Novel mental workload scale application using fuzzy sets theory

Hanna J. Neroj[1] & Jim Nixon[2]

[1]Think Research Ltd, [2]Centre for Safety and Accident Investigation, Cranfield University

ABSTRACT

A new scale measuring pilot subjective mental workload was developed as part of the Horizon 2020, Future Sky Safety Project: The validity of new measures needs to be established for effective and reliable implementation. This article investigates the validity of this new scale through assessment of psychological distances between linguistic variables on the scale and subsequent transformation from discrete to continuous scale using fuzzy sets theory. Although the new scale has claimed ordinal properties for the items, no evidence exists to support the interval properties of the scale and any subsequent analysis using statistical methods. The results in this article show linear progression of the items on the scale, supporting the order of the items. To establish the interval properties of the scale, the distances between the ordered items were evaluated using fuzzy sets theory using the Fuzzy Logic Designer in MATLAB. Transformation to continuous scale using fuzzy sets may allow the capture of the dispersion of the data around specific points, showing the degree to which each score belongs to different linguistic category of mental workload. Finally, a proposal for incorporating the validated performance measure was presented.

KEYWORDS

Mental workload, fuzzy sets theory, semantic distance, validity

Introduction

Workload has been an important topic in the aviation industry and is a widely used concept in the human factors performance assessment (Bailey and Iqbal, 2008; Wickens et al., 2013; Young et al., 2015). Subjective rating scales are used to assess mental workload, the canonical example being the NASA task load index (NASA-TLX) (Hart and Staveland, 1984). Workload scales that require participants to assign a qualitative description of workload are also available, for example the Bedford Workload scale. These scales have the advantage that workload can be described and communicated. These scales often have ordinal properties but no data to support the interval properties of the scale or indeed the magnitude of the intervals between items. This problem renders the results of any statistical analysis dubious at best. In order to assign appropriate interval scale characteristics, formal analysis must be conducted to establish the distances between scale points needs to be established.

This article addresses this problem. In the recent Horizon 2020, Future Sky Safety programme (Future Sky Safety, 2016), a new scale was proposed to measure pilot workload in the 'Human Performance Envelope' part of this project. This scale currently claims ordinal properties. This research establishes and demonstrates the ordinal *and* interval properties of this scale through an evaluation of the semantic distances between the categories and change in the dimensions' inclusion applying fuzzy sets theory. The idea of applying fuzzy sets theory for mental workload measurement has been established in this area of inquiry (for example Moray et al., 1988). Fuzzy sets may allow for reflecting the magnitude of mental workload sensation due to transformation of the discrete scale to continuous. The output reflects not only a numerical value, but also a degree to which the answer belongs to different categories of workload. This allows to capture the dispersion of the data around a specific point and prevents from choosing the closest category applicable, that may not reflect true sensation of mental workload. Fuzzy sets give as well the possibility to transform the scale to multidimensional. This may allow for less time-consuming procedure for the participant.

Methodology

Participants

Fifty participants with flying experience gave informed consent to participate in the research. The subjects were divided into two groups, 29 subjects with private pilot licenses (PPL) and 21 subjects with airline pilot licenses (ATPL). The participants comprised males and females, with male predominating. The more detailed breakdown by age and gender is presented in Table *1* below.

Table 1: Participants of the study

Experience	Gender	Mean Age (sd)	Count
ATPL	Women	29.6 (2.51)	5
	Men	33.5 (10.06)	15
PPL	Women	25.0 (4.90)	6
	Men	33.2 (9.24)	23
	Prefer not to say	25.0 (n/a)	1

Design

Participants assessed the distances between the scale items using slider-scales presented in the Qualtrics online survey environment. Analysis was conducted using a 2 × 6 mixed ANOVA. The independent variables were experience [PPL, ATPL] and each scale point from the Performance Curve [Relaxed, Focused, Under Pressure, Struggling, Failing, and Lost It] comprising the within-subjects factor. Planned linear contrasts were conducted following the omnibus F-test.

