instructions from Hart and Staveland (1988) and Hart (2006).

Participants who attended the experiments and also two professors were invited to form an individual pairwise comparison matrix using the linguistic expressions given in Table 1; those raters are chosen because of their expertise in cognitive workload evaluation in VR applications: participants knew what was important for them because of their in-person experience in the test, and two external raters performed pairwise comparison based on professional knowledge. A detailed procedure about the rating method could be found in Gumus (2009). The comparison among the criteria was conducted by considering which one was more important concerning the cognitive demand. To this end, participants were asked to fill the comparison questionnaire given in Appendix A, which led to a fuzzy comparison matrix. As explained in subsection 2.2, the normalised weightings of the aggregated responses were derived from the fuzzy synthetic extent analysis.

5. Results

After the experiments, we collected all the questionnaires. We first computed the weighting coefficients using the *Hart* weighting and the fuzzy weighting approaches. Then, with these weighting coefficients available, we computed the *RCC* based on the TOPSIS method and the
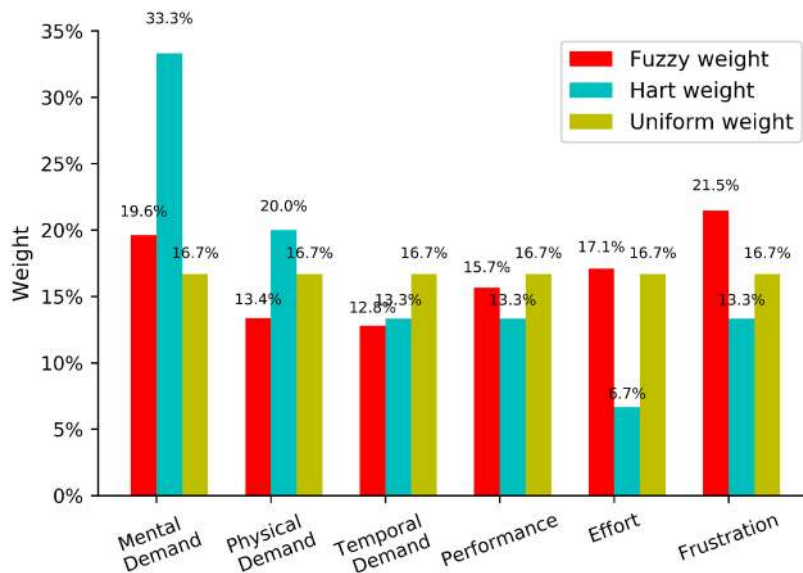

Fig. 7. Aggregated normalised weights of the NASA-TLX criteria from three different approaches: the fuzzy AHP approach, the *Hart* weighting approach and the uniform weighting approach.

Table 3
Effect of factors on workload from different determination methods. Significance level:.05 (*),.01 (**)

	Fuzzy weight		Hart weight		Uniform weight	
TLX						
Factors	Platform	Scenario	Platform	Scenario	Platform	Scenario
Sum sq	7.46	22.93	4.88	27.47	6.94	16.42
NumDF	1	2	1	2	1	2
DenDF	236.09	236.17	235.96	236.04	236.02	236.11
F	2.10	3.23	1.46	4.12	2.15	2.55
p-value	.14	.04*	.23	.02*	.14	.08
η_p^2	.01	.03	.01	.03	.01	.02
RCC						
Factors	Platform	Scenario	Platform	Scenario	Platform	Scenario
Sum sq	0.0008	0.07	0.0003	0.08	0.0003	0.05
NumDF	1	2	1	2	1	2
DenDF	236.03	236.11	235.91	235.99	235.95	236.23
F	.12	4.72	.04	5.37	.05	4.12
p-value	.73	.009**	.84	.005**	.83	.02*
η_p^2	<.01	.04	<.01	.04	<.01	.03

TLX_{Hart} score according to the weighted sum method. A statistical analysis was then performed to compare the RCC and TLX_{Hart}.

5.1. Weighting coefficients

Figure 7 illustrates the final weighting values each NASA-TLX criterion according to the weighting determination approach including the fuzzy AHP, the classical comparison process (*Hart*) and the uniform weighting. The uniform weights were obtained by setting the same weighting coefficient to all indices, and were used in this study as a control group, which helped to find the effect of different weighting coefficients determination approaches. From Fig. 7, we can observe that the weights of the NASA-TLX indices strongly vary depending on the approaches. From the fuzzy AHP, *Frustration* and *Mental Demand* were the top two important factors, with a weighting coefficient of 21.5% and 19.6% respectively; the *Effort* weight was relatively low (17.1%) followed by the *Performance* weight (15.7%); *Temporal Demand* and *Physical Demand* had the lowest weights, 12.8% and 13.4% respectively. In parallel, the classical pairwise comparison indicated the highest weighting coefficient for *Mental Demand* (33.3%), followed by *Physical Demand* (20.0%); *Temporal Demand, Performance* and *Frustration* got equal weights (13.3%), while *Effort* got the lowest weight (6.7%).

