Measuring Decision Accuracy and Confidence of Mock Air Defence Operators

Jade E. Adams-White[1], Jacqueline M. Wheatcroft[1], and Michael Jump[1]

[1]The University of Liverpool, UK

Author note

Jade E. Adams-White, Institute of Psychology, Health & Society. Department of Psychological Sciences; Jacqueline M. Wheatcroft, Institute of Psychology, Health & Society, Department of Psychological Sciences; Michael Jump, School of Engineering.

Correspondence concerning this article should be addressed to Jade E. Adams-White, E-mail: j.adams-white@liverpool.ac.uk

Abstract

This study aimed to understand more fully some of the factors that influence decisions as related to air defence in a naval vessel's Operation Room. The study considered the impact of decision criticality (DC) and Task Load (TL) on measures of accuracy, confidence, and within-subjects confidence-accuracy (W-S C-A; a measure of metacognition). Personality constructs, workload, and situation awareness were also assessed. Participants were allocated to either a high, moderate, or low TL condition. Each took part in a computer-generated simulated air defence scenario where they were required to make a range of decisions and provide a corresponding confidence rating for each decision taken. Results showed that low DC increased confidence in decisions and high DC increased decision accuracy. Thereby DC significantly impacts on decision confidence and decision accuracy. In addition, those less tolerant of ambiguity were less accurate in their decision-making. Future studies should take account of these factors.

Keywords: Decision-making; Command and Control; Military; Metacognition.

General Audience Summary

Air defence decision-making is often conducted in a complex and uncertain environment. It is therefore important that the individuals faced with this task are able to make accurate and confident decisions under varying degrees of stress and criticality (i.e., the consequence associated with a decision). The purpose of this study was to examine external factors (e.g., task duration/stress) and internal factors (e.g., personality constructs) that may impact on air defence operator's decision-making abilities. In this study a measure of Within-Subjects Confidence-Accuracy was used. This measure considers the relationship between decision confidence and decision accuracy by assessing individual awareness of the accuracy of decisions made. For the task, a realistic set of scenarios, which varied in task difficulty, were designed with subject matter experts. Participants were required to make a range of decisions which varied in criticality and then rate how confident they were that they had made the best decision given the situation. The results demonstrated that the criticality of the decision impacted on both decision accuracy and confidence. Low decision criticality increased confidence in decisions and high decision criticality increased decision accuracy. The implications of this research include an increased understanding of decision criticality on decision-making in critical environments and the introduction of a novel method which has potential application in terms of informing the selection and in the training of personnel who are required to make accurate and confident decisions under conditions of uncertainty and stress. It is important to note that these inferences are based on findings from a novice sample and that non-trained staff are unlikely to make decisions in critical environments.

Measuring Decision Accuracy and Confidence of Mock Air Defence Operators

A ship's operation's room (OR) is the focal point for air defence decision-making. Large amounts of information must be attended to and managed to make tactical war-fighting decisions. Trained operators must detect, locate, and identify potential air threats, coupled with complex and cognitively demanding decisions in the uncertain environment of naval warfare. This often involves information overload and ambiguous information.

Lipshitz and Strauss (1997, p.150) define uncertainty as a "sense of doubt that blocks or delays action." Previous fatal air defence incidents emphasised the need to better understand how decisions are made under uncertain conditions. For instance, human error, including poor decision-making, was one of the main factors that led to the USS Vincennes shooting down an airliner after mistaking it for a hostile aircraft (Fogarty, 1988). To help militate against the impact of stress on decision-making, research was needed to gain an understanding of decision-making in critical and uncertain environments (Cannon-Bowers & Salas, 1998). One paradigm which aims to understand decision-making in such environments is Naturalistic Decision Making (NDM).