Secondly, the intervals between scale points was interrogated using fuzzy sets theory. The design of the fuzzy logic system was conducted in MATLAB 2021a software. Fuzzy sets theory could be a solution for the lost magnitude of workload when transforming into a unidimensional scale. The principle of fuzzy logic is that the output is not binary, i.e., either 0 or 1, but belongs to the set [0, 1] with 0 being absolute false and 1 being absolute truth. In social sciences it could be translated as the answer is not 0 or 1 (or 'yes' or 'no'), but it lies somewhere in between, e.g., 0.3 (or 'partially yes'). In this study, two input variables were modelled, namely the workload score from the workload scale from the first section of this study, and the performance. The linguistic variables for the performance were applied according to the HPE project (Future Sky Safety, 2016), reflecting three levels [Good, Acceptable, Degraded]. The defuzzification method chosen was a centroid.

Materials and Procedure

The research instrument was an online questionnaire prepared using Qualtrics software. Ethical approval was given for the research from Cranfield University Research Ethics Committee. After giving informed consent, participants were presented with questions about age, gender, and flight experience. Participants then completed a practice question being asked to distribute emotional states from '*Sad*' to '*Joyful*' on a straight line. The experimental task consisted of one question, where participants were presented with six words that relate to how they may feel in the cockpit.

Those words were obtained from the HPE project (Future Sky Safety, 2016), namely: *Relaxed, Focused*, *Under Pressure*, *Struggling*, *Failing*, and *Lost It*.

Figure 1: Scaling of a new workload scale task presented in Qualtrics

The participants were asked to distribute the scale labels on a line from the lowest workload on the left and the highest on the right. This was done using a slider (Figure 1). The participants were advised to treat the question as if they were marking all those words on a common scale. This configuration was not available to the researchers due to the limitations of the software. The lowest possible value to choose was 0.0 and the highest was 100.0 in decimal intervals. The numbers on the axis were not visible to the participants to reduce the possibility of following the tendency to choose equal intervals. Moreover, they were asked to consider the distances between each word based on their personal experience in the cockpit, e.g., how long does the 'Relaxed' feeling last in comparison to 'Focused' and each other state. At the end of the survey, the pilots were asked whether the words were clear or unclear using a 5-point Likert scale, i.e., from 'Extremely clear' to 'Extremely unclear'. Additionally, they could share any other comments about the experiment. The survey was distributed online amongst pilots. Each of the participants completed the survey anonymously on their personal devices, firstly providing informed consent.

Results

To assess the distances between the scale points, descriptive statistics analysis was held. Table 2 presents the mean values for the position of each scale point along with the standard deviation.

Table 2: Scale points descriptive statistics

Scale Point	Mean	SD
Relaxed	14.01	24.80
Focused	34.03	24.93
Under Pressure	55.93	21.39
Struggling	63.65	19.11
Failing	75.95	23.02
Lost It	88.73	25.25

Data presented in Table 2 shows the mean value (Mean) with standard deviation (SD) of each scale point calculated from all 50 participants. The lowest point on the scale is 'Relaxed' (M=14.01, SD=24.80), and the highest is 'Lost it' (M=88.73, SD=25.25). It is visible that the standard deviation is similar for each of the points.

