# Weight watchers: NASA-TLX weights revisited

Kai Virtanen, Heikki Mansikka, Helmiina Kontio, and Don Harris



**Citation**
Virtanen, K., Mansikka, H., Kontio, H. and Harris, D., 2021. Weight watchers: NASA-TLX weights revisited. *Theoretical Issues in Ergonomics Science* (In Press)

https://doi.org/10.1080/1463922X.2021.2000667

DOI 10.1080/1463922X.2021.2000667

ISSN 1463-922X
ESSN 1464-536X

Publisher: Taylor & Francis

THEORETICAL ISSUES IN ERGONOMICS SCIENCE
https://doi.org/10.1080/1463922X.2021.2000667

Taylor & Francis
Taylor & Francis Group



# Weight watchers: NASA-TLX weights revisited

Kai Virtanen[a,b], Heikki Mansikka[a,b,c], Helmiina Kontio[a] and Don Harris[d]

[a]Department of Mathematics and Systems Analysis, Aalto University, Helsinki, Finland; [b]Department of Military Technology, Finnish National Defence University, Helsinki, Finland; [c]Insta DefSec, Tampere, Finland; [d]Faculty of Engineering Enviroment and Computing, Coventry University, Coventry, United Kingdom of Great Britain and Northern Ireland

**ABSTRACT**

National Aeronautics and Space Administration Task Load Index (NASA-TLX) is a popular method to evaluate mental workload. NASA-TLX assesses mental workload across six load dimensions. When the dimensions are not assumed to be approximately equally important, they are weighted by conducting a pairwise comparison for every dimension pair, followed by the normalisation of weights reflecting the importance of the dimensions. This original NASA-TLX weighting method creates some challenges that are difficult to identify when the weights are being assigned. First, the original NASA-TLX weighting does not allow directly expressing two or more dimensions as equally important. Second, if pairwise comparisons are conducted consistently, there exists only one possible importance order for the dimensions. Third, with consistently conducted pairwise comparisons, a weight of 0.33 is artificially forced on the most important dimension. Swing and Analytic Hierarchy Process weighting methods for eliciting the weights of the dimensions are proposed as a solution to these challenges. The advantages of applying these methods in NASA-TLX are introduced theoretically and demonstrated empirically using data from virtual air combat simulations. The objective of this paper is to help scholars and practitioners to use NASA-TLX in mental workload assessments such that the discussed weighting issues are avoided.

**ARTICLE HISTORY**
Received 14 June 2021
Accepted 27 October 2021

**KEYWORDS**
Analytic hierarchy process weighting method; mental workload; NASA-TLX; Swing weighting method; virtual air combat simulation

## Relevance to human factors/ergonomics theory

This paper highlights the relevance of load dimension weighting in the NASA-TLX mental workload measurement technique and reveals the fundamental challenges of the original weighting method, as well as those of ignoring the weights altogether. The practice of using NASA-TLX is improved by introducing the use of weighting methods from the field of multi-criteria decision analysis to overcome the challenges associated with the original weighting method. The superiority of these methods over the original NASA-TLX weighting is demonstrated theoretically. Moreover, the same is illustrated in practice using weighting data obtained from qualified F/A-18 fighter pilots who attended virtual air combat simulations.

**CONTACT** Heikki Mansikka heikki.mansikka@aalto.fi

## Introduction

Mental workload (MWL) characterises the balance between a person's limited cognitive resources and the task demands when a desired level of performance is to be maintained (Wilson 2008). As an unbalanced MWL has potential to degrade performance (Mansikka, Virtanen, Harris, et al. 2021; Mansikka et al. 2016; Mansikka, Virtanen, and Harris 2019), the assessment of MWL is at the core of human-system performance studies (Young et al. 2015; Tsang and Vidulich 2006). MWL can be assessed using physiological, behavioural, and/or subjective measures – each having their own strengths and weaknesses (O'Donnell, Eggemeier, and Thomas 1986). Despite criticism, subjective MWL measures are sometimes the only suitable ones in applied field settings, such as simulated air combat (Mansikka et al. 2021a, 2021b; Mansikka, Virtanen, and Harris 2019).

National Aeronautics and Space Administration Task Load Index (NASA-TLX) is a widely used subjective method to assess MWL (Wei, Bolton, and Humphrey 2021; Bogg et al. 2021; Mourali and Lakhal 2021; Mansikka, Virtanen, and Harris 2019; Schreiter et al. 2019; Perry et al. 2008; Hart 2006; Hart and Staveland 1988). NASA-TLX is a multidimensional method, where the MWL is assessed across six dimensions: mental demand (MD), physical demand (PD), temporal demand (TD), performance (OP), effort (EF) and frustration level (FR). When NASA-TLX is administered, the subjects provide two types of information about each dimension: weights and scores. The weights represent the subjective importance of each dimension as a source of MWL in the task of interest, whereas the scores express the subjectively sensed magnitude of MWL with respect to each dimension. To obtain the weights, subjects conduct a pairwise comparison for every dimension pair. In each comparison, the dimension that contributes more to MWL is given a score of one, whereas the other dimension is given zero. Once all 15 pairwise comparisons have been completed, the total score given to each dimension ranges from zero to five. The subjects engage in a task of interest and rate each workload dimension based on their experienced MWL. The range for OP is from 0 (good) to 100 (poor). In all other dimensions the range is from 0 (low) to 100 (high). Rating scales other than 0-100 (e.g. 1-10 or 1-7) have also been used to administer NASA-TLX. However, as this paper discusses specifically the original version of NASA-TLX, the rating scale introduced by Hart and Staveland (1988) is used. A weighted MWL index for every dimension is calculated by multiplying each dimension's score by its weight, i.e. points from the pairwise comparisons. Finally, an overall MWL index is calculated by adding up all weighted MWL scores and by dividing the sum by 15. In other words, the overall MWL index is a weighted sum of the dimension scores, where weights have been normalised to sum of one. The order at which the scores and weights should be assigned to the workload dimensions is problematic, and even the original NASA-TLX user guide (NASA (National Aeronautics and Space Administration), 2021) is ambiguous about the preferred order. If the weights are given first, they are likely to affect the dimensions' expected scores. In contrast, using a reverse order increases the likelihood of the weights being affected by the scores. Overall, to avoid a confusion between the weights and scores, it is probably preferable to assign the weights first.

Despite the popularity and extensive use of NASA-TLX, a closer look at the NASA-TLX weighting method raises three fundamental challenges, which originate partly from the way the NASA-TLX uses ordinal importance information about the workload dimensions

(see, e.g. Chang and Chen 2006). While Nygren (1991) has elegantly discussed most of these challenges, so far no one has provided pragmatic recommendations for weighting the dimensions in an alternative way. In fact, there have been suggestions that the weights should be ignored altogether (Hendy, Hamilton, and Landry 1993; Byers, Bittner, and Hill 1989). One such solution, known as Raw-TLX, uses the arithmetic mean of dimensions' scores as the overall MWL index. While Raw-TLX is an effective and straightforward version of the original NASA-TLX, its use is justified only when the importance of the dimensions is a priori known to be approximately equal in a task of interest. This, however, is not always the case. It is commonly accepted that different load dimensions often contribute differently to MWL and representative weights are needed to express these differences. More recently, fuzzy integrals have been proposed as an alternative method for aggregating the dimensions' scores (Mouzé-Amady et al. 2013). Unfortunately, due to the difficulties related to the interpretation and use of fuzzy integrals, this method adds little, if any, value to the use of NASA-TLX in applied settings (Torra and Narukawa 2006). Finally, there is also an ongoing debate about whether NASA-TLX, and subjective rating scales in general, are scientific methods in the first place (Matthews, Winter, and Hancock 2020; Annett 2002). That dispute, however, is beyond the scope of this paper. Instead, this paper acknowledges that MWL has a central role in human factors and ergonomics research and NASA-TLX is a widely used method for its evaluation. From that perspective, this paper seeks to improve the practice of using NASA-TLX such that the weighting issues inherent in the original version are avoided.

The first challenge is related to a task in which a person assesses two or more dimensions as equally important. With the NASA-TLX weighting method, such preference information concerning the dimensions cannot be expressed directly as the method requires a strict importance order in all the pairwise comparisons, i.e. one of the two dimensions must be stated to be more important than the other. As a result, a rational person is forced to imply a difference in the importance of the dimensions even when s/he feels there is not any. That said, it is still possible to make pairwise comparisons such that equally important dimensions get equal weights. However, this requires inconsistent pairwise comparisons (e.g. $A \succ B$, $B \succ C \succ D \succ E \succ F$, where $\succ$ means the importance order of two dimensions A, B or C) as discussed next.

The second challenge is more severe and considers an inconsistency of the pairwise comparisons. Since one must make 15 comparisons, it is human to end up with inconsistent results simply by accident – especially as the NASA-TLX weighting method does not highlight inconsistent comparisons. On the other hand, if a person wants to make sure two or more dimensions get the same weights, s/he must deliberately make inconsistent pairwise comparisons – something that a person may find confusing. In fact, the only case where the comparisons are consistent is when they imply the perfect preference order of the dimensions, i.e. $A \succ B \succ C \succ D \succ E \succ F$. Then, the resulting weights are $w_A = \frac{5}{15} \approx 0.33$, $w_B = \frac{4}{15} \approx 0.27$, $w_C = \frac{3}{15} \approx 0.20$, $w_D = \frac{2}{15} \approx 0.13,$, $w_E = \frac{1}{15} \approx 0.07$, and $w_F = \frac{0}{15} \approx 0.00$. In other words, should the weights differ from these values, the underlying pairwise comparisons are inconsistent. In short, if the actual weights of the dimensions are not exactly those of the perfect order, a person must choose between accurately expressing his/her preferences of the importance of the dimensions or making consistent comparisons.