NDM aims to understand the way people use their experience to make decisions in field settings (Zsambok & Klein, 1997). It is domain specific and strives for high ecological validity. NDM investigates how experts make decisions in environments that have been defined as ill-structured, uncertain, ill-defined, high stakes, include feedback loops, organizational goals and norms, and are time stressed (Orasanu & Connolly, 1993). NDM attempts to understand human capabilities and the decision-making processes, not just outcomes. NDM models are therefore descriptive. A range of methods have been used to help obtain a better understanding of decision-making

processes in these environments, including knowledge-elicitation techniques (Kaempf, Klein, Thorsden & Wolf, 1996) and *microworlds* (Brehmer & Dorner, 1993).

The term metacognition refers to an awareness of ones' performance, and the ability and willingness to reflect on ones' thinking processes (Parker & Stone, 2014). Previous NDM metacognition research used qualitative methods, such as think-aloud protocols (Cohen, Freeman & Wolfe, 1996). However, more experimentally-based methods may benefit NDM research (Lipshitz, Klein, Orasanu & Salas, 2001). These methods allow more controlled testing to enhance understanding of variables involved in the decision-making process. Furthermore, experimental designs within the NDM paradigm may help to understand psychological constructs involved in decision-making (Elliot, Welsh, Nettelbeck & Mill, 2007). This paper's method uses realistic decision-making scenarios and a combination of subjective measures of confidence alongside objective scores of accuracy to investigate the metacognitive abilities of mock air defence operators. It could therefore advance NDM methodologies by using NDM concepts in conditions more akin to experimental paradigms.

Arguably, metacognitive confidence should be included in studies of decision-making because it is an important indicator of real-world outcomes (Jackson & Kleitman, 2014) and critical to performance (Rousseau, Tremblay, Banbury, Breton & Guitouni, 2010). Ensuring confidence is correctly placed has important implications. Overconfidence has been linked to underestimation of risk which could have a direct impact on the evaluation of future events (Lovallo & Kahneman, 2003). However, it is not only how confident one is in a decision, but the corresponding accuracy of the decision that is relevant (Wheatcroft & Woods, 2010). Strong positive relationships between confidence and accuracy are highly beneficial as they demonstrate an

individual's ability to weight information and subsequent decisions appropriately (Stichman, 1967).

Given the above, metacognition can be assessed by using decision confidence. The relationship between decision confidence and accuracy can provide a quantitative measure of metacognition (Fleming & Lau, 2014). One measure which has been used to assess this relationship is the within-subjects confidence-accuracy (W-S C-A) relation. The measure of W-S C-A has been defined as a "calculation which enables expression of individual confidence in each incorrect or correct response made" (Wheatcroft & Woods, 2010; p.195).

W-S C-A has been used successfully in domains such as forensic, investigative, and legal psychology (Wheatcroft & Woods, 2010; Wheatcroft, Kebbell & Wagstaff, 2004), perceptual tasks (Koriat, 2011), and general knowledge tasks (Buratti, Allwood & Kleitman, 2013). Recently W-S C-A has been used to examine the suitability of supervisory personnel for unmanned aircraft systems (Wheatcroft, Breckell, Jump & Adams-White, 2017).

The W-S C-A measure can add value to NDM in the assessment of individual awareness of the accuracy of decisions made. This approach is potentially similar to type 2 Signal Detection Theory (SDT) which assesses individual confidence in correct/incorrect responses (Clarke, Birdsall &Tanner, 1959). However, this approach remains to be consistently established empirically (Maniscalo & Lau, 2012). Whilst it is a subjective metacognitive measure, it has potential to affect the amount of resources applied to an action (Bingi, Turnipseed & Kasper, 2001) - crucial in air defence environments.

Air defence decisions may be influenced by both environmental and individual factors. Prior research has demonstrated potentially influential environmental factors to the relationship between confidence and accuracy. For example, both difficulty of decision (Wheatcroft, Wagstaff & Manarin, 2015) and decision danger (Wheatcroft et al., 2017) have shown to impact W-S C-A. This highlights the potential for W-S C-A to aid the understanding of external factors influencing the decision maker – such as the criticality of the decision to be made and the level of stress (Task Load, TL) experienced. Research has found criticality influences performance (Hanson, Bliss, Harden & Papelis, 2014). Decision criticality (DC) refers to the associated consequence of that decision. Hence, both DC and TL are crucial factors in an OR.