The obtained weighting values from the three approaches were then provided to the TOPSIS model in order to get the RCC with Eq. 18. The resulting overall workload scores were then compared with the classical weighted sum method (denoted TLX_{Hart} in Eq. 1).

5.2. Workload assessment

Three participants performed only part of the eighteen conditions due to their availability. Therefore, this was not a perfect within-subject experiment because of the unbalanced samples, and we decided to analyze the variance with the mixed-effects model which is an extension of the repeated-measures ANOVA but with more flexibility (Galecki and Burzykowski, 2013). Independently from the platform type and scenario density, individual differences (e.g., the way of thinking, judgment criteria) can be a potential factor that affects the overall workload score. In order to reduce these individual effects, we set the effect of individual difference as random factor while we set the effect of the scenario density and the platform type as fixed effects. Statistics were conducted in R (R Core Team, 2020) along with the related packages: lme4, afex, lmerTest and effectsize.

A mixed-effects model was conducted to determine the influence of two independent variables (platform, scenario) on cognitive workload considering both the TLX_{Hart} and RCC methods (Table 3). Normality of the data was checked. Three coefficient weighting approaches were applied for each cognitive workload determination method. Considering first the TLX_{hart} evaluation, the level of scenario density was found to have significant effects on the overall workload only when it was measured with the fuzzy AHP ($F_{2,236.17} = 3.23, p = .04, \eta_p^2 = .03$) and the *Hart* approaches ($F_{2,236.04} = 4.12, p = .02, \eta_p^2 = .03$). On the contrary, the VR platform did not have any significant effect whatever the weighting determination approach (fuzzy weighting: $F_{1,236.09} = 2.10, p = .14, \eta_p^2 = .01$, *Hart* weighting: $F_{1,235.96} = 1.46, p = .23, \eta_p^2 = .01$ and uniform weighting: $F_{1,236.02} = 2.15, p = .14, \eta_p^2 = .01$). With the RCC evaluation, the level of scenario density revealed a significant effect with all three weighting approaches (fuzzy weighting: $F_{2,236.11} = 4.72, p < .01, \eta_p^2 = .04$, *Hart* weighting: $F_{2,235.99} = 5.37, p < .01, \eta_p^2 = .04$, uniform weighting: $F_{2,236.23} = 4.12, p = .02, \eta_p^2 = .03$); the effect of the platform still showed no significant difference whatever the weighting approach (fuzzy weighting: $F_{1,236.03} = 0.12, p = .73, \eta_p^2 < .01$, *Hart* weighting: $F_{1,235.91} = 0.04, p = .84, \eta_p^2 < .01$ and uniform weighting: $F_{1,235.95} = 0.05, p = .83, \eta_p^2 < .01$).

Post-hoc analyses were performed to understand the differences between the three levels of scenario density. Pairwise comparisons were done using Tukey HSD tests. Results are shown in Table 4. In addition to p-values, we considered the confidence interval (CI) to provide additional information that p-values do not convey, for example the actual mean difference between groups. The width of the CI for the difference reveals the precision of the estimate, and narrower intervals suggest a more precise estimate (Lee, 2016). With the TLX_{hart} evaluation,

Table 4
Post-hoc analysis for the scenario type considering different determination methods; Lower and Upper represent the boundaries of the 95% confidence interval (CI). Significance level:.05 (*),.01 (**)

	Fuzzy weight			Hart weight			Uniform weight		
TLX									
Scenario type	low	low	middle	low	low	middle	low	low	middle
Scenario type	middle	high	high	middle	high	high	middle	high	high
Mean difference	-0.44	-0.74	-0.30	-0.52	-0.81	-0.29	-0.38	-0.63	-0.25
Lower	-1.01	-1.33	-0.89	-1.07	-1.37	-0.86	-0.92	-1.18	-0.81
Upper	0.12	-0.16	0.29	0.03	-0.24	0.29	0.16	-0.07	0.31
p-value	.12	.01*	.31	.06	.005**	.32	.17	.02*	.39
η_p^2	.01	.03	<.01	.01	.03	<.01	.01	.02	<.01
RCC									
Scenario type	low	low	middle	low	low	middle	low	low	middle
Scenario type	middle	high	high	middle	high	high	middle	high	high
Mean difference	-0.03	-0.04	-0.01	-0.03	-0.04	-0.01	-0.02	-0.03	-0.01
Lower	-0.05	-0.07	-0.04	-0.05	-0.07	-0.04	-0.05	-0.06	-0.04
Upper	-0.001	-0.01	0.01	-0.003	-0.02	0.01	-0.001	-0.01	0.01
p-value	.04*	.003**	.31	.03*	.002**	.33	.05*	.006**	.39
η_p^2	.02	.04	<.01	.02	.04	<.01	.02	.03	<.01