The third challenge deals with the limited value range of the weights provided by the method. The possible range of the dimensions' weights is from 0.00 to 0.33 – regardless of the results of the pairwise comparisons. Indeed, it is not possible to obtain a weight higher than 0.33 even if the dimension's importance would warrant that. If the dimensions are in perfect order, the weight of the most important dimension is 0.33 and the weight of the least important dimension is 0.00 as stated above. Thus, when the comparisons are consistent, a weight of 0.00 is always assigned for the least important dimension and the computation of the overall MWL index is based on the other five dimensions with weights more than 0.00. In practice, this essentially makes NASA-TLX a five-dimensional method, which strongly contradicts the original idea of using six load dimensions. The highest weight is also 0.33 when one dimension wins all its pairwise comparisons – regardless of the other pairwise comparisons. Similarly, the lowest weight is 0.00 if one dimension loses all its comparisons regardless of the comparisons between the other dimensions. When this occurs, the weighted index for that dimension is zero and the dimensions' contribution to the overall MWL index is removed – even if a person gives that dimension a MWL score greater than zero. It should be noted that the perfect importance order of the dimensions leads to the weight distribution with the highest variation that can be revealed using the method. In case where the pairwise comparisons result in the minimum weight 0.00 or the maximum weight 0.33, the weight does not necessarily reflect the original preferences of a person in a valid way. For instance, if five dimensions are equally important and one dimension is slightly less important, the least important dimension gets a zero weight, and the other five dimensions get a weight higher than the average weight 0.17. On the other hand, if five dimensions are equally important but one dimension is slightly more important, the most important dimension gets a weight 0.33, and the remaining five dimensions get a weight lower than the average weight 0.17. Furthermore, as mentioned previously, obtaining equal weights for two or more dimensions requires strongly inconsistent pairwise comparisons.

The objective of this paper is to suggest new methods for the weighting of the dimensions to promote the use of the NASA-TLX. The paper does not contribute to the lively debate about how task performance and task induced MWL are associated or disassociated (see, e.g. Mansikka et al. 2019; Galy, Cariou, and Mélan 2012; Oron-Gilad et al. 2008; De Rivecourt et al. 2008; Wickens 2008; Nickel and Nachreiner 2003; Haga, Shinoda, and Kokubun 2002; Miyake 2001; Veltman and Gaillard 1998; Brookings, Wilson, and Swain 1996; Hanckock 1996; Wickens and Yeh 1983; Yeh and Wickens 1988; Vidulich and Wickens 1986; Gopher and Donchin 1986). This has previously been discussed in Theoretical Issues in Ergonomics Science (see, e.g. Mansikka, Virtanen, Harris, et al. 2021; Guastello et al. 2015; Young and Stanton 2007; Kaber and Endsley 2004). Purely from the pragmatic perspective, the human factors/ergonomics community needs a means to describe the importance of the dimensions using weights in a justifiable manner when MWL is assessed in a task where each dimension's importance is not regarded to be approximately similar. While the administration of NASA-TLX with alternative weighting methods described in this paper is more complex than the administration of Raw-TLX, the alternatives are not more complex than the original NASA-TLX weighting method. More importantly, the alternative weighting methods avoid the challenges of the original NASA-TLX weighting method. The challenges are particularly concerning as they are not easy to notice when NASA-TLX is administered. As such, the alternative weighting methods increase the internal validity of NASA-TLX in situations

where the importance of the dimensions is not equal and the dispersion of the distribution is large. To the authors' knowledge, this paper is the first one in which the weighting methods from the field of multi-criteria decision analysis (MCDA) (see, e.g. Eisenführ, Weber, and Langer 2010; Keeney and Raiffa 1993; Keeney 1992) are proposed for the weighting of NASA-TLX dimensions. In addition, the implications of applying different weighting methods in NASA-TLX are demonstrated using data from a virtual air combat simulator exercise. The findings are discussed, and recommendations are made accordingly.

## Alternative weighting methods

Two alternative weighting methods, Analytic Hierarchy Process (AHP) (Saaty 2000) and Swing (von Winterfeldt and Edwards 1985), are introduced. Both AHP and Swing have been widely used, elaborated and validated in the field of MCDA (see, e.g. Lienert, Duygan, and Zheng 2016; Montibeller and von Winterfeldt 2015; Schuwirth, Reichert, and Lienert 2012; Bottomley and Doyle 2001; Pöyhönen and Hämäläinen 2001; Fischer 1995). The original NASA-TLX and MCDA weighting methods differ in the manner preference information regarding the importance of load dimensions is elicited, and what kind of numerical scales are used for answers from the elicitation questions. Despite their differences, they all result in comparable, normalised weights describing the relative importance of the dimensions. Such normalised weights have also been used in many MCDA studies where different weighting methods have been compared and validated in various experimental settings with real-life decision-making problems (see, e.g. Danielson and Ekenberg 2019; Riabacke, Danielson, and Ekenberg 2012; Eisenführ, Weber, and Langer 2010; Hämäläinen and Alaja 2008; Salo and Hämäläinen 1997).

From the many different MCDA weighting methods, AHP was selected as it is similar to, but more versatile than the original NASA-TLX weighting method. In this method, the weights are obtained through pairwise comparisons between decision criteria, i.e. the workload dimensions in the context of NASA-TLX. This should not be confused with the Subjective WORkload Dominance (SWORD) technique, that utilises AHP to conduct relative comparisons between task conditions (see, e.g. Tsang and Vidulich 1994; Vidulich, Ward, and Schueren 1991; Vidulich 1989) or with the Subjective Workload Assessment Technique (SWAT) (Reid and Nygren 1988) which is based on conjoint analysis. When AHP is applied to weight the load dimensions, the relative importance of two dimensions is described on a scale from 1 to 9, where 1 implies indifference and 9 an extreme importance difference. As in the original NASA-TLX weighting method, 15 pairwise comparisons are required when AHP is used. However, the importance difference for each dimension pair is expressed using cardinal importance information in contrast to just the ordinal information used in the NASA-TLX weighting. To illustrate, the associations between verbal judgements and the numerical values of the pairwise comparison of PD and MD dimensions are presented in Table 1. When AHP is used, a similar comparison is conducted for each dimension pair. The values of the importance differences between the dimensions are represented as a pairwise comparison matrix. Then, the weights of the dimensions are obtained by calculating and normalising the principal eigenvector of this matrix.

Nevertheless, the comparison results might not always be consistent. In addition to ordinal inconsistency, AHP comparisons can also be inconsistent in the cardinal sense. For example, consider three dimensions A, B and C being compared using AHP. Assume that

**Table 1.** Verbal descriptions of the pairwise comparison of PD and MD load dimensions, and the corresponding numerical values. Similar comparison is conducted for each dimension pair.

| Verbal description | Numerical value |
|---|---|
| PD and MD are equally important | 1 |
| – | 2 |
| PD is slightly more important than MD | 3 |
| – | 4 |
| PD is more important than MD | 5 |
| – | 6 |
| PD is strongly more important than MD | 7 |
| – | 8 |
| PD is absolutely more important than MD | 9 |

based on the pairwise comparisons A is twice as important as B, B is twice as important as C, and A is three times as important as C. The comparisons are inconsistent on the cardinal scale since according to the first two comparisons A is four times more important than C, although the third comparison states that A is three times more important than C. Inconsistent comparison results can be revealed using the consistency ratio (*CR*). According to a convention widely accepted in the AHP literature, $CR > 0.1$ indicates inconsistent pairwise comparisons, and the comparisons should be considered again. If $CR < 0.1$, the pairw=ise comparisons are sufficiently consistent. However, $CR < 0.1$ does not mean that there would be no inconsistencies in the pairwise comparisons. Instead, it means that while there may be slight inconsistencies, the pairwise comparison are acceptable. The interested reader is referred to Salo and Hämäläinen (1997) for other consistency measures and verbal scales for AHP.

Like AHP, Swing is a widely utilised method in MCDA for determining weights for decision criteria. For this study, Swing was selected mainly for pragmatic reasons; Swing is easy and simple to use and can provide high quality information about the decision maker's preferences (Bottomley and Doyle 2001). In the Swing method, the importance of changing the levels of the decision criteria from the worst to the best level is considered and weights are given accordingly. In the context of the NASA-TLX, the workload dimensions are considered as decision criteria. The most important dimension is given 100 points, the less important dimensions are given between 99 to 0 points, and equally important dimensions are given equal points. As a result, Swing points can be perceived as percentages. Similar to the AHP method, Swing points represent the dimensions' cardinal importance information. Finally, the weights of the six dimensions are obtained by normalising the sum of the points to one.

### *Equally important load dimensions*

The utilisation of the cardinal importance information about the load dimensions enables the AHP and Swing methods to overcome the NASA-TLX weighting method's challenges in several ways. Where the NASA-TLX method has shortcomings in describing equally important dimensions, both the Swing and AHP methods allow such information to be straightforwardly expressed. In the Swing method, the equally important dimensions are simply given equal points. Consequently, and without a need to make inconsistent pairwise comparisons, the dimensions get equal normalised weights. In the AHP method, the equally

important dimensions are both given a value of 1 in their pairwise comparison. It should be noted, though, that the value of 1 in the pairwise comparison does not imply that the weights of these two dimensions are the same in all cases. The equally important dimensions get the same weights only if they get the same values in every pairwise comparison.