Research is required to increase understanding of individual differences that impact on air defence decision-making and in highlighting internal factors that influence effective decision-making. Individual differences, such as personality, play a key role (Jackson & Kleitman, 2014). Personality traits are important to decision-making as they can influence how people think, feel, and behave (Roberts, 2009). Wheatcroft et al. (2017) found that neuroticism was negatively related, and conscientiousness positively related, to confidence. It was also found that intolerance of ambiguity was negatively related to W-S C-A. Research by Jøsok et al. (2016) argues individual metacognitive ability is a relatively stable personality trait. It may therefore be beneficial to understand the relationship between metacognitive ability alongside other traits.

In summary, the aim of this paper is to introduce a metacognitive methodology that can be used to increase understanding of air defence decision-making. Further, the paper aims to begin to uncover some of the factors (TL & DC) related to decision-making in an OR air defence role and their implications on confidence, accuracy, and W-S C-A. Individual differences in personality and decision-related tendencies are also

considered. The study also aims to establish how W-S C-A aligns to the wider measurements currently used in human-machine interaction decision-making literature (mental workload and situational awareness).

In light of these points the following key hypotheses were formulated:

I. DC will impact on decision accuracy, confidence and W-S C-A.

II. High TL will reduce decision accuracy confidence, and W-S C-A.

III. Psychometric assessments of personality characteristics will demonstrate variable relationships in respect of accuracy, confidence and the W-S C-A relationship.

Method

Participants

Sixty participants were recruited through opportunity sampling from the University of Liverpool. The participants consisted of 30 females and 30 males with a mean age of 26 years (*SD* = 3.96). None of the participants had any prior experience in naval warfare operations as the study was initially interested in the OR role and novice capacity to the task. The sample size was decided upon by design, power, and previous studies using G power analysis with an effect size of 0.8 and significance level of .05 (Faul, Erdfelder, Lang & Buchner, 2007). The study received approval from the University of Liverpool's Institute of Psychology Health and Society Ethics Committee, and a favourable opinion from the Ministry of Defence Research Ethics Committee.

Design

A mixed measures quasi-experimental design was employed. Independent variables (IV) were Task Load and Decision Criticality. As such, a 3 (Task Load: Low,

Moderate, High) X 3 (Decision Criticality: Low, Medium, High); with repeated measures on the last factor. The dependent variables (DV) were confidence, accuracy, W-S C-A, personality constructs of openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (NEO-PI-R, Costa & McCrae, 1992), decision tendencies (i.e., tolerance to ambiguity; Budner, 1961; decision style; Roets & Van Hiel, 2007), subjective mental workload (NASA TLX; Hart & Staveland, 1988) and situational awareness (SART; Taylor, 1990).

Materials

Decision Logs. To ensure as high a level of ecological validity as possible in a quasi-experimental design, an air defence scenario was created with the guidance and assistance of subject matter experts (SMEs). The use of SMEs to assist in the experimental design is highly beneficial as SMEs are able to provide a unique insight into the appropriate and relevant situations that are likely to be met and applied in the study context. Four (4) SMEs with extensive knowledge of naval warfare and many years of experience in both Air Warfare Officer (AWO) and Principal Warfare Officer (PWO) roles were used to acquire the domain specific knowledge needed to provide the optimum ecologically valid options for decision making. The scenario uses a realistic set of events within a Peace Enforcement (PE) operation. A series of events and associated event decision logs were also created and agreed by SMEs. The event decision logs specify three decision options of reasonable equivalence for each event presented to the operator. SMEs agreed one option per decision made as the 'optimal/best' decision option given the current situation.