Table 5
Descriptive statistics with the mean (M), the standard deviation (SD) and the CV with the different approaches

		Fuzzy weighting			Hart weighting			Uniform weighting		
TLX_{Hart}										
		M	SD	CV	M	SD	CV	M	SD	CV
Platform	HTC	7.53	2.90	0.39	7.4	2.93	0.40	7.48	2.83	0.38
	CAVE	7.26	3.16	0.44	7.21	3.22	0.45	7.22	3.10	0.43
Scenario type	low	7.04	2.74	0.39	6.92	2.80	0.40	7.05	2.74	0.39
	middle	7.45	2.91	0.39	7.37	2.89	0.39	7.4	2.84	0.38
	high	7.79	3.42	0.43	7.72	3.50	0.45	7.69	3.31	0.43
RCC										
Platform	HTC	0.40	0.13	0.33	0.38	0.14	0.37	0.40	0.13	0.33
	CAVE	0.40	0.14	0.35	0.38	0.15	0.39	0.40	0.14	0.35
Scenario type	low	0.38	0.13	0.34	0.36	0.13	0.36	0.39	0.13	0.33
	middle	0.40	0.13	0.33	0.39	0.13	0.33	0.41	0.12	0.29
	high	0.42	0.16	0.38	0.40	0.17	0.43	0.42	0.15	0.36

Table 6
Statistical analysis of CV for the different evaluation methods and weighting approaches. Significance level:.05 (*),.01 (**)

Mixed-effects model		Sum Sq	NumDF	DenDF	F	p-value	η_p^2
	TLX v.s. RCC	0.30	1	30	38.23	.00**	.56
	Weighting methods	0.005	2	30	3.54	.04*	.19
Multi-comparison		Mean Difference	Lower	Upper		p-value	η_p^2
	TLX RCC	-0.06	-0.08	-0.04		.00**	.56
	Fuzzy weight Hart weight	-0.02	-0.04	0.003		.09	.09
	Fuzzy weight Uniform weight	0.01	-0.01	0.03		.39	.02
	Hart weight Uniform weight	0.03	0.007	0.05		.01*	.19

statistical significance was observed only between the low and high densities and with two weighting approaches (the fuzzy weighting: $p = .01$, 95% $CI : -1.33 \sim -0.16, \eta_p^2 = .03$, *Hart* weighting: $p < .01$, 95% $CI : -1.37 \sim -0.24, \eta_p^2 = .03$, uniform weighting: $p = .02$, 95% $CI : -1.18 \sim -0.07, \eta_p^2 = .02$). In contrast, the RCC evaluation reported more significant effects. Statistical significance was found between the low and middle densities (fuzzy weighting: $p = .04$, 95% $CI : -0.05 \sim -0.001, \eta_p^2 = .02$, *Hart* weighting: $p = .03$, 95% $CI : -0.05 \sim -0.003$, $\eta_p^2 = .02$ and uniform weighting: $p = .05$, 95% $CI : -0.05 \sim -0.001$, $\eta_p^2 = .02$), as well as between the low and high densities also with all three weighting methods (fuzzy weighting: $p < .01$, 95% $CI : -0.07 \sim -0.01, \eta_p^2 = .04$, *Hart* weighting: $p < .01$, 95% $CI : -0.07 \sim -0.02$, $\eta_p^2 = .04$ and uniform weighting: $p < .01$, 95% $CI : -0.06 \sim -0.01, \eta_p^2 = .03$). We can also remark that the CIs from the RCC are much narrower than with TLX_{Hart}.

5.3. Advantage of TOPSIS over the weighted sum method

When the overall workload was measured with the weighted sum method, the range of the CIs was wider than with the TOPSIS method, as shown in Table 4, revealing that TOPSIS provided more precise estimates than TLX_{Hart}. As a supplementary support, we considered the CV in addition to the mean (M) and the standard deviation (SD), obtained for each platform and scenario type with the three different weighting approaches and the two evaluation methods (Table 5). The CV is defined as the ratio of SD to M and represents a statistical measure of the dispersion of the data (Kesteven, 1946; Lovie, 2005). The higher the CV, the higher the dispersion and the less reliable the measure.