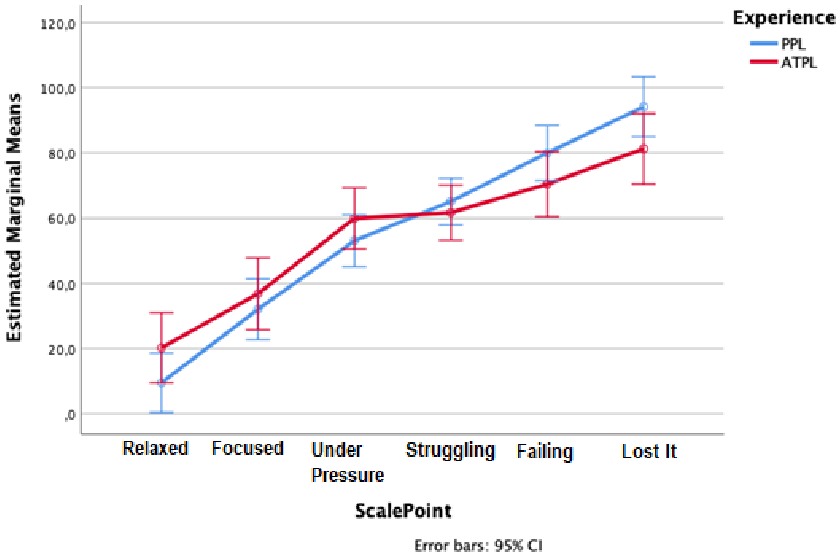

Figure 2: Scale points values based on the experience

Figure 2 shows the distribution of the mean values of the scale points on the graph based on the responses of two groups of pilots: ATPL and PPL. The data presented in Table 3 shows unequal distances between the scale points. To assess the trend of the points of the scale, a linear contrasts calculation was done using IBM SPSS software. Due significant sphericity, the Greenhouse-Geisser corrected df was used. The within-subjects effects test showed significant difference between each of the scale points ($F_{2.38,114.27}=76.25$, $p<0.001$, $\eta_p^2=0.61$), i.e., the scale points rise linearly from 'Relaxed' to 'Lost it'.

The analysis was then continued with a between-subjects effects test with the factor: Experience, to assess the level of interaction between flying experience and the values of the scale points. The test showed that there is no significant interaction between scale points values and level of experience ($F_{2.38,114.27}=2.40$, $p=0.086$, $\eta_p^2=0.05$). Therefore, no follow-up statistical analysis was conducted in that matter. However, when looking at a plot of scale points differentiated by the experience (Figure 13), there are some visible differences worth further discussion. The graph presented in Figure 2 for the PPL pilots differs from the ATPL pilots. The reason for the p-value approaching significance could be a non-significant difference between scale points 'Under Pressure' and 'Struggling' for the ATPL pilots group ($F_{1,48}=0.14$, $p=0.707$, $\eta_p^2=0.003$). The result for the same contrast for the PPL pilots showed a significant difference ($F_{1,48}=9.23$, $p=0.004$, $\eta_p^2=0.16$).

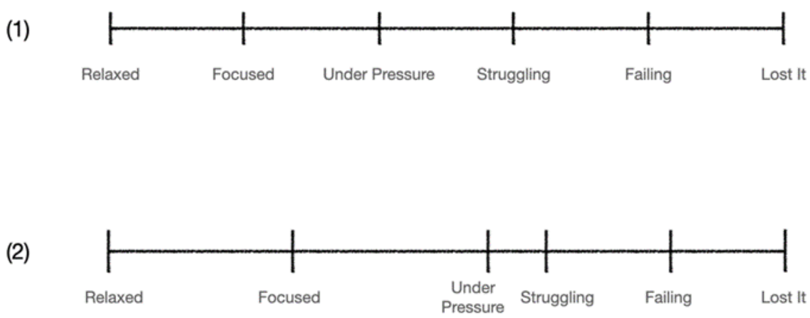

Figure 3: Novel workload scale based on Performance Curve before and after scaling

Figure 3 presents a visualisation of how the novel workload scale could be presented to research participants after conducted scaling activity (2) in comparison with assumed equal distances (1). Presentation (2) may be beneficial because it could give the participant the sense of the range of each of the categories.

The scale after scaling activity reflects more accurately the range of each workload category, however the magnitude of each of those experiences may be lost when using a discrete scale. By applying fuzzy sets theory, the scale can be treated as continuous, allowing for multiple options for the participants. Moreover, the chosen point on the scale will belong in varying degree to different categories. This contributes to higher scale sensitivity. The membership function will inform of the degree with which a chosen point belongs to different categories. The scale was established by linking linguistic variables (from ‘Relaxed’ to ‘Lost It’) with the triangular fuzzy numbers derived from statistical analysis, i.e., the mean values and standard deviations.