Computer Scenario .The visual display used as the stimulus for the experiment was created using VAPS XT (Virtual Avionics Prototyping Software) software. The screen depicted a pseudo-realistic radar screen which included an airlane, a No Fly Zone (NFZ), a coastline, and a border. A textbox to display additional information to assist decision-making and a timer which counted down from 20 seconds at each decision event was also included (see Figure 1). The algorithms used to animate the visual display symbols were created using Matlab/Simulink. The symbology used is as specified by APP-6c (NATO, 2008). Microsoft Movie Maker was used to edit the video (e.g., to apply timer). SMEs verified the display as sufficiently realistic.

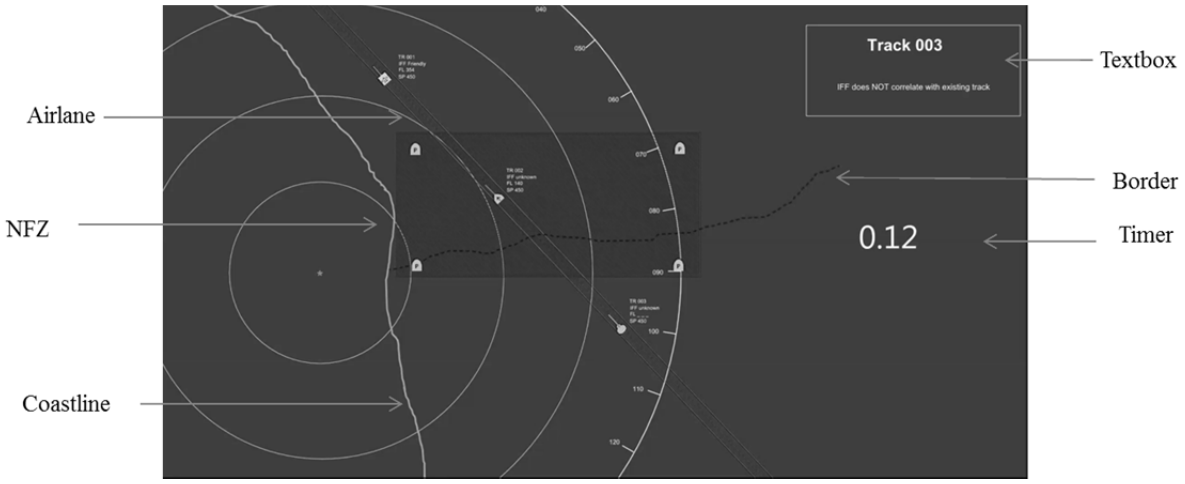

Figure 1. Visual display of the radar screen used in the experiment. The figure is taken from the moderate TL condition which shows three aircraft tracks as well as the airlane, No fly Zone (NFZ), coastline, textbox, border, and timer.

Questionnaires. The Situation Awareness Rating Technique (SART; Taylor 1990) was used to measure SA. To measure WL the NASA-TLX (Hart & Staveland, 1988) was utilised. NASA-TLX is a subjective workload assessment tool. Personality was assessed by the NEO-PI-R (Costa & McCrae, 1992).

Procedure

Participants were randomly allocated to a high, moderate, or low TL condition. Participants first completed participant demographic forms which collected data on age, gender, and occupation. Participants were also asked to complete paper-based questionnaires to gauge the relevance of a number of measures across groups (e.g., general personality constructs, thinking and reasoning) where they may be relevant to particular questions. Following this, participants were provided with the task booklet to read. The task booklet provided participants with information needed to assist them in the decision-making task, including air defence terminology and symbols. Once they had read the booklet, participants undertook a practice trial. The practice trial involved a series of decision events which allowed the participant to familiarise themselves with the task and the procedure. The duration of the trial was kept limited so as not to fatigue the participants before the experimental task (Barnes-Yallowley; personal communication, 2015). The questionnaire booklet presented three separate decision options based on the events of the scenario. One choice was required to be selected by placing a tick by the option they believed to be the 'best option given the current situation'. Participants were then required to rate how confident they were in the options chosen on a Likert scale, where 0 = *not at all confident* to 5 = *extremely confident*. After 20 seconds, the screen was blanked out to signal to the participants that the allocated decision time has ended. All participants then undertook the experimental air defence scenario, following the same procedure as described for the practice.