Another mixed-effects model was carried out to analyze the CV from Table 5, where the random effects were the platform type and the scenario density and the fixed effects were the evaluation methods (TLX_{Hart}, RCC) and the weighting approaches (fuzzy weighting, *Hart* weighting, uniform weighting). Table 6 presents the results of the statistical analysis. A significant difference was found between the evaluation methods in terms of CV, $F_{1,30} = 38.23, p < .01, \eta_p^2 = .56$. The weighting approach
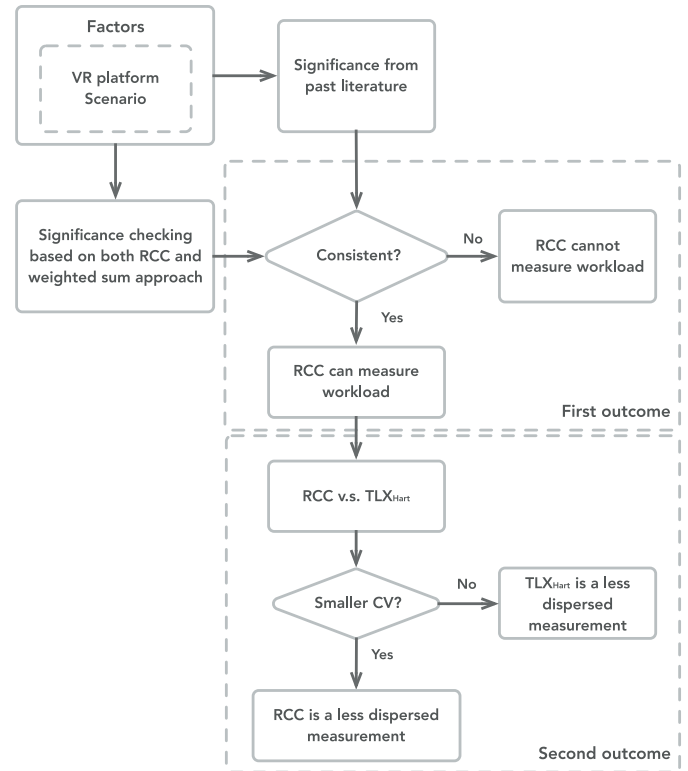

Fig. 8. Flowchart of the validation process for the two factors considered here: first to check if there is a consistent result to show the proposed method working; second to compare the dispersion of measurements based on the CV.

also influenced the CV significantly, $F_{2,30} = 3.54, p = .04, \eta_p^2 = .19$. Post-hoc analyses revealed that the RCC led to significantly smaller CVs than with TLX_{Hart}, $p < .01, \eta_p^2 = .56$. Furthermore, it was found that the CV from the *Hart* weighting was significantly higher than with the uniform

weighting, $p = .01,\ 95\%\ CI : 0.007 \sim 0.05, \eta_p^2 = .19$. However, as an exploratory study, we didn't find difference at the given significance level between the fuzzy weighting and *Hart* weighting, $p = .09,\ 95\%\ CI : -0.04 \sim 0.003, \eta_p^2 = .09$.

Therefore, the evaluation from TOPSIS illustrated a significantly lower dispersion of the data compared to TLX_{Hart}, which confirms our expectation that TOPSIS provides enhanced precision of subjective evaluation.

6. Discussion

We proposed to use the *RCC* as an enhancement for the measurement of cognitive workload, with the expected outcomes that: the *RCC* is a more precise metric than classical methods to quantify cognitive workload resulting from different tasks or systems. In the considered use case, we could investigate the effectiveness of the *RCC* to discriminate cognitive workload in different VR platforms and scenario densities. Fig. 8 shows the two steps operated to study the performance of our approach. First, evaluation results from TOPSIS were checked for consistency with results from the literature (R1). Then, the precision of the evaluation methods was checked by comparing the CVs between *RCC* and TLX_{Hart} (R2).

6.1. Validation of R1

No evidence was found that different VR platforms can lead to significantly different levels of workload, neither the *RCC* nor TLX_{Hart} showed such effect. In this sense, our results conformed to the findings of past research (Freitas, 2018; Porssut and Chardonnet, 2017; Riley and Kaber, 1999; Sevinc and Berkman, 2020) and confirmed that the *RCC* could represent well the level of cognitive workload, with respect to the VR platforms. Concerning the scenario density, past work proved that the type of scenario can affect cognitive workload (Parsons et al., 2009). From Table 3, both the *RCC* and TLX_{Hart} methods behaved consistently.

No significant difference was found between the middle and the high-density scenarios whatever the evaluation method. Past research showed that depending on the task difficulty, there exist three workload states that are cognitive under-load, adequate workload and cognitive saturation (Harrison et al., 2014; McKendrick et al., 2019). In our experiment, we set densities to produce these different states. From the *RCC* results, when participants performed the task in the low-density scenario, as it was less demanding, they were in an under-load state, resulting in significantly lower cognitive load than in the middle density case. As the scenario density increased, their cognitive ability to receive and process the spatial information was around overload or had already become overloaded, which explains why there was no significant difference between the middle and the high-density scenarios, whatever the evaluation method.