### *Consistency of weight assessments*

Considering the inconsistency challenge of the NASA-TXL weighting method previously discussed, it is a lot easier to be consistent when using the AHP or Swing method. Furthermore, there is no need to make inconsistent pairwise comparisons on purpose when two or more dimensions are considered as being equally important. With the NASA-TLX weighting, one much choose between making consistent comparisons and accurately expressing preferences on the importance order of the dimensions. More importantly, with AHP and Swing, inconsistent comparisons are not required to utilise all six workload dimensions, i.e. to obtain non-zero weights for all the dimensions. Although the use of the AHP method requires as many pairwise comparisons as the NASA-TLX weighting method, *CR* indicates whether the comparison results are inconsistent. In this case, important differences can be re-assessed until sufficient consistency is achieved. With Swing, inconsistency is not an issue at all, since the points are given directly for each dimension, rather than through pairwise comparisons.

### *Limited value range of weights*

Both the AHP and Swing methods avoid the challenges of the NASA-TLX weighting method relating to the narrow weight range of 0.00-0.33. When considering the six dimensions, the highest possible weight that a dimension can get with the AHP method is 0.64 and the lowest possible weight is 0.02. The Swing method does not limit the upper or lower value of the weights. In other words, the possible weights obtained with Swing range from 0 to 1. As the AHP and Swing methods allow a wider value range of weights than the NASA-TLX method, the variation in the resulting weight distribution provided by the NASA-TLX weighting method may naturally be smaller than the one determined with the AHP or Swing methods. In addition to the larger variation of the weight distributions, the use of the cardinal importance information in AHP and Swing enables flexible generation of weights. For instance, as discussed earlier, in the case of five equally important dimensions and one dimension slightly less (or more) important, the NASA-TLX weighting yields discrepantly the weight 0.00 for the less (or weight 0.33 for the more) important dimension. When AHP or Swing is used, this slightly less (or more) important dimension can get a weight that is clearly closer to the weights of the equally important dimensions.

### *Raw-TLX*

It should be stressed that according to Hart (2006) and Hart and Staveland (1988), determining an overall MWL index with NASA-TLX requires weighting of the dimensions. Moreover, as Hart and Staveland (1988) state, the diagnostic power of NASA-TLX lies within weights and the magnitude of ratings of the individual dimensions. However, NASA-TLX has also been widely used without considering this weighting, simply by averaging the

scores of the dimensions with the average weights, i.e. $1/6\kappa$. This Raw-TLX is an effective and straightforward approach when the importance of dimensions is roughly equal. Researchers have justified its use with high correlations between the weighted overall MWL and Raw-TLX indices (see, e.g. DiDomenico and Nussbaum 2008; Noyes and Bruneau 2007; Nygren 1991). In the light of the discussed challenges associated with the NASA-TLX weighting method, it is not surprising that high correlations have been found. In fact, Raw-TLX is essentially a special case of the NASA-TLX, when extremely inconsistent pairwise comparisons result in an average weight for every load dimension. A limitation of Raw-TLX is that it assumes that the importances of the dimensions are approximately similar. If these importances are actually different, the use of Raw-TLX may lead to biased conclusions.

As discussed earlier, the more pairwise comparisons the most important dimension wins, the closer its weight gets to the maximum value of 0.33. But the fewer comparisons that dimension wins, the closer its weight gets to the average value of 0.17 – resulting in a decrease in the variation of the weight distribution. Respectively, the more pairwise comparisons the least important dimension loses, the closer its weight gets to the minimum value of 0.00. But the more comparisons that dimension wins, the closer its weight gets to the average value, i.e. 0.17 – resulting also in a smaller variation. Naturally, when the variation of the weight distribution decreases, an overall MWL index determined with these weights approaches the Raw-TLX index in which the average values are used. This does not mean, however, that weighting the dimensions would be insignificant. Instead, it manifests the importance of using an appropriate method for eliciting the weights.

### *Validity of the NASA-TLX weighting method*

The methodological challenges of the original NASA-TLX weighting method are not always realised. It should be stressed, however, that when the challenges do occur, the internal validity of the weighting method becomes questionable. Once the validity becomes an issue, the condition is unlikely to be noticed by either the person assigning the dimensions' weights or by the one administering NASA-TLX. The following demonstration highlights the implications of using the alternative weighting methods that avoid the shortcomings of the original NASA-TLX weighting method.

## NASA-TLX data collection

### *Participants*

Weighting data for NASA-TLX load dimensions were obtained from 20 qualified F/A-18 fighter pilots. NASA-TLX rating data of the dimensions were collected from 16 pilots as four of the pilots were not able to participate in the flying task. The pilots' mean flight experience with F/A-18 was 548.5 flight hours (SD = 218.2). All pilots were male. The pilots were participants in a distributed simulator exercise as part of their normal flight training. The flying tasks were beyond visual range (BVR) air combat missions. For each mission, eight pilots from the group of participants were assigned into two flights of four pilots based on their availability.

### *Weighting procedure*

The purpose of the demonstration was to evaluate how the different weighting methods contributed to the resulting overall MWL index. To avoid the participants from confusing the weights and the scores, and to reduce the likelihood of the weights being affected by the scores, the participants assigned the weights before the flying task. Before the exercise, the participants were briefed about NASA-TLX, including the workload dimensions and the rating scale definitions. Then, written instructions on how to express the importance of the dimensions using the Swing, AHP and original NASA-TLX weighting methods were provided. Participants weighted the TLX workload dimensions using all these methods. To make these weightings, the participants were asked to consider a typical BVR air combat mission of a flight when expressing their preferences with respect to each dimensions' importance. Separate forms were prepared to support Swing, AHP and NASA-TLX weighting procedures. These were simply called just 'First', 'Second' and 'Third' weighting method throughout the material given to the participants.

Each form had detailed instructions on how to enter the workload dimension importance information. The forms were designed to indicate to the participant any unacceptable, missing or unacceptable data entry. For example, an error indication was displayed if both dimensions in a pairwise comparison were given a point in the NASA-TLX pairwise comparison, or more than 100 points were entered for a load dimension in the Swing form. In addition, once a participant had completed the AHP pairwise comparisons, a *CR* value was displayed. If the *CR* was more than 0.1, the participant was informed that the pairwise comparisons were too inconsistent and should be revised. After completing the AHP form, the participants were asked if they managed to achieve an acceptable *CR* on a first try.

The participants were allowed to complete the AHP, Swing and NASA-TLX forms at their own pace and in any desired order. The normalised weights were computed in the background as the data were entered. To eliminate the possibility of the participants adjusting their statements to get similar weights with each weighting method, they were not shown the results.

### *Rating procedure*

The composition of the flights was varied between missions as dictated by the participants' training curriculums and schedules. The flights' task was to fly air combat missions against hostile fighter aircraft. All hostile aircraft were computer programmed entities which followed predefined scripts. Before each mission, the flights received an air tasking order which provided the tactical details for the mission. After studying the air tasking order, the flights conducted standard mission briefings. Once the briefings were completed, the participants entered the flight training devices (FTDs) and the simulation was started.

FTDs are routinely used for basic and advanced fighter pilot training. A total of eight FTDs were used in the demonstration. Three of them had a 135-degree visual display and the rest of the devices were equipped with virtual reality goggles providing a 360-degree visual display. Twenty-six different scenarios were used for the simulations. Hostile aircraft followed the same scripts in every scenario. Each simulation was let to evolve freely and was stopped once there were either no friendly or hostile aircraft left, or nine minutes had

elapsed. For the pilots, each simulation, including the ones with similar starting scenarios, was different, as the hostile aircraft reacted dynamically to friendly aircraft's manoeuvres. Immediately after the simulation was stopped, the participants provided NASA-TLX scores for the six workload dimensions. The original NASA-TLX rating scale 0-100 was used. However, AHP and Swing weighting methods are insensitive to the rating scale used and do not limit the use of other rating scales. Scores were collected from a total of 715 missions.

## Results

Data analysis was conducted with IBM SPSS Statistics software, version 27. In addition to analysing the data with repeated measures ANOVA over the whole sample, it was also analysed for individual pilots. Moreover, data from three pilots (ID4, ID7 and ID16) were examined in more detail to illustrate in greater detail inconsistences and potential advantages when using the different weighting approaches.

The means, maximums and minimums of the normalised weights obtained with the NASA-TLX, AHP and Swing methods for each dimension are summarised in Table 2. Regarding the limited weight range of 0.00 – 0.33 produced by the NASA-TLX weighting method, Table 2 shows that the maximum NASA-TLX weights of MD, TD and OP dimensions were 0.33. The corresponding maximum weights with AHP were higher (0.48, 0.46 and 0.48), but no dimension got a Swing weight higher than 0.31. The maximum weights in Table 2 highlight how the original NASA-TLX restricts weight values more than 0.33 even when that would be preferred.