Thirty decision events were presented during the experimental simulation. A decision event was defined as an occasion where a decision may need to be made by an operator. For example, an unknown data link track appears on the screen. Decision Criticality was varied across the decision events presented (i.e., 10 high, 10 medium,

and 10 low; DC). Decision criticality related to the consequence of that decision. The decision held a higher criticality if there was a greater risk should the decision taken be incorrect in the high DC decisions (e.g., aircraft demonstrating hostile intent) compared to low DC (e.g., new track on radar screen) and the event occurrences varied depending on TL condition. The scenario stimulus ran for 20 minutes, 30 minutes, or 45 minutes for the high, moderate, and low conditions, respectively. The high condition involved an increased frequency of decisions, multiple decisions, and more aircraft tracks to monitor on the screen in comparison to the low condition which was characterised by one aircraft track to monitor at a time, decisions based only on the one track, reduced frequency of decision events, and longer periods of no action.

Once the scenario had finished, participants completed Situational Awareness and Workload questionnaires. Participants were fully debriefed to ensure each understood the nature of the study and given the opportunity to ask further questions.

Results

To assess the differences in means a number of statistical analyses were performed on the data for accuracy, confidence and W-S C-A using Analysis of Variance (ANOVA). A manipulation check was carried out to assess the differences in TL (see analysis of Workload and Situational Awareness). No significant differences were found in WL, there were differences in SA. The TL manipulation was not significant. An alpha level of .05 was used for all statistical tests.

Accuracy

The accuracy of the decisions was decided on by the SMEs. When designing the decision log and generating the decision options one of the decision options was voted

to be the best decision given the current situation. Participants scored '1' for *an accurate response* or '0' for *an incorrect response*. The maximum total score was 30 and the maximum mean for each DC was 10. To examine the mean differences between TL and DC in accuracy an ANOVA was conducted.

A 3 (Task Load [TL]: High, Moderate, Low) x 3 (Decision Criticality [DC]: High, Medium, Low) mixed ANOVA, with repeated measures on the last factor, was conducted on the data (see Table 1).

A main effect of DC was found, $F(2, 114) = 16.71$, $p = .01$, $\eta_p^2 = .23$. Bonferroni corrected post hoc tests showed participants were more accurate in high DC decisions ($M = 5.25$, $SD = 1.94$) than low DC decisions ($M = 3.55$, $SD = 2.00$). Additionally, participants were more accurate in medium DC decisions ($M = 4.88$, $SD = 1.79$) than low DC decisions [both $p < .01$].

However, no main effect of TL was found $F(2, 57) = 2.03$, $p = .14$, $\eta_p^2 = .07$. Additionally, no interaction effect was observed $F(4, 114) = 1.77$, $p = .14$, $\eta_p^2 = .06$.

These findings lend support to hypothesis I that DC would impact on decision accuracy. Participants were more accurate in the decisions taken which were highly critical in comparison to the low criticality decisions. However, hypothesis II was not supported, no differences were found in accuracy of decision between the TL conditions.