From these observations, our first expected outcome was achieved.

6.2. Improved workload measurement by reducing dispersion

Much research has been done over the last decades to measure cognitive workload during a task, especially in fields related to human-computer interaction (Gevins and Smith, 2003), driving (Patten et al., 2006) and flight (Kantowitz and Casper, 2017; Sterman and Mann, 1995), in which users have to process amounts of information simultaneously, resulting in high cognitive workload. Many past literature methods try to apply biosensors to measure cognitive workload, while evaluation through subjective questionnaires did not progress significantly. During the subjective evaluation process, participants find it difficult to quantify their impressions towards the experience, but they are forced to give answers, leading to subjective results with high uncertainties (Katicic et al., 2015). However, we believed that a proper analysis approach can address this drawback and enhance the reliability

of individual feedback. Therefore, we introduced an alternative evaluation method to the straightforward weighted sum method to quantify workload based on subjective evaluation results more precisely. We compared the weighted sum method to the TOPSIS method considering three different weighting approaches.

From the experimental results, regarding the scenario type, with the classical weighted sum calculation method, statistical significance could be detected only between the low and high densities, implying a lack of precision of this method to discriminate the overall workload in smaller scales. In contrast, with the TOPSIS method, more significant effects were detected, as differences between low and middle densities were found. TOPSIS could therefore more precisely detect small changes in cognitive workload. The reason lies in the significantly smaller CV of the *RCC*, meaning a lower dispersion of the data and therefore an improved quality of discrimination among data.

Regarding the weighting approaches, we did not find a significant difference between the fuzzy AHP and the *Hart* weighting approaches as both of them gave similar significant results when applied to TLX_{hart} and *RCC*, but the *Hart* weighting resulted in a higher CV than with the uniform weighting approach. In other words, the fuzzy AHP and the *Hart* weighting approaches provided effective results, while particularly the fuzzy weighting could be used as an alternative to the *Hart* weighting and the uniform weighting approaches to determine the weighting coefficients for the overall cognitive workload score. This finding implies that applying weights to each criterion according to its importance is an essential step for computing the overall workload.

The comparison of CV validated our second outcome R2. It suggested that the cognitive workload measured from the TOPSIS was more reliable, which would be particular important for those wishing to compare and manage the workload difference in smaller scales.

7. Conclusion

We introduced a new method to improve the precision and reduce the CV of cognitive workload quantification thanks to the TOPSIS. The model was tested with three different weighting approaches: fuzzy weighting, *Hart* weighting and uniform weighting. Thanks to its hierarchy structure, the fuzzy AHP method for computing the fuzzy weighting extended the possibility to measure the workload with other questionnaires instead of the NASA-TLX. The proposed method was applied in a VR user experiment to validate its performance by studying two factors in a navigation task: the type of VR platform and the scenario density. Results were compared with a classical weighted sum method. The *RCC* computed from TOPSIS was found to be a comprehensive metric for quantifying and comparing the level of workload among various VR applications, with a reduced CV on subjective evaluation compared to the classical weighted sum method. Validation results were consistent with the corresponding literature, which suggested that our new framework can be useful in assessing workload measurement by reducing subjective uncertainty and improving measuring quality.

Because of the increasing popularity of VR, it is important to consider cognitive workload in this domain. Therefore, the TOPSIS method for measuring cognitive workload was designed and validated in the field of VR, while it can also benefit to the measurement of cognitive workload arising from other domains by following our experimental steps. Nevertheless, one substantial limitation was that only two factors were considered to determine the effectiveness of our approach; future research will consider more factors as well as more use cases to test the generality of this approach entirely.

CRediT authorship contribution statement

Yuyang Wang: Writing - original draft, Formal analysis, Methodology. **Jean-Rémy Chardonnet:** Writing - review & editing, Conceptualization. **Frédéric Merienne:** Writing - review & editing, Supervision, Resources, Funding acquisition.