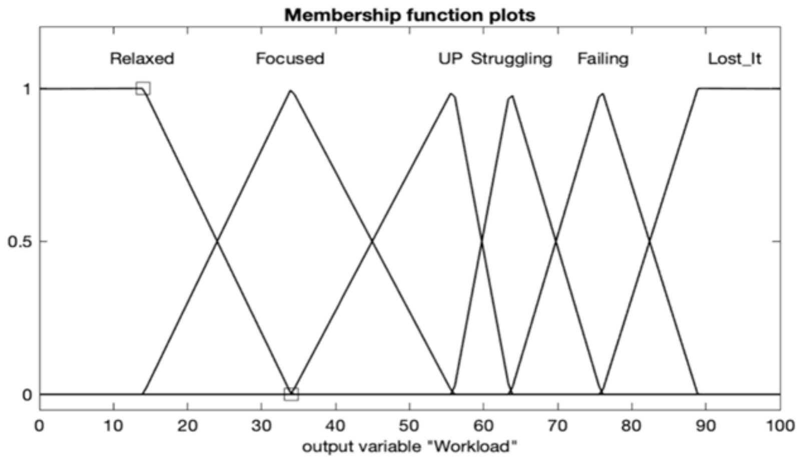

Figure 4: Membership function for the novel workload scale

In order to re-scale the workload, a Single Input Single Output (SISO) Mamdani fuzzy system is considered. Defuzzification is done by the Centre of Gravity (centroid). Since the fuzzy Likert scale allows the partial agreement to a scale point, in comparison with the traditional Likert scale, the responses can be obtained to a decimal place, following a similar procedure as proposed by Li (2013). In this technique, the survey respondent can choose any point on the scale, including space in between the scale points. Since the partial agreement can be detected, the dispersion of the data around a specific point can be captured and can give some valuable results (Li, 2013). If the participants were presented with a traditional Likert scale, and they would not agree with the statement that they felt, for example, ‘Under Pressure’, they would choose the closest possible option. In that case, the data dispersion would not be captured, and it could be concluded from the data that there is full consensus between the participants in terms of their workload.

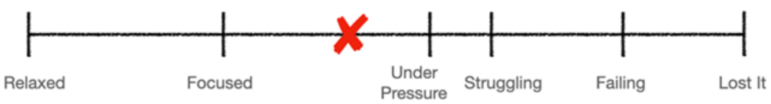

Figure 5: Example answer by the participant of their subjective mental workload

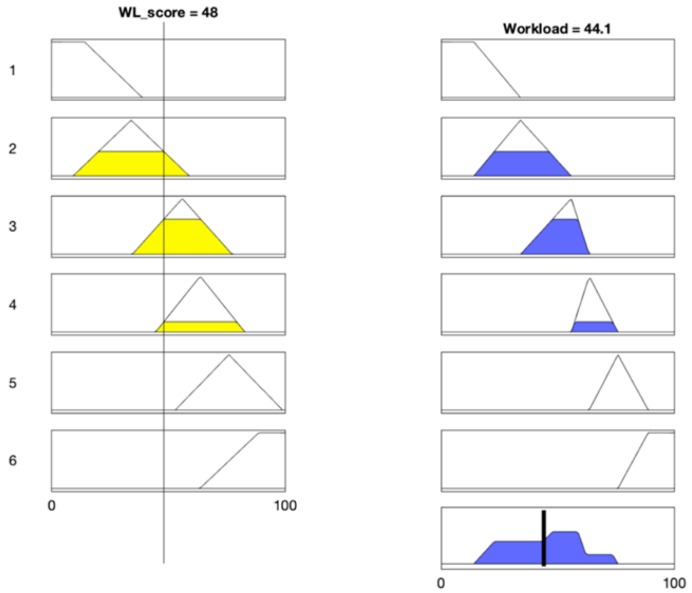

Figure 6: Input and output modelling for input WL_score=48, Fuzzy Logic Designer view

Figure 5 an example of mental workload chosen by the respondent on the linguistic scale is presented. The chosen linguistic value is between points ‘Focused’ and ‘Under Pressure’. Figure 6 presents modelled in MATLAB input and output functions, where WL_score=48 corresponds to the place on a scale chosen by the participant. The number from 1 to 6 correspond with rules assigned according to the fuzzy sets theory, which describe the values assigned per each of the workload categories, from 1 corresponding to ‘Relaxed’ category, and 6 corresponding to ‘Lost It’ category. The figures coloured in yellow are the visualisation of the input value, divided per membership of participant’s input to each rule, and the blue figures are the visualisation of the membership degree of the output separated per each rule. The final output is the blue figure on the right side of the Figure 6 at the bottom.. In this case, the value 48 could be interpreted either as ‘Slightly Focused’, ‘Very Under Pressure’, or ‘A Little Struggling’. The bottom figure on the right side of the graph presents the area of the membership function (Figure 6) that reflects the degree of adhesion to each of the three scale points. Afterwards, using the centroid method, the overall workload score is obtained. If the score would be obtained with a traditional Likert type scale, this magnitude would be lost, and the respondent would have to choose one of the three points on a scale.