The dispersion of the weight distribution was analysed with three measures commonly used in the MCDA literature (see, e.g. Pöyhönen and Hämäläinen 2001), i.e. variance, weight ratio and spread of the weights. The weight ratio is the ratio of the weights of the first and the second most important NASA-TLX dimension. The spread of weights is measured by the ratio of the weights of the most important and the second least important dimension. If consistent comparisons are made, the lowest weight with the NASA-TLX weighting method is always zero and hence results in an infinite spread. Therefore, the

**Table 2.** Means (M), maximums (Max) and minimums (Min) of the normalised weights obtained with the NASA-TLX, AHP and Swing weighting methods (N = 20).

| | | M | Max | Min |
|---|---|---|---|---|
| NASA-TLX | MD | 0.27 | 0.33 | 0.07 |
| | PD | 0.01 | 0.13 | 0.00 |
| | TD | 0.18 | 0.33 | 0.07 |
| | OP | 0.21 | 0.33 | 0.07 |
| | FR | 0.13 | 0.27 | 0.00 |
| | EF | 0.2 | 0.27 | 0.13 |
| AHP | MD | 0.3 | 0.48 | 0.06 |
| | PD | 0.03 | 0.07 | 0.02 |
| | TD | 0.19 | 0.46 | 0.05 |
| | OP | 0.21 | 0.48 | 0.06 |
| | FR | 0.12 | 0.32 | 0.04 |
| | EF | 0.15 | 0.3 | 0.09 |
| Swing | MD | 0.22 | 0.31 | 0.14 |
| | PD | 0.08 | 0.12 | 0.03 |
| | TD | 0.18 | 0.31 | 0.09 |
| | OP | 0.18 | 0.26 | 0.10 |
| | FR | 0.16 | 0.24 | 0.09 |
| | EF | 0.18 | 0.23 | 0.12 |

weight of the second least important dimension is used for determining the spread. The mean, maximum and minimum variances, weight ratios and spreads of the weights are summarised in Table 3.

As shown in Table 3, the weights obtained with AHP had the largest mean variance, weight ratio and spread, whereas the corresponding statistics for the Swing weights were lowest. The measures of the dispersion for the NASA-TLX weights were between the values of the Swing and AHP weights. In short, Table 3 shows how the different weighting methods resulted in different weights and how the maximum weight restriction of the NASA-TLX weighting limits the dispersion of the weight distributions.

The means, maximums, minimums and standard deviations of the overall MWL indices obtained with the NASA-TLX, AHP and Swing weights, as well as of the Raw-TLX indices, are summarised in Table 4. AHP and NASA-TLX weighting resulted in higher overall MWL indices in contrast to the indices with Swing and Raw-TLX. This was caused by the AHP and NASA-TLX weights having the biggest variances, weight ratios and spreads of the weight distributions (see Table 3).

The means of the overall indices in Table 4 were analysed using repeated measures ANOVA with Greenhouse-Geisser correction. When analysing the data over all 16 pilots who provided the rating data and the missions they flew, a significant difference between the means obtained with the different weighting methods was found ($F(1.69, 1208.10)=920.11$, $p < 0.001$, partial $^{G}=0.56$). Based on the paired t-tests, there was a statistically significant difference between all means. The results of these pairwise comparisons are summarised in Table 5. The results clearly show that all the weighting methods, including Raw-TLX, lead to significantly different overall indices.

The data were also analysed separately for each pilot with repeated measures ANOVA. Table 6 summarises the results of F-tests which were used to analyse the differences in the means of the overall MWL indices with the different weighting methods. The F-test results revealed that for every pilot, there was a statistically significant difference between the means. Table 7 summarises the results of the pairwise comparisons of these means. Of the 96

**Table 3.** Mean (M), maximum (Max) and minimum (Min) variances, weight ratios and spreads of the normalised weights obtained with the NASA-TLX, AHP and Swing weighting methods (N = 20).

| | | NASA-TLX | AHP | Swing |
|---|---|---|---|---|
| Variance | M | 0.012 | 0.018 | 0.004 |
| | Max | 0.013 | 0.024 | 0.011 |
| | Min | 0.007 | 0.009 | 0.001 |
| Weight ratio | M | 1.270 | 1.720 | 1.160 |
| | Max | 1.670 | 2.510 | 1.430 |
| | Min | 1.000 | 1.030 | 1.000 |
| Spread | M | 4.630 | 7.400 | 2.060 |
| | Max | 5.000 | 10.410 | 3.330 |
| | Min | 2.000 | 2.970 | 1.250 |

**Table 4.** Means (M), maximums (Max), minimums (Min) and standard deviations (SD) of the overall MWL indices over all pilots and missions with the different weighting methods (N = 715).

| Weighting method | M | Max | Min | SD |
|---|---|---|---|---|
| NASA-TLX | 50.32 | 84.67 | 8.67 | 14.04 |
| AHP | 49.85 | 86.31 | 9.58 | 14.27 |
| Swing | 47.02 | 83.79 | 7.56 | 13.15 |
| Raw-TLX | 43.00 | 84.17 | 6.67 | 12.34 |

**Table 5.** Pairwise comparisons between the overall MWL indices obtained with the different weighting methods (N = 715).

| Weighting methods | *t*-value | *p*-value |
|---|---|---|
| NASA-TLX – AHP | 4.23 | <0.001 |
| NASA-TLX – Swing | 27.22 | <0.001 |
| NASA-TLX – Raw-TLX | 42.71 | <0.001 |
| AHP – Swing | 17.92 | <0.001 |
| AHP – Raw-TLX | 30.43 | <0.001 |
| Swing – Raw-TLX | 32.20 | <0.001 |

**Table 6.** Test statistics of F-tests on the means of the overall MWL indices with the different weighting methods. N denotes the number of missions flown by each pilot.

| ID | Df | *F*-value | *p*-value | η2 | N |
|---|---|---|---|---|---|
| 1 | 1.12 | 40.97 | <0.001 | 0.47 | 47 |
| 2 | 1.17 | 19.98 | <0.001 | 0.35 | 39 |
| 3 | 1.93 | 227.21 | <0.001 | 0.84 | 43 |
| 4 | 1.13 | 120.69 | <0.001 | 0.72 | 47 |
| 5 | 1.28 | 177.13 | <0.001 | 0.79 | 47 |
| 6 | 1.21 | 97.34 | <0.001 | 0.7 | 43 |
| 7 | 1.20 | 139.17 | <0.001 | 0.77 | 43 |
| 8 | 1.06 | 143.84 | <0.001 | 0.76 | 47 |
| 9 | 1.25 | 70.07 | <0.001 | 0.6 | 47 |
| 10 | 1.07 | 34.05 | <0.001 | 0.43 | 47 |
| 11 | 1.11 | 98.04 | <0.001 | 0.68 | 47 |
| 12 | 1.25 | 331.18 | <0.001 | 0.89 | 43 |
| 13 | 1.45 | 155.13 | <0.001 | 0.79 | 43 |
| 14 | 1.69 | 66.73 | <0.001 | 0.61 | 43 |
| 15 | 1.69 | 83.75 | <0.001 | 0.65 | 47 |
| 16 | 1.22 | 134.00 | <0.001 | 0.77 | 42 |

conducted pairwise comparisons, 90 pairs indicated a significant difference. This demonstrates how the different weighting methods caused different overall MWL indices for the individual pilots.

Dimension scores from three pilots (ID4, ID7 and ID16) were analysed in more detail to explain the rationales behind the differences in workload magnitudes and orders of the overall MWL indices provided by the alternative weighting methods. While this analysis is not statistical in nature, it extends and explains the earlier theoretical discussion in a transparent, traceable and understandable manner by explicitly highlighting the internal validity issues when using the original NASA-TLX weighting method in real-life MWL assessments. In Figure 1a, the data containing the mean scores and the dimension weights for pilot ID4 is shown as an example to demonstrate a scenario where the NASA-TLX weights produced the highest mean index and the AHP weights the second highest for a single pilot. The NASA-TLX weight of the most important dimension (MD) is 0.33, and the corresponding AHP weight is higher than that. This result highlights the unnatural upper bound of the original NASA-TLX weighting. Figure 1b describes the mean scores and the dimension weights for pilot ID7. Moreover, it demonstrates a situation where there is no significant difference between the overall MWL indices provided by the NASA-TLX and AHP weighting methods. In Figure 1c, the data including the mean scores and weights of the workload dimensions for pilot ID16 is provided as an example to demonstrate a case where the highest mean of the overall MWL index is produced by the AHP weights and the second highest by the NASA-TLX weights for a single pilot. The dimension with the highest weight (OP) also

**Table 7.** Pairwise comparisons' statistical significance of the means obtained with the different weights. Bolded p-values highlight the pairwise comparisons without a significant difference ($p > 0.05$) in the means.

| ID | NASA-TLX - AHP | NASA-TLX - Swing | NASA-TLX - Raw-TLX | AHP- Swing | AHP- Raw-TLX | Swing Raw-TLX |
|---|---|---|---|---|---|---|
| 1 | <0.05 | <0.001 | <0.001 | <0.001 | <0.001 | <0.01 |
| 2 | <0.001 | <0.001 | <0.001 | **>0.05** | <0.05 | <0.001 |
| 3 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 4 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 5 | **>0.05** | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 6 | **>0.05** | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 7 | **>0.05** | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 8 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 9 | <0.01 | <0.01 | <0.001 | <0.001 | <0.001 | <0.001 |
| 10 | <0.001 | <0.01 | <0.001 | <0.001 | **>0.05** | <0.001 |
| 11 | <0.01 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 12 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 13 | **>0.05** | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 14 | <0.01 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| 15 | <0.001 | <0.001 | <0.001 | <0.01 | <0.001 | <0.001 |
| 16 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |

has the highest mean score. The highest NASA-TLX weight is again limited to 0.33, whereas the AHP weight is higher than that.

When the NASA-TLX weighting was used, all pilots assigned a zero weight for the least important load dimension. This meant that with the NASA-TLX weighting method, the overall MWL indices were essentially calculated based on five, not six dimensions. To examine the relevance of this effectively excluded dimension, its proportion in the overall MWL index with the Swing weights was determined for each pilot. The proportion is of the form: Proportion = ((Wi * Mean score of dimension i)/Mean of overall MWL index with Swing) * 100%, where i refers to a dimension whose NASA-TLX weight was zero. Here, the Swing weight of the dimension i denoted by Wi was multiplied by the mean score of the same dimension i. The result was divided by the overall MWL index obtained with the Swing weights and multiplied by 100%. These proportions are presented in Table 8.