Table 7

Fuzzy comparison table to be filled with the linguistic expressions (MD: mental demand; PD: physical demand; TD: temporal demand; P: performance; F: frustration)

	MD	PD	TD	P	E	F
MD	EI					
PD		EI				
TD			EI			
P				EI		
E					EI	
F						EI

Table 8

One filled questionnaire from the experiment (MD: mental demand; PD: physical demand; TD: temporal demand; P: performance; F: frustration); the cells with "-" are automatically filled during the data analysis as it has a reciprocal relationship with the item from the other side of the diagonal

	MD	PD	TD	P	E	F
MD	EI	-	WI	WI	WI	EI
PD	WI	EI	WI	-	-	SI
TD	-	-	EI	-	-	WI
P	-	SI	SI	EI	EI	-
E	-	WI	SI	-	EI	SI
F	-	-	-	WI	-	EI

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Fuzzy comparison questionnaire

A1. Questionnaire design

- Name:
- The Nasa Task Load Index (NASA-TLX) is a widely used, subjective, multidimensional assessment tool that rates perceived workload in order to assess a task, system, or team's effectiveness or other aspects of performance; more precisely: mental demand, physical demand, temporal demand, performance, effort, frustration.
- In order to create individual weighting of these sub-scales, please fill the following table (see Table 7) using the pre-defined language expressions considering: which one do you think is more important during the task you have just performed?
- Language expressions: equally important (EI), weakly more important (WI), strongly more important (SI), very strongly more important (VI), absolutely important (AI).

A2. Example of a filled questionnaire

See Table 8.

References

Behzadian, M., Otaghsara, S.K., Yazdani, M., Ignatius, J., 2012. A state-of the-art survey of topsis applications. Expert Systems with Applications 39 (17), 13051–13069. https://doi.org/10.1016/j.eswa.2012.05.056.

Bellini, H., Chen, W., Sugiyama, M., Shin, M., Alam, S., Takayama, D., 2016. Virtual and augmented reality: Understanding the race for the next computing platform, the goldman sachs group. The Goldman Sachs Group,Inc.

Blackwood, W.O., 1900. Human factors in the design of tactical display systems for the individual soldier. National Academies Press.

Cannavò, A., De Pace, F., Salaroglio, F., Lamberti, F., 2020. A visual editing tool supporting the production of 3d interactive graphics assets for public exhibitions. International Journal of Human-Computer Studies 102450. https://doi.org/10.1016/j.ijhcs.2020.102450.

Casner, S.M., Gore, B.F., 2010. Measuring and evaluating workload: A primer. NASA Technical Memorandum 216395, 2010.

Chang, D.-Y., 1996. Applications of the extent analysis method on fuzzy AHP. European Journal of Operational Research 95 (3), 649–655. https://doi.org/10.1016/0377-2217(95)00300-2.

Chang, S., Chen, W., 2017. Does visualize industries matter? a technology foresight of global virtual reality and augmented reality industry. 2017 International Conference on Applied System Innovation (ICASI), pp. 382–385. https://doi.org/10.1109/ICASI.2017.7988432.

Eraslan, E., Can, G.F., Atalay, K.D., 2016. Mental workload assessment using a fuzzy multi-criteria method. Tehnicki vjesnik - Technical Gazette 23 (3), 667–674. https://doi.org/10.17559/TV-20140401112509.

Framingham, M., 2016. Worldwide revenues for augmented and virtual reality forecast to reach $162 billion in 2020, according to idc. IDC Report 15.

Freitas, J.D.C., 2018. KAVE - Kinect Cave Design, tools and comparative analysis with other VR technologies. Universidade da Madeira. Ph.D. thesis.

Galecki, A., Burzykowski, T., 2013. Linear Mixed Effects Models Using R. : A Step-by-Step Approach. https://doi.org/10.1007/978-1-4614-3900-4.

Gerry, L., Ens, B., Drogemuller, A., Thomas, B., Billinghurst, M., 2018. Levity: A virtual reality system that responds to cognitive load. Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–6. https://doi.org/10.1145/3170427.3188479.

Gevins, A., Smith, M.E., 2003. Neurophysiological measures of cognitive workload during human-computer interaction. Theoretical Issues in Ergonomics Science 4 (1-2), 113–131.

Gevins, A., Smith, M.E., Leong, H., McEvoy, L., Whitfield, S., Du, R., Rush, G., 1998. Monitoring working memory load during computer-based tasks with eeg pattern recognition methods. Human factors 40 (1), 79–91.

Gumus, A.T., 2009. Evaluation of hazardous waste transportation firms by using a two step fuzzy-AHP and TOPSIS methodology. Expert Systems with Applications 36 (2), 4067–4074. https://doi.org/10.1016/j.eswa.2008.03.013.

Harper, R.P., Cooper, G.E., 1986. Handling qualities and pilot evaluation. Journal of Guidance, Control, and Dynamics 9 (5), 515–529.

Harris, D., Wilson, M., Vine, S., 2019. Development and validation of a simulation workload measure: the simulation task load index (SIM-TLX). Virtual Reality 1–10. https://doi.org/10.1007/s10055-019-00422-9.