Table 1

Means and Standard Deviations as influenced by accuracy according to Task load and Decision Criticality

Task Load	Overall	High DC	Medium DC	Low DC
High	13.80 (4.16)	5.15 (2.78)	4.60 (1.79)	4.05 (2.21)
Moderate	12.30	4.50	4.70	3.20

	(3.61)	(1.64)	(1.84)	(2.02)
Low	14.85	6.10	5.35	3.40
	(3.80)	(1.65)	(1.73)	(1.73)
Total	13.65	5.25	4.88	3.55
	(3.94)	(1.94)	(1.79)	(2.00)

Note. Standard deviations are in parenthesis

Confidence

Participants were asked to rate confidence in each decision - '0' being *not confident at all* and '5' *being extremely confident*. The maximum confidence score in total was 150 and for each DC 50. A 3 x 3 mixed ANOVA was carried out to assess the impact of TL and DC on decision confidence. As Mauchly's test of sphericity was found to be significant, the Greenhouse-Geisser estimate for *df* was used (see Table 2).

A main effect of DC was found $F(2, 88) = 3.29$, $p = .05$, $\eta_p^2 = .55$. A Bonferroni corrected post hoc test showed that participants were significantly more confident in low DC decisions ($M = 37.25$, $SD = 9.60$) than medium DC decisions ($M = 35.05$, $SD = 7.23$), $p = .02$. No significant differences were found between high DC and low DC or medium DC and high DC, $p > .05$.

No main effect of the TL condition was found, $F(2, 57) = 1.32$, $p = .27$, $\eta_p^2 = .04$. Similarly, no interaction effect was observed, $F(4, 88) = 2.13$, $p = .10$, $\eta_p^2 = .07$. These results demonstrate that DC impacted on decision confidence thus supporting hypothesis I. Participants were more confident in the low DC decisions compared to the medium DC decisions. No support was found for hypothesis II; the results show no differences for TL or decision confidence.

Table 2

Means and Standard Deviations for Confidence as influenced by Task Load and Decision Criticality

Task Load	Overall	High DC	Medium DC	Low DC
High	102.10 (26.81)	32.45 (11.93)	32.90 (8.78)	36.70 (13.65)
Moderate	111.90 (21.67)	38.75 (7.36)	36.00 (6.99)	37.40 (8.59)
Low	113.00 (14.90)	38.55 (5.60)	36.30 (5.40)	37.65 (5.06)
Total	109.00 (21.88)	36.58 (9.06)	35.07 (7.23)	37.25 (9.60)

Note. Standard deviations are in parenthesis

W-S C-A

A 3 x 3 mixed ANOVA was performed on the relationship between TL and DC on individuals within-subjects confidence-accuracy (W-S C-A).

There was no main effect of DC shown, $F(2, 98) = 0.62$, $p = .54$, $\eta_p^2 = .01$ and no main effect of TL, $F(2, 49) = 1.61$, $p = .20$, $\eta_p^2 = .06$ was observed. No interaction was observed, $F(4, 98) = 0.61$, $p = .66$, $\eta_p^2 = .23$. An observation of the descriptive statistics showed individual W-S C-A was found to be lowest between participants in moderate TL and within participants in medium DC (see Supplementary Material). W-S C-A was found to be highest between participants in the low TL condition and within participants in high DC. Overall W-S C-A scores were very low and not negative ($M = .02$). The findings do not support the hypotheses that DC and TL would impact on W-S-C-A.

Percentage Confidence in Correct and Incorrect Responses

To consider the variation in the data and examaine the low correlations displayed for W-S C-A, percentage confidence in correct and incorrect responses was calculated. W-S C-A demonstrates the relationship between confidence and accuracy; however, a high correlation suggests both being highly confident in correct decisions as well as low confidence in incorrect decisions. Similary, a negative correlation would suggest that individuals are highly confident in incorrect reposnses or not confident in correct

responses. Thus, by examining the percenatge confidence in incorrect or correct responses the direction of the confidence (over/under confidence) can be displayed.

To do this the number of correct responses was recorded and the confidence in those decisions calculated to produce percentage confidence in correct responses. The same method was applied to incorrect responses (see Table 3). Interestingly, all data suggests a high degree of confidence in decisions (M = 70.58, SD = 18.42).