In the demonstration, the contribution of the dimension excluded by the NASA-TLX weighting varied from 0.08% to 13.04% of the overall index when using the Swing weights. Over 10% of the mean overall indices of pilots ID1 and ID13 originated from the dimension assigned with a zero NASA-TLX weight.

## Discussion

The weights assigned to six load dimensions are essential to the use of NASA-TLX. Unfortunately, there are limitations associated with the original weighting method used in NASA-TLX. The strict importance order required in all pairwise comparisons does not allow directly expressing two or more dimensions as equally important. In addition, weight values representing exactly the perfect importance order of the dimensions are the only ones that can be obtained without inconsistent comparisons being made. Thus, one must choose between accurately expressing preferences on the importance order and making consistent comparisons. Moreover, should the person be consistent, only five dimensions are considered when calculating the overall workload index, which contradicts the premise

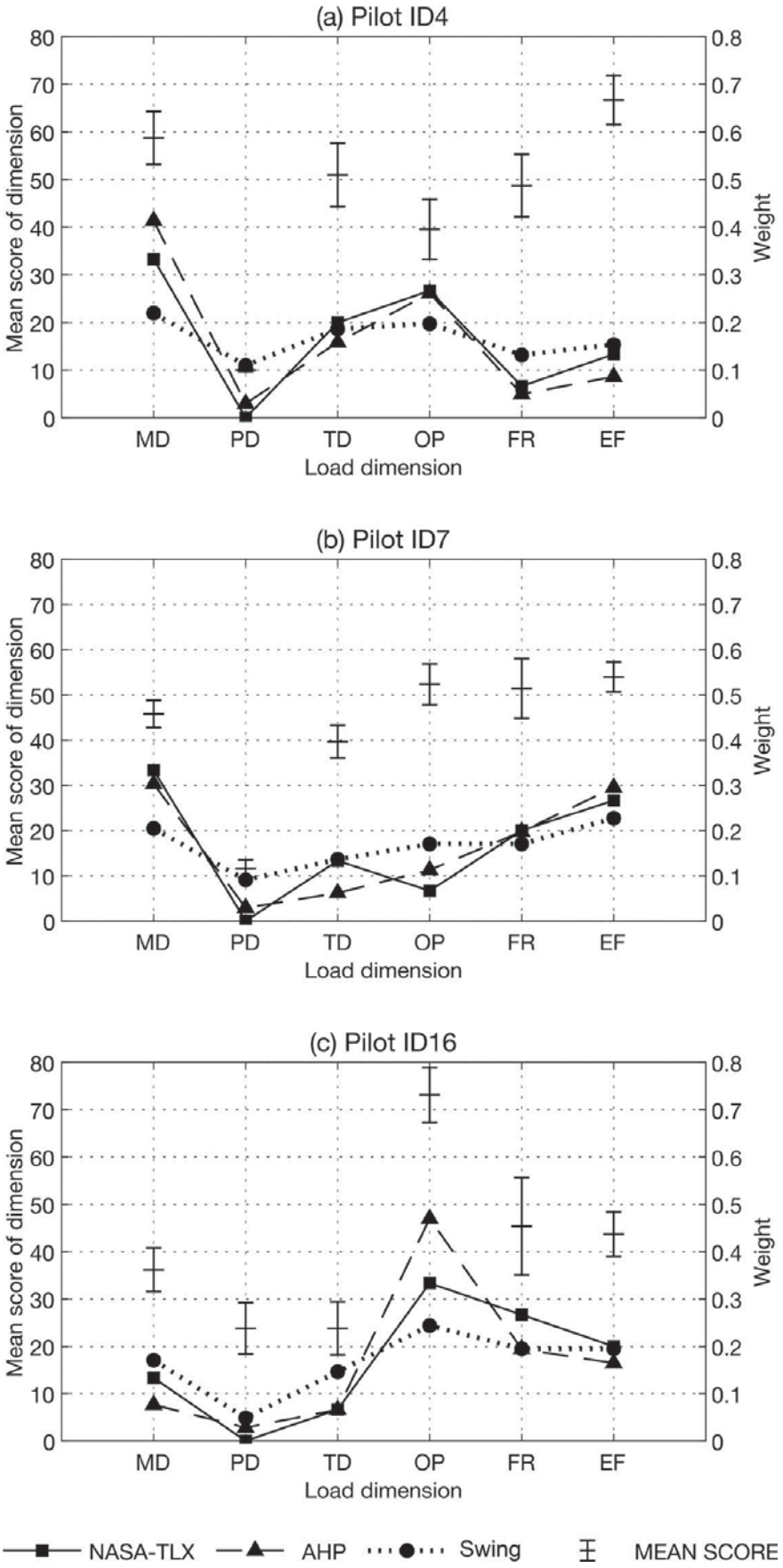


**Figure 1.** Means of the dimension scores with 95% confidence intervals and the dimension weights for pilot (a) ID4 (score N = 47), (b) ID7 (score N = 43) and (c) ID16 (score N = 42). N denotes the number of missions flown by each pilot.

**Table 8.** Proportion of the dimension with a zero NASA-TLX weight in the overall MWL index when using the Swing weights.

| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Proportion (%) | 13.04 | 1.77 | 0.08 | 2.29 | 0.67 | 2.4 | 2.3 | 0.56 |
| ID | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Proportion (%) | 1.54 | 0.63 | 2.04 | 2.29 | 11.1 | 1.15 | 2.73 | 2.52 |

of using the six load dimensions. Finally, the NASA-TLX weighting method limits the dimensions' weights from 0.00 to 0.33 regardless of the results of the pairwise comparisons. As the results of the demonstration highlighted, these limitations induce issues associated with the use of NASA-TLX in a real-life setting.

## *Load dimension weights*

### *Limited value range of weights*

The upper bound limitation of the original NASA-TLX weighting method was evident in the demonstration, as 17 pilots obtained the highest possible weight 0.33 for some dimension. With AHP, 16 of these 17 pilots got weights higher than 0.33, ranging from 0.35 to 0.48. On the other hand, the corresponding Swing weights were lower, ranging from 0.20 to 0.29. The weights of the most important dimension with the AHP approach ranged from 0.29 to 0.48. In most cases, the maximum weight was clearly higher than the weight of any other dimension, whereas the differences between the weights of the less important dimensions were less pronounced. This type of weight distribution thus differs from that found as a product of the perfect order of the NASA-TLX. The maximum weights with the Swing method were closer to the average weight 0.17, ranging from 0.20 to 0.31. With Swing, the differences between the weights of the dimensions were smaller than with the other two methods. This result has also been observed in the context of MCDA where Swing yielded considerably smaller mean weight ratio and spread of weights than AHP (Pöyhönen and Hämäläinen 2001).

No dimension got a zero weight when Swing or AHP was used. Thus, all pilots considered every dimension to have at least some importance as a source of MWL in the given task. This exemplifies the weakness of the NASA-TLX weighting method which by default assigns the least important dimension a zero weight, thus eliminating it completely. Moreover, after a zero weight has been assigned, NASA-TLX still expects the pilot to rate this dimension. This limitation was highlighted in the demonstration, where 18 pilots out of 20 gave a zero weight for PD and the remaining two gave a zero weight for FR when using the NASA-TLX weighting. For instance, a zero PD weight was assigned by pilots ID4, ID7 and ID6, which is pointed out in Figures 1a-1c. Not a single pilot assigned a zero weight to PD or FR when AHP or Swing was used. There were 672 missions where PD and 43 missions where FR were assigned a weight of zero when the NASA-TLX weighting was used. However, in 629 of these missions, PD or FR was rated higher than zero.

### *Equally important load dimensions*

Regardless of the weighting method used, many pilots ended up with equal or almost equal weights for two or more dimensions. Here, the expression 'almost equal' refers to a weight

difference of 0.02 or less. When the NASA-TLX weighting method was used, five out of 20 pilots obtained equal or almost equal weights for at least two dimensions. As noted earlier, this can only be achieved by making inconsistent pairwise comparisons. When AHP or Swing was used, equal or almost equal weights occurred in 19 and 18 cases, respectively. In the pairwise comparisons of AHP, two dimensions were considered to be equally important in 19 out of 300 comparisons, i.e. the ratio of their importance was stated to be 1. However, this does not imply that the weights are equal since the weights depend also on the pairwise comparisons of the other dimensions. For these 19 pairs of the dimensions, the AHP weights were equal or almost equal in 10 out of 19 cases. Similarly, the Swing weights were equal or almost equal in nine out of 19 cases. Based on these observations, the possibility to give two dimensions equal weights seems to be important – but impossible with the NASA-TLX weighting when the pairwise comparisons are consistent.

### *Consistency of weight assessments*

When using the NASA-TLX weighting method, inconsistency clearly is an issue. Out of 20 pilots, 15 were consistent in the pairwise comparisons since the weights they obtained were those of the perfect order. Examples of such pilots are ID4, ID7 and ID16 as the NASA-TLX weights in Figures 1a-1c reflect the perfect order. However, five pilots made inconsistent comparisons. One pilot made five comparisons that conflicted with the order of the resulting weights. For example, EF won the pairwise comparison against FR, but the resulting weights of these two dimensions were equal. The remaining four pilots made three inconsistent comparisons each. For instance, one pilot stated that EF was more important than FR, FR was more important than OP, and OP was more important than EF. However, the resulting weights of these three dimensions were equal. When the dimensions' weights were determined using AHP, 11 pilots out of 20 needed to make the assessment more than once to get a consistency ratio of 0.1 or less. Among the 300 pairwise comparisons performed by 20 pilots, there were a few ordinally inconsistent comparisons, i.e. comparisons that were conflicting with the order of the resulting AHP weights. Finally, it is impossible to be inconsistent when using Swing since the importance points dictating the weights are given directly for each dimension.