Discussion and conclusions

The analyses conducted support the correctness of the order of linguistic variables on the new workload scale that was assumed by the Future Sky Safety project #6. The variables follow a trend from lower to higher, i.e., from ‘Relaxed’ to ‘Lost it’. The contrast analysis showed linear progression of the scale categories and the scaling activity assured of the exact distances between scale categories. However, when looking at the data differentiating the values by experience, there is a visible difference between ATPL pilots’ and PPL pilots’ responses to scale points ‘Under Pressure’ and ‘Struggling’. Due to that fact, it was decided to do further contrasts analysis for each group of the pilots, and the results showed no significant difference between above mentioned two scale points for the ATPL pilots ($p=0.707$, $\eta_p^2=0.003$).The reason for such a result may be the psychological meaning of the linguistic variables. Whereas ‘Focused’ has generally a positive tone, ‘Under Pressure’ may be understood in twofold ways. It could mean that when airline pilots were feeling under any sort of pressure, they were already feeling like they are struggling, therefore it could have a negative meaning for them. Moreover, this finding is aligned with the theory of uneven distribution of workload in the cockpit, where pilots have peaks of high workload during situations that they do not encounter on a daily basis, such as difficult landing in adverse weather conditions or emergencies. Due to the mid-point semantic clustering, further research should be undertaken regarding modification of those scale points, to investigate possibility of greater differentiation. Furthermore, sensitivity of the scale in terms of capturing nuances of the responses around middle points should be investigated, given the likelihood of the bell curve responses of the pilots.

The original scale developed as part of Future Sky Safety Project comprised *two* dimensions: mental workload rating and performance. The fuzzy sets theory could be applied as well to incorporate those two inputs in the overall score, transforming the scale to multidimensional. The linguistic variable plot was modelled using the mean and standard deviation values (Table 2) that were the results of the scaling activity performed by pilots through the survey. The membership functions have a trapezoidal and triangular shape, which was chosen in accordance with previous research (Liou and Wang, 1994; Chen, 1996; Yong, 2011). The performance was modelled following Future Sky Safety Project #6, where 'Good' level of performance corresponded with values of linguistic variables 'Relaxed' and 'Focused', 'Acceptable' level of performance corresponded with values of linguistic variables 'Under Pressure' and 'Struggling', and 'Degraded' level of performance corresponded with values of linguistic variables 'Failing' and 'Lost It'. However, this requires further research on the relationship between workload and performance.

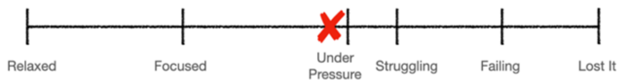

Figure 7: Example of the participant's response

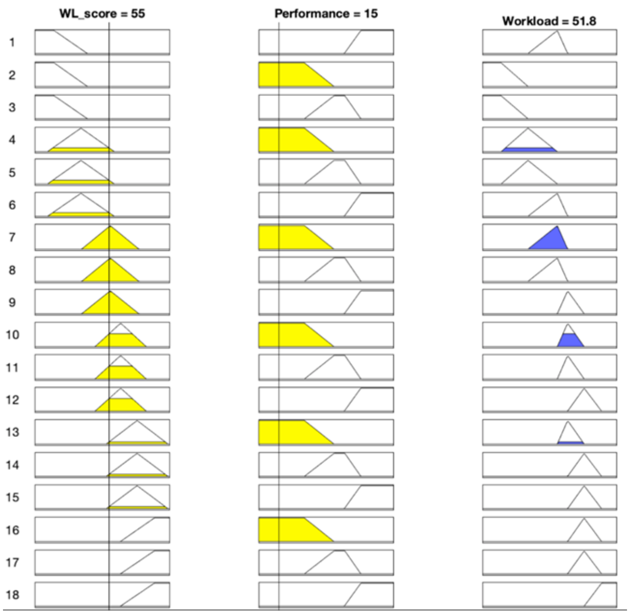

Figure 8: Workload score between 'Focused' and 'Under Pressure' with 'Good' performance

Figure 7 shows the possible response of the participant to rate their workload level. Additionally, the participant's performance should be evaluated, either as subjective rating, or measured by objective means. In this example, the model assumes that the performance is assessed subjectively by the participant by choosing from three possible categories, i.e. 'Good', 'Acceptable', and 'Degraded'.