Harrison, J., Izzetoglu, K., Ayaz, H., Willems, B., Hah, S., Ahlstrom, U., Woo, H., Shewokis, P.A., Bunce, S.C., Onaral, B., 2014. Cognitive workload and learning assessment during the implementation of a next-generation air traffic control technology using functional near-infrared spectroscopy. IEEE Transactions on Human-Machine Systems 44 (4), 429–440. https://doi.org/10.1109/THMS.2014.2319822.

Hart, S.G., 2006. Nasa-task load index (nasa-tlx); 20 years later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50 (9), 904–908. https://doi.org/10.1177/154193120605000909.

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

Hong, Y., Zeng, X., Bruniaux, P., Chen, Y., Zhang, X., 2018. Development of a new knowledge-based fabric recommendation system by integrating the collaborative design process and multi-criteria decision support. Textile Research Journal 88 (23), 2682–2698. https://doi.org/10.1177/0040517517729383.

Hoover, A., Singh, A., Fishel-Brown, S., Muth, E., 2012. Real-time detection of workload changes using heart rate variability. Biomedical Signal Processing and Control 7 (4), 333–341. https://doi.org/10.1016/j.bspc.2011.07.004.

Hwang, C.-L., Yoon, K., 1981. Methods for multiple attribute decision making. Multiple attribute decision making: methods and applications a state-of-the-art survey. Springer, pp. 58–191. https://doi.org/10.1007/978-3-642-48318-9_3.

Kamaraj, D.C., Dicianno, B.E., Mahajan, H.P., Buhari, A.M., Cooper, R.A., 2016. Stability and workload of the virtual reality–based simulator-2. Archives of physical medicine and rehabilitation 97 (7), 1085–1092.

Kantowitz, B.H., Casper, P.A., 2017. Human workload in aviation. Human Error in Aviation. Routledge, pp. 123–153.

Katicic, J., Häfner, P., Ovtcharova, J., 2015. Methodology for emotional assessment of product design by customers in virtual reality. Presence: Teleoperators and Virtual Environments 24 (1), 62–73. https://doi.org/10.1162/PRES_a_00215.

Kesteven, G., 1946. The coefficient of variation. Nature 158 (4015), 520–521. https://doi.org/10.1038/158520c0.

Kilmer, K.J., Bateman, R., Malzahn, D., 1988. Techniques of subjective assessment: A comparison of the swat and modified cooper-harper scales. Proceedings of the Human Factors Society Annual Meeting 32 (2), 155–159.

Kosch, T., Hassib, M., Buschek, D., Schmidt, A., 2018. Look into my eyes: Using pupil dilation to estimate mental workload for task complexity adaptation. Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. LBW617:1–LBW617:6. https://doi.org/10.1145/3170427.3188643.

Lee, D.K., 2016. Alternatives to p value: confidence interval and effect size. Korean journal of anesthesiology 69 (6), 555. https://doi.org/10.4097/kjae.2016.69.6.555.

Longo, L., 2014. Formalising Human Mental Workload as a Defeasible Computational Concept. Technological University Dublin, Dublin,Ireland. Ph.D. thesis.

Lovie, P., 2005. Coefficient of variation. Encyclopedia of statistics in behavioral science. https://doi.org/10.1002/0470013192.bsa107.

Ma, R., Kaber, D.B., 2006. Presence, workload and performance effects of synthetic environment design factors. International Journal of Human-Computer Studies 64 (6), 541–552. https://doi.org/10.1016/j.ijhcs.2005.12.003.

McKendrick, R., Feest, B., Harwood, A., Falcone, B., 2019. Theories and methods for labeling cognitive workload: Classification and transfer learning. Frontiers in Human Neuroscience 13, 295. https://doi.org/10.3389/fnhum.2019.00295.

Merel, T., 2017. After a mixed year, mobile ar to drive $108 billion vrar market by 2021. Digi-Capital.

Metalis, S., 1991. Heart period as a useful index of pilot workload in commercial transport aircraft. The International Journal of Aviation Psychology 1 (2), 107–116. https://doi.org/10.1207/s15327108ijap0102_2.

Miller, S., 2001. Workload measures. National Advanced Driving Simulator. Iowa City, United States.

Milleville-Pennel, I., Charron, C., 2015. Do mental workload and presence experienced when driving a real car predispose drivers to simulator sickness? An exploratory study. Accident Analysis and Prevention 74, 192–202. https://doi.org/10.1016/j.aap.2014.10.021.

Mouzé-Amady, M., Raufaste, E., Prade, H., Meyer, J.-P., 2013. Fuzzy-tlx: using fuzzy integrals for evaluating human mental workload with nasa-task load index in laboratory and field studies. Ergonomics 56 (5), 752–763. https://doi.org/10.1080/00140139.2013.776702.