### *Overall MWL indices*

As shown in the Results section, when the overall MWL index is determined, the dimensions' weights and the method by which they are elicited with clearly make a difference. As the individual pilots' means of the overall MWL indices obtained with different weighting methods were compared, only six pairs did not have a significant difference. On the other hand, the results of the demonstration implied that the use of Raw-TLX consistently generated the lowest overall MWL index when compared to those provided by the Swing, AHP and NASA-TLX weighting methods. Except for pilot ID10, all the differences were statistically significant. Based on these findings, one should be cautious before ignoring the weighting procedure altogether, even though this approach is used in numerous studies using NASA-TLX. As the Swing weighting method provided the weightings with the smallest variation, the overall MWL indices with Swing were closest to the Raw-TLX indices.

Recall that the AHP and NASA-TLX weightings led to the largest dispersion of weight distributions, which in turn resulted in higher overall MWL indices compared to the indices obtained with Swing and Raw-TLX. The larger the dispersion of the score distribution of

the load dimensions, the more aggravated this phenomenon is. Compared to NASA-TLX, the AHP weighting method resulted in a higher mean of the overall MWL index for four pilots (ID1, ID3, ID6 and ID16), whereas the opposite was true for the other 12 pilots. This might suggest that it makes little difference whether the NASA-TLX or AHP weighting method is used. However, a closer investigation of the examples provided in Figures 1a and 1c leads to a different conclusion. For pilot ID16, the mean of the overall index based on the AHP weights was greater than the one based on the NASA-TLX weights, see Figure 1c. The NASA-TLX weight for the OP dimension had reached its upper limit 0.33, while the AHP weight for that dimension was 0.47. As OP had also the highest mean score of all the dimensions, the limited weight range of the NASA-TLX weighting effectively exaggerated the difference between the overall MWL indices based on the AHP and NASA-TLX weights. Regarding pilot ID4, Figure 1a provides another example of the same issue – this time resulting in the overall MWL using the original NASA-TLX weights being higher than the one based on the AHP weights. It can be argued that for pilot ID4 the NASA-TLX weight for the MD dimension had been artificially limited to the value of 0.33. This then contributed to the increase of the weights of the other dimensions as well. For example, the NASA-TLX weight for the EF dimension was higher than the weight obtained with AHP. The combined effect of the potentially biased NASA-TLX weight and the high mean score on the EF dimension (see Figure 1a) contributed to the mean of the overall MWL index. As a result, the index with the NASA-TLX weights was higher than the one based on the AHP weights.

As noted above, a larger dispersion of dimension score distributions strengthens the impact of weight distributions on the MWL index. In contrast, the weighting method used and the resulting weight distributions have little impact on the overall index if there is little variation within the dimension scores – as was the case with pilot ID7 (see Figure 1b). The score distribution of pilot ID7 was generally uniform, except for a notably lower score for PD. Due to this shape of the score distribution as well as to the low NASA-TLX and AHP weights of the PD dimension, there was no statistically significant difference between the overall indices with AHP and NASA-TLX. However, the Swing weight – and naturally the average weight - of that dimension were higher leading to the lower Swing based and Raw-TLX indices.

Finally, when the NASA-TLX weighting method is used in a consistent fashion, the least important dimension obtains a zero weight. Thus, that dimension does not contribute to the computation of the overall MWL index regardless of the dimension's score. In the demonstration, the contribution of the least important dimension was completely ignored by the original NASA-TLX weighting method – although its contribution was relevant to the overall MWL indices. This highlights the fact that excluding a dimension completely – which is one of the inherent characteristics of the NASA-TLX weighting method – may have a misleading effect on the evaluation of MWL using NASA-TLX.

As the experiment of this paper demonstrated, both Swing and AHP are suitable methods for assigning weights for NASA-TLX's dimensions. Some people may find it more natural to assign weights using AHP's relative pairwise comparisons. In general, however, there are several reasons why Swing is often preferred over AHP in MCDA studies. Compared to relative pairwise comparisons, people usually find it easier to determine weights using an absolute scale. In addition, mapping from the standard measurement scale used in AHP comparisons to resulting normalised weights is nonlinear, which complicates how the levels

of the scale and their verbal expressions can be understood. Moreover, Swing has proven to provide stable weights over time in a time dependent decision environment (Lienert, Duygan, and Zheng 2016) and similar weights in test-retest experiments (Bottomley and Doyle 2001) and may avoid some cognitive biases (Montibeller and von Winterfeldt 2015).

### *Future work*

This paper opens versatile avenues for future research, which should focus on, e.g. behavioural phenomena in using NASA-TLX, alternative weighting approaches, as well as on interpretation and definition of weights, rating scales and scoring of workload dimensions. A pre-task weight given to a dimension is likely to affect person's expectation of that dimension's load in the task of interest. On the other hand, if the weighting of the dimensions is conducted after the load dimensions have been scored, the weights may be affected by the respective scores. This issue could be overcome by introducing a formal interpretation for the weights and scores of the dimensions. Such an interpretation exists in MCDA, but not in the NASA-TLX literature.

In the field of MCDA, a model called an additive value function (Keeney 1992; Keeney and Raiffa 1993) is widely used when decision alternatives are compared to support decision making. This function provides the overall value of an alternative as a weighted sum of partial values representing the alternative's value in respect to each individual criterion. It reveals a natural interpretation for criteria weights which is associated with the ranges of the worst and best measurement levels of the criteria in a set of available decision alternatives. That is, the weight of the given criterion is equal to the change of the overall value of an alternative when the measurement level of this criterion moves from the worst level to the best level. In other words, the weights represent the importance order of the changes of the criterion levels within the alternatives under consideration. Furthermore, the ratio of two weights reflects the trade-off between the partial values of the corresponding criteria. On the other hand, regarding the theoretical rationale of the additive value function, its functional form, i.e. the weighted sum of partial values, is motivated and argued by underlying preference axioms. When a decision maker accepts these axioms, then there exists an additive value function that captures the decision maker's preferences. If the decision maker does not accept the axioms, then more complex value functions, e.g. a multiplicative function, should be used (Dyer and Sarin 1979). Similar axiomatization and interpretation for NASA-TLX would clarify the actual meaning of the weights, rating scales and scores of load dimensions. In addition, they would aid explaining their meaning to a person conducting weighting. Clearly, this kind of insight would promote assessment of dimension weights and scores independently of each other. Moreover, the formal interpretation would reveal MWL measurement situations in which workload dimensions are interdependent in such a way that the weighted sum of scores is not a valid functional form to determine the overall MWL index – another theme that is ignored in the NASA-TLX literature. In this case, a more complex functional form, e.g. multiplicative mentioned above, could be used for aggregating scores into the overall index.

It is acknowledged in the field of MCDA that all weighting methods have potential for cognitive biases (Hämäläinen and Alaja 2008), anchoring effects (Montibeller and von Winterfeldt 2015) and path dependencies (Lahtinen, Hämäläinen, and Jenytin 2020).

Luckily, there is already vast MCDA literature on such behavioural phenomena (Franco et al. 2021), which can be applied to future studies on NASA-TLX. Furthermore, NASA-TLX would benefit from the use of weight elicitation methods enabling incomplete preference information (see, e.g. Harju, Liesiö, and Virtanen 2019; Mattila and Virtanen 2015; Salo and Hämäläinen 2001, 1992; Weber 1985). Such methods would allow, e.g. point values of AHP pairwise comparisons (Salo and Hämäläinen 1995) and Swing points (Mustajoki, Hämäläinen, and Salo 2005) to be expressed as intervals.

When MCDA weighting methods are utilised to support real-life decision making, sensitivity analyses are conducted to reveal whether the rank of decision alternatives changes as weights of criteria and partial values of the alternatives are varied. Future research should also explore the potential of a similar approach when assessing MWL with NASA-TLX.

## Conclusions

In this paper, the implications of applying the original NASA-TLX, Swing and AHP weighting methods in NASA-TLX as well as Raw-TLX were described theoretically and demonstrated in practice using data from 715 air combat simulator missions. The main implications were that there is a clear requirement to be able to (1) use equal weights for some load dimensions; (2) use weights not reflecting the perfect importance order of the dimensions; (3) use weight values beyond the range of 0.00-0.33, as well as to (4) avoid a zero weight for the least important dimension. In addition, the impact of the alternative weighting methods on the resulting overall MWL indices was clearly shown. The importance of using load dimension weightings in NASA-TLX in situations where the use of Raw-TLX was not suitable and superiority of the AHP and Swing weighting methods over the original NASA-TLX weighting method were elaborated from several perspectives. Most importantly, when the validity issues of the NASA-TLX weighting arise, they are likely to go unnoticed by both the person assigning the weights and the person administering NASA-TLX. To conclude, if it is possible that the load dimensions have different contributions to MWL, the alternative weighting methods discussed in this paper should be applied instead of the original NASA-TLX weighting method when NASA-TLX is administered.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Notes on contributors

***Kai Virtanen*** is Professor of Operations Research at Department of Mathematics and System Analysis, Aalto University, Finland and at Department of Military Technology, National Defence University of Finland. His research includes decision analysis, dynamic optimisation, simulation-optimisation and human performance in complex systems.

***Heikki Mansikka*** is retired Air Force fighter pilot (LtCol). He holds a PhD from Coventry University, Aviation Human Factors and a MA from King's College London, Defence Studies. Heikki is a member of the Chartered Institute of Ergonomics and Human Factors. His research includes human performance in air combat. Heikki is currently an Adjunct Professor at Department of Military

Technology, National Defence University of Finland and a Chief Air Combat Scientist at Insta Group, Tampere, Finland.