The WL_score is modelled as a continuous variable, i.e., the software allows for a WL_score input with an accuracy of one decimal place. The main advantage of this approach is that the participant will have the flexibility of choosing values in between specific scale points, which will give higher reliability of the output value. For example, if the subject feels 'Very Focused' or 'Almost Under Pressure' (Figure 7), they will have the ability to mark that point on a scale instead of choosing between 'Focused' and 'Under Pressure' like on a typical Likert scale.

Data analysis using fuzzy logic gives the ability of not losing the magnitude of subjective workload feeling as well as gives to the possibility to add another dimension, such as performance, without a time-consuming procedure for the participant. The fuzzy sets give the ability to interpret individual scores looking also at the membership degree to each category, which gives a more detailed indication of what the subject may be going through.

This research has shown that there is potential in using this scale. However, there are some limitations that require further research, such as the distance between 'Under Pressure' and 'Struggling' scale points, or consideration of another scale point addition to reflect underload conditions, based on literature review (e.g. Kantowitz, 2000; Montani et al., 2020). In order to translate the linguistic result of mental workload rating into a numerical value, the re-scaling using fuzzy logic can be applied. That way the dispersion of respondents' answers amongst specific scale points can be captured as well as a more accurate representation of the perceived mental workload of individuals.

References

Bailey, B. P. & Iqbal, S. T. (2008). 'Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management', ACM Transactions on Computer-Human Interaction, 14(4). doi: 10.1145/1314683.1314689.

Chen, S. M. (1996) 'New methods for subjective mental workload assessment and fuzzy risk analysis', *Cybernetics and Systems*, 27(5), pp. 449–472. doi: 10.1080/019697296126417.

Future Sky Safety (2016) *Test report preliminary testing with system pilots cognitive task analysis*.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklade, A. L., Christ, R. E. (1992) 'Comparison of four subjective workload rating scales', *Human Factors*, 34(4), pp. 429–439. doi: 10.1177/001872089203400405.

Kantowitz, B. H. (2000). Attention and Mental Workload. Proceedings of the IEA 2000/HFES 2000 Congress.

Li, Q. (2013) 'A novel Likert scale based on fuzzy sets theory', *Expert Systems with Applications*, 40(5), pp. 1609–1618. Doi: 10.1016/j.eswa.2012.09.015.

Liou, T. S. and Wang, M. J. J. (1994) 'Subjective assessment of mental workload - A fuzzy linguistic multi-criteria approach', *Fuzzy Sets and Systems*, 62(2), pp. 155–165. Doi: 10.1016/0165-0114(94)90055-8.

Miller, S. (2001) *Literature Review Workload Measures*. Doi: 10.1201/9780429019579.

Montani, F., Vandenderghe, C., Khedhaouria, A., Courcy, F. (2020). Examining the inverted U-shaped relationship between workload and innovative work behavior: The role of work engagement and mindfulness. *Human Relations*, 73(1), pp. 59-93.

Moray, N., Eisen, P., Money, L., Turksen, I. B. (1988) 'Fuzzy Analysis of Skill and Rule-Based Mental Workload', *Advances in Psychology*, 52, pp. 289-304.

Wickens, C. D., Hollands, J. G., Banbury, S., Parasuraman, R. (2013) *Engineering Psychology and Human Performance*. 4th edition. Pearson Education, Inc.

Yong, D. (2011) 'Subjective mental workload assessment based on generalized fuzzy numbers', *Cybernetics and Systems*, 42(4), pp. 246–263. doi: 10.1080/01969722.2011.583596.

Young, M. S., Brookhuis, K. A., Wickens, C. D., Hancock, P. A. (2015) 'State of science: mental workload in ergonomics', *Ergonomics*. Taylor & Francis, pp. 1–17. doi: 10.1080/00140139.2014.956151.