Mulder, G., 1986. The concept and measurement of mental effort. Energetics and human information processing. Springer, pp. 175–198.

Novák, V., Perfilieva, I., 1999. Evaluating linguistic expressions and functional fuzzy theories in fuzzy logic. Computing with Words in Information/Intelligent Systems 1. Springer, pp. 383–406. https://doi.org/10.1007/978-3-7908-1873-4_17.

Paas, F., Tuovinen, J.E., Tabbers, H., Van Gerven, P.W., 2003. Cognitive load measurement as a means to advance cognitive load theory. Educational psychologist 38 (1), 63–71.

Parsons, T.D., Cosand, L., Courtney, C., Iyer, A., Rizzo, A.A., 2009. Neurocognitive workload assessment using the virtual reality cognitive performance assessment test. In: Harris, D. (Ed.), Engineering Psychology and Cognitive Ergonomics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 243–252.

Patten, C.J., Kircher, A., Östlund, J., Nilsson, L., Svenson, O., 2006. Driver experience and cognitive workload in different traffic environments. Accident Analysis & Prevention 38 (5), 887–894. https://doi.org/10.1016/j.aap.2006.02.014.

Porssut, T., Chardonnet, J.-R., 2017. Asymetric telecollaboration in virtual reality. 2017 IEEE Virtual Reality (VR). IEEE, pp. 289–290. https://doi.org/10.1109/VR.2017.7892290.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Reid, G.B., Nygren, T.E., 1988. The subjective workload assessment technique: A scaling procedure for measuring mental workload. Advances in psychology, 52. Elsevier, pp. 185–218.

Riley, J.M., Kaber, D.B., 1999. The effects of visual display type and navigational aid on performance, presence, and workload in virtual reality training of telerover navigation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 43 (22), 1251–1255. https://doi.org/10.1177/154193129904302218.

Saaty, R., 1987. The analytic hierarchy process-what it is and how it is used. Mathematical Modelling 9 (3), 161–176. https://doi.org/10.1016/0270-0255(87)90473-8.

Sevinc, V., Berkman, M.I., 2020. Psychometric evaluation of Simulator Sickness Questionnaire and its variants as a measure of cybersickness in consumer virtual environments. Applied Ergonomics 82, 102958. https://doi.org/10.1016/j.apergo.2019.102958.(July 2018)

Shakouri, M., Ikuma, L.H., Aghazadeh, F., Nahmens, I., 2018. Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: The case of highway work zones. International Journal of Industrial Ergonomics 66, 136–145. https://doi.org/10.1016/j.ergon.2018.02.015.

Singh, H., Gupta, M. M., Meitzler, T., Hou, Z.-G., Garg, K. K., Solo, A. M., Zadeh, L. A., 2013. Real-life applications of fuzzy logic. doi:10.1155/2013/581879.

Son, J., Park, S., 2011. Cognitive workload estimation through lateral driving performance. Technical Report. SAE Technical Paper.

Sterman, M., Mann, C., 1995. Concepts and applications of eeg analysis in aviation performance evaluation. Biological psychology 40 (1-2), 115–130.

Tubbs-Cooley, H.L., Mara, C.A., Carle, A.C., Gurses, A.P., 2018. The nasa task load index as a measure of overall workload among neonatal, paediatric and adult intensive care nurses. Intensive and Critical Care Nursing 46, 64–69. https://doi.org/10.1016/j.iccn.2018.01.004.

Wickens, C.D., Hollands, J.G., Banbury, S., Parasuraman, R., 2015. Engineering psychology and human performance. Psychology Press.

Wilson, M.R., Poolton, J.M., Malhotra, N., Ngo, K., Bright, E., Masters, R.S., 2011. Development and validation of a surgical workload measure: the surgery task load index (surg-tlx). World journal of surgery 35 (9), 1961. https://doi.org/10.1007/s00268-011-1141-4.

Yoon, K.P., Hwang, C.-L., 1995. Multiple attribute decision making: an introduction, 104. Sage publications.

Yu, C.-S., 2002. A gp-ahp method for solving group decision-making fuzzy ahp problems. Computers & Operations Research 29 (14), 1969–2001. https://doi.org/10.1016/S0305-0548(01)00068-5.

Zhang, L., Wade, J., Bian, D., Fan, J., Swanson, A., Weitlauf, A., Warren, Z., Sarkar, N., 2017. Cognitive load measurement in a virtual reality-based driving system for autism intervention. IEEE transactions on affective computing 8 (2), 176–189. https://doi.org/10.1109/TAFFC.2016.2582490.

Zhou, R., Chan, A.H., 2017. Using a fuzzy comprehensive evaluation method to determine product usability: A proposed theoretical framework. Work 56 (1), 9–19.