***Helmiina Kontio*** is a research assistant at the Department of Mathematics and Systems Analysis, Aalto University, Finland. She holds a BSc (Tech.) from Aalto University, Mathematics and Systems Sciences.

***Don Harris*** is Professor of Human Factors at Coventry University. He is a Fellow of the Chartered Institute of Ergonomics and Human Factors and a Chartered Psychologist. Don is heavily involved in the design and development of the next generation of advanced pilot interfaces and the certification of flight decks.

## ORCID

Heikki Mansikka http://orcid.org/0000-0002-7261-6408
Don Harris http://orcid.org/0000-0002-2113-8848

## References

Annett, J. 2002. "Subjective Rating Scales: science or Art?" *Ergonomics* 45 (14): 966–987. doi:10.1080/00140130210166951.

Bogg, A., S. Birrell, M. Bromfield, and A. Parkes. 2021. "Can we Talk? How a Talking Agent Can Improve Human Autonomy Team Performance." *Theoretical Issues in Ergonomics Science* 22 (4): 488–509. Advance online publication. doi:10.1080/14639X.2020.1827080.

Bottomley, P., and J. Doyle. 2001. "A Comparison of Three Weight Elicitation Methods: good, Better, and Best." *Omega* (29) (6): 553–560. doi:10.1016/S0305-0483(01)00044-5.

Brookings, J., G. Wilson, and C. Swain. 1996. "Psychophysiological Responses to Changes in Workload during Simulated Air Traffic Control." *Biological Psychology* 42 (3): 361–377. doi:10.1016/0301-0511(95)05167-8.

Byers, J., A. Bittner, and S. Hill. 1989. "Traditional and Raw Task Load Index (TLX) Correlations: Are Paired Comparisons Necessary?." In *Advances in Industrial Ergonomics and Safety*, edited by A. Mital, 481–485. New York, NY: Taylor & Francis.

Chang, S.-Y., and T.-H. Chen. 2006. "Discriminating Relative Workload Level by Data Envelopment Analysis." *International Journal of Industrial Ergonomics* 36 (9): 773–778. doi:10.1016/j.ergon.2006.06.003.

Danielson, M., and L. Ekenberg. 2019. "An Improvement to Swing Techniques for Elicitation in MCDM Methods." *Knowledge-Based Systems* 168: 70–79. doi:10.1016/j.knosys.2019.01.001.

De Rivecourt, M., M. N. Kuperus, W. J. Post, and L. J. M. Mulder. 2008. "Cardiovascular and Eye Activity Measures as Indices for Momentary Changes in Mental Effort during Simulated Flight." *Ergonomics* 51 (9): 1295–1319. doi:10.1080/00140130802120267.

DiDomenico, A., and M. Nussbaum. 2008. "Interactive Effects of Physical and Mental Workload on Subjective Workload Assessment." *International Journal of Industrial Ergonomics* 38 (11-12): 977–983. doi:10.1177/154193120204601320.

Dyer, J., and R. Sarin. 1979. "Measurable Multiattribute Value Functions." *Operations Research* 27 (4): 810–822. doi:10.1287/opre.27.4.810.

Eisenführ, F., M. Weber, and T. Langer. 2010. *Rational Decision Making*. Berlin, Germany: Springer Verlag.

Fischer, G. 1995. "Range Sensitivity of Attribute Weights in Multiattribute Value Models." *Organizational Behavior and Human Decision Processes* 62 (3): 252–266. doi:10.1006/obhd.1995.1048.

Franco, L., R. Hämäläinen, E. Rouwette, and I. Leppänen. 2021. "Taking Stock of Behavioural or: A Review of Behavioural Studies with an Intervention Focus." *European Journal of Operational Research* 293 (2): 401–418. Advance online publication. doi:10.1016/j.ejor.2020.11.031.

Galy, E., M. Cariou, and C. Mélan. 2012. "What is the Relationship between Mental Workload Factors and Cognitive Load Types?" *International Journal of Psychophysiology : official Journal of the International Organization of Psychophysiology* 83 (3): 269–275. doi:10.1016/j.ijpsycho.2011.09.023.

Gopher, D., and E. Donchin. 1986. "Workload an Examination of the Concept." In *Handbook of Perception and Human Performance*. Vol. II, edited by K. Boff, L. Kaufman and J. Thomas, 41-1–41-49. New York, NY: Wiley.

Guastello, S., A. Shircel, M. Malon, and P. Timm. 2015. "Individual Differences in the Experience of Cognitive Workload." *Theoretical Issues in Ergonomics Science* 16 (1): 20–52. doi:10.1080/1463922X.2013.869371.

Haga, S., H. Shinoda, and M. Kokubun. 2002. "Effects of Task Difficulty and Time-on-Task on Mental Workload." *Japanese Psychological Research* 44 (3): 134–143. doi:10.1111/1468-5884.00016.

Hämäläinen, R., and S. Alaja. 2008. "The Threat of Weighting Biases in Environmental Decision Analysis." *Ecological Economics* 68 (1-2): 556–569. doi:10.1016/j.ecolecon.2008.05.025.

Hanckock, P. 1996. "Effects of Control Order, Augmented Feedback, Input Device and Practice on Tracking Performance and Perceived Workload." *Ergonomics* 39 (9): 1146–1162. doi:10.1080/00140139608964535.

Harju, M., J. Liesiö, and K. Virtanen. 2019. "Spatial Multi-Attribute Decision Analysis: Axiomatic Foundations and Incomplete Preference Information." *European Journal of Operational Research* 275 (1): 167–181. doi:10.1016/j.ejor.2018.11.013.

Hart, S. 2006. "NASA-Task Load Index (NASA-TLX); 20 Years Later." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50 (9): 904–908. doi:10.1177/154193120605000909.

Hart, S., and L. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." *Advances in Psychology* 52: 139–183. doi:10.1016/S0166-4115(08)62386-9.

Hendy, K., K. Hamilton, and L. Landry. 1993. "Measuring Subjective Workload: When is One Scale Better than Many?" *Human Factors: The Journal of the Human Factors and Ergonomics Society* 35 (4): 579–601. doi:10.1177/001872089303500401.

Kaber, D., and M. Endsley. 2004. "The Effects of Level of Automation and Adaptive Automation on Human Performance, Situation Awareness and Workload in a Dynamic Control Task." *Theoretical Issues in Ergonomics Science* 5 (2): 113–153. doi:10.1080/1463922021000054335.

Keeney, R. 1992. *Value Focused Thinking: A Path to Creative Decision Making*. Cambridge, MA: Harvard University Press.

Keeney, R., and H. Raiffa. 1993. *Decision with Multiple Objectives: Preferences and Value Tradeoffs*. New York, NY: Cambridge University Press. doi:10.1017/CBO9781139174084.

Lahtinen, T., R. Hämäläinen, and C. Jenytin. 2020. "On Preference Elicitation Processes Which Mitigate the Accumulation of Biases in Multi-Criteria Decision Analysis." *European Journal of Operational Research* 282 (1): 201–210. doi:10.1016/j.ejor.2019.09.004.

Lienert, J., M. Duygan, and J. Zheng. 2016. "Preference Stability over Time with Multiple Elicitation Methods to Support Wastewater Infrastructure Decision-Making." *European Journal of Operational Research* 253 (3): 746–760. doi:10.1016/j.ejor.2016.03.010.

Mansikka, H., P. Simola, K. Virtanen, D. Harris, and L. Oksama. 2016. "Fighter pilots' heart rate, heart rate variation and performance during instrument approaches." *Ergonomics* 59 (10): 1344–1352. 2016. doi:10.1080/00140139.2015.1136699.

Mansikka, H., K. Virtanen, and D. Harris. 2019. "Comparison of NASA-TLX Scale, Modified Cooper-Harper Scale and Mean Inter-Beat Interval as Measures of Pilot Mental Workload during Simulated Flight Tasks." *Ergonomics* 62 (2): 246–254. doi:10.1080/00140139.2018.1471159.

Mansikka, H., K. Virtanen, and D. Harris. 2019. "Dissociation between Mental Workload, Performance, and Task Awareness in Pilots of High Performance Aircraft." *IEEE Transactions on Human-Machine Systems* 49 (1): 1–9. doi:10.1109/THMS.2018.2874186.

Mansikka, H., K. Virtanen, D. Harris, and M. Jalava. 2021. "Measurement of Team Performance in Air Combat–Have we Been Underperforming?" *Theoretical Issues in Ergonomics Science* 22 (3): 338–359. doi:10.1080/1463922X.2020.1779382.

Mansikka, H., K. Virtanen, D. Harris, and J. Salomaki. 2021a. "Live – Virtual – Constructive Simulation for Testing and Evaluation of Air Combat Tactics, Techniques, and Procedures, Part 1: Assessment Framework." *The Journal of Defense Modeling and Simulation* Advance online publication. 18 (4): 285–293. doi:10.1177/1548512919886375.

Mansikka, H., K. Virtanen, D. Harris, and J. Salomaki. 2021b. "Live – Virtual – Constructive Simulation for Testing and Evaluation of Air Combat Tactics, Techniques, and Procedures, Part 2: Demonstration of the Framework." *The Journal of Defense Modeling and Simulation* Advance online publication. 18 (4): 295–308. doi:10.1177/1548512919886378.