To examine the differences in TL a one way ANOVA was conducted on percentage confidence in correct responses. No main effect of TL on percentage confidence in correct decisions was observed, $F(2, 57) = 0.49$, $p = .62$, $\eta_p^2 = .13$. Similarly, an ANOVA was conducted with percentatge confidence in incorrect responses. Again, no main effect percentage incorrect was found, $F(2, 57) = 1.20$, $p = .31$, $\eta_p^2 = .25$.

Percentage Accuracy

To examine how confidence related to the accuracy of decision, percentage accuracy was calculated and analysed. A further ANOVA found no main effects of TL, $F(2, 57) = 2.77$, $p = .07$, $\eta_p^2 = .09$. See also Table 3. The results for percenatge accuracy and confidence further demonstrate that TL did not make sufficient impact to reach significance; thus not supporting the experimental hypothesis.

Table 3

Means and Standard Deviations for % confidence (correct and incorrect) and % accuracy according to Task Load

Task Load	% Confidence Correct	% Confidence Incorrect	% Accuracy
High	67.40 (17.20)	67.45 (18.55)	45.17 (13.83)
Moderate	73.10 (15.54)	74.85 (14.40)	43.83 (12.04)

Low	71.25 (22.33)	73.53 (14.68)	46.27 (12.67)
Total	70.58 (18.42)	71.88 (16.04)	45.06 (13.15)

Note. Standard deviations are in parenthesis

Workload (WL) and Situational Awareness (SA)

To assess the relationship between workload and SA, a series of Pearson's correlations were calculated. A significant negative relationship was found between SA and WL, r (58) = -.53, p = .001. Higher levels of reported WL were related to lower feelings of SA.

A one-way ANOVA was conducted to assess the relationship between SA and TL. There was a significant effect of TL condition on SA, F(2, 57) = 6.44, p = .001. Participants in the low TL condition reported higher levels of subjective SA (M = 21.40, SD = 4.67) than participants in high TL (M = 14.30, SD = 5.42) p = .001.

No significant relationship was found between WL and TL, F(2, 57) = 3.00, p = .06. As a non-significant relationship was found this would suggest that the manipulation check was not successful.

As an exploratory analysis, the 6 dimensions of the NASA TLX (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration) were also examined to determine whether differences existed across the conditions. One way ANOVAs were conducted with TL across the different dimensions of workload (see Table 4).

Table 4

Means and Standard Deviations for dimensions of WL as influenced by Task Load

Task Load	Overall SA	Overall WL	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration
High	14.30 (5.42)	64.00 (14.41)	257.50 (122.73)	0.75 (2.45)	299.00 (89.92)	168.25 (108.23)	133.50 (89.24)	101.50 (87.39)
Mod	18.70 (8.29)	60.79 (15.42)	250.50 (113.53)	6.00 (24.58)	220.75 (118.70)	146.25 (113.22)	162.00 (87.24)	133.00 (126.91)
Low	21.40 (4.65)	53.32 (12.40)	224.00 (130.35)	3.25 (13.40)	184.75 (98.76)	185.75 (102.83)	103.75 (89.45)	100.75 (101.61)
Total	18.13 (6.87)	59.37 (14.60)	244.00 (121.18)	3.33 (16.10)	234.83 (112.23)	166.75 (107.57)	133.10 (90.37)	111.75 (105.83)

Note. Standard deviations are in parenthesis

There was a significant difference found for Temporal Demand, $F(2, 57) = 6.41$, $p < .003$, $\eta_p^2 = .18$. Comparisons show that there was a significant difference between the conditions high ($M = 299.00$, $SD = 89.92$) and moderate ($M = 220.75$, $SD = 118.70$) ($p = .02$) and high and low ($M = 184.75$, $SD = 98.76$) ($p = .001$). The findings suggest that participants felt more time pressured due to the rate and pace at which the task elements occurred in the high and moderate TL conditions in comparison to the low task condition. No significant differences were found