Mansikka, H., K. Virtanen, V. Uggeldahl, and D. Harris. 2021. "Team Situation Awareness Accuracy Measurement Technique for Simulated Air combat-Curvilinear Relationship between Awareness and Performance." *Applied Ergonomics* 96: 103473. doi:10.1016/j.apergo.2021.103473.

Matthews, G., J. De Winter, and P. Hancock. 2020. "What Do Subjective Workload Scales Really Measure? Operational and Representational Solutions to Divergence of Workload Measures." *Theoretical Issues in Ergonomics Science* 21 (4): 369–396. doi:10.1080/1463922X.2018.1547459.

Mattila, V., and K. Virtanen. 2015. "Ranking and Selection for Multiple Performance Measures Using Incomplete Preference Information." *European Journal of Operational Research* 242 (2): 568–579. doi:10.1016/j.ejor.2014.10.028.

Miyake, S. 2001. " Multivariate Workload Evaluation Combining Physiological and Subjective Measures." *International Journal of Psychophysiology: official Journal of the International Organization of Psychophysiology* 40 (3): 233–238. doi:10.1016/S0167-8760(00)00191-4.

Montibeller, G., and D. von Winterfeldt. 2015. "Cognitive and Motivational Biases in Decision and Risk Analysis." *Risk Analysis : An Official Publication of the Society for Risk Analysis* 35 (7): 1230–1251. doi:10.1111/risa.12360.

Mourali, H., and L. Lakhal. 2021. "Mental Workload Measurement, the Case of Stock Market Traders." *Theoretical Issues in Ergonomics Science* 22 (4): 409–433. Advance online publication. doi:10.1080/1463922X.2020.1818866.

Mouzé-Amady, M.,. E. Raufaste, H. Prade, and J.-P. Meyer. 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. doi:10.1080/00140139.2013.776702.

Mustajoki, J., R. Hämäläinen, and A. Salo. 2005. "Decision Support by Interval SMART/SWING - Methods to Incorporate Uncertainty into Multiattribute Analysis." *Decision Sciences* 36 (2): 317–339. doi:10.1111/j.1540-5414.2005.00075.x.

NASA (National Aeronautics and Space Administration). 2021. *NASA Task Load Index (TLX) Paper and Pencil Package v. 1.0*. Human Performance Research Group. California, CA: NASA Ames Research Center. Accessed June 10 2021. https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX_pappen_manual.pdf.

Nickel, P., and F. Nachreiner. 2003. "Sensitivity and Diagnosticity of the 0.1-Hz Component of Heart Rate Variability as an Indicator of Mental Workload." *Human Factors* 45 (4): 575–590. doi:10.1518/hfes.45.4.575.27094.

Noyes, J., and D. Bruneau. 2007. "A Self-Analysis of the NASA-TLX Workload Measure." *Ergonomics* 50 (4): 514–519. doi:10.1080/00140130701235232.

Nygren, T. 1991. "Psychometric Properties of Subjective Workload Measurement Techniques: Implications for Their Use in the Assessment of Perceived Mental Workload." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 33 (1): 17–33. doi:10.1177/001872089103300102.

O'Donnell, R., F. Eggemeier, and F. Thomas. 1986. "Workload Assessment Methodology." In *Handbook of Perception and Human Performance 2 (42)*, edited by K. Boff, L. Kaufman, and J. Thomas, 1–49. New York, NY: Wiley-Interscience.

Oron-Gilad, T., J. Szalma, S. Stafford, and P. Hancock. 2008. "The Workload and Performance Relationship in the Real World: A Study of Police Officers in a Field Shooting Exercise."

*International Journal of Occupational Safety and Ergonomics : JOSE* 14 (2): 119–131. doi:10.1080/10803548.2008.11076757.

Perry, C., M. Sheik-Nainar, N. Segall, R. Ma, and D. Kaber. 2008. "Effects of Physical Workload on Cognitive Task Performance and Situation Awareness." *Theoretical Issues in Ergonomics Science* 9 (2): 95–113. doi:10.1080/14639220600959237.

Pöyhönen, M., and R. Hämäläinen. 2001. "On the Convergence of Multiattribute Weighting Methods." *European Journal of Operational Research* 129 (3): 569–585. doi:10.1016/S0377-2217(99)00467-1.

Reid, G., and T. Nygren. 1988. "The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload." In *Advances in Psychology*, vol. 52, edited by P. Hancock and N. Meshkati, 185–218. Amsterdam, Netherlands: Elseveir. doi:10.1016/S0166-4115(08)62387-0.

Riabacke, M., M. Danielson, and L. Ekenberg. 2012. "State-of-the-Art Prescriptive Criteria Weight Elicitation." *Advances in Decision Sciences* 2012: 1–24. doi:10.1155/2012/276584.

Saaty, T. 2000. *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarcy Process*, volume 6. RWS Publications, Pittsburgh, PA.

Salo, A., and R. Hämäläinen. 1992. "Preference Assessment by Imprecise Ratio Statements." *Operations Research* 40 (6): 1053–1061. doi:10.1287/opre.40.6.1053.

Salo, A., and R. Hämäläinen. 1995. "Preference Programming through Approximate Ratio Comparisons." *European Journal of Operational Research* 82 (3): 458–475. doi:10.1016/0377-2217(93)E0224-L.

Salo, A., and R. Hämäläinen. 1997. "On the Measurement of Preferences in the Analytic Hierarchy Process." *Journal of Multi-Criteria Decision Analysis* 6 (6): 309–319. doi:10.1002/(SICI)1099-1360(199711)6:6<309::AID-MCDA163>3.0.CO;2-2.

Salo, A., and R. Hämäläinen. 2001. "Preference Ratios in Multiattribute Evaluation (PRIME) – Elicitation and Decision Procedures under Incomplete Information." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 31 (6): 533–545. doi:10.1109/3468.983411.

Schreiter, K., S. Müller, R. Luckner, and D. Manzey. 2019. "A Flight Simulator Study of an Energy Control System for Manual Flight." *IEEE Transactions on Human-Machine Systems* 49 (6): 672–683. 2019. doi:10.1109/THMS.2019.2938138.

Schuwirth, N., P. Reichert, and J. Lienert. 2012. "Methodological Aspects of Multi-Criteria Decision Analysis for Policy Support: A Case Study on Pharmaceutical Removal from Hospital Wastewater." *European Journal of Operational Research* 220 (2): 472–483. doi:10.1016/j.ejor.2012.01.055.

Torra, V., and Y. Narukawa. 2006. "The Interpretation of Fuzzy Integrals and Their Application to Fuzzy Systems." *International Journal of Approximate Reasoning* 41 (1): 43–58. doi:10.1016/j.ijar.2005.08.001.

Tsang, P., and M. Vidulich. 1994. "The Roles of Immediacy and Redundancy Relative Subjective Workload Assessment." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 36 (3): 503–513. doi:10.1177/001872089403600307.

Tsang, P., and M. Vidulich. 2006. "Mental Workload and Situation Awareness." In *Handbook of Human Factors and Ergonomics*, edited by G. Salvendy, 243–268. Hobogen, NJ: John Wiley & Sons, Inc. doi:10.1002/0470048204.ch9.

Veltman, J., and A. Gaillard. 1998. "Physiological Workload Reactions to Increasing Levels of Task Difficulty." *Ergonomics* 41 (5): 656–669. doi:10.1080/001401398186829.

Vidulich, M. 1989. "The Use of Judgment Matrices in Subjective Workload Assessment: The Subjective WORkload Dominance (SWORD) Technique." *Proceedings of the Human Factors Society Annual Meeting* 33 (20): 1406–1410. doi:10.1177/154193128903302009.

Vidulich, Michael A., G. Frederic Ward, and James Schueren. 1991. "Using the Subjective Workload Dominance (SWORD) Technique for Projective Workload Assessment." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 33 (6): 677–691. doi:10.1177/001872089103300605.

Vidulich, M., and C. Wickens. 1986. "Causes of Dissociation between Subjective Workload Measures and Performance: Caveats for the Use of Subjective Assessments." *Applied Ergonomics* 17 (4): 291–296. doi:10.1016/0003-6870(86)90132-8.

von Winterfeldt, D., and W. Edwards. 1985. *Decision Analysis and Behavioural Research*. Cambridge, MA: Cambridge University Press.

Weber, M. 1985. "A Method of Multiattribute Decision Making with Incomplete Information." *Management Science* 31 (11): 1365–1371. doi:10.1287/mnsc.31.11.1365.

Wei, J., M. Bolton, and L. Humphrey. 2021. "The Level of Measurement of Trust in Automation." *Theoretical Issues in Ergonomics Science* 22 (3): 274–295. doi:10.1080/1463922X.2020.1766596.

Wickens, C. 2008. "Multiple Resources and Mental Workload." *Human Factors* 50 (3): 449–455. doi:10.1518/001872008X288394.

Wickens, C., and Y.-Y. Yeh. 1983. "The Dissociation between Subjective Workload and Performance: A Multiple Resource Approach." *Proceedings of the Human Factors Society Annual Meeting* 27 (3): 244–248. doi:10.1177/1541931283027003.

Yeh, Y.-Y., and C. Wickens. 1988. "Dissociation of Performacne and Subjective Measures of Mental Workload." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 30 (1): 111–120. doi:10.1177/001872088803000110.

Young, M., K. Brookhuis, C. Wickens, and P. Hancock. 2015. "State of Science: mental Workload in Ergonomics." *Ergonomics* 58 (1): 1–17. 2015. doi:10.1080/00140139.2014.956151.

Young, M., and N. Stanton. 2007. "Miles Away: determining the Extent of Secondary Task Interference on Simulated Driving." *Theoretical Issues in Ergonomics Science* 8 (3): 233–253. doi:10.1080/14639220601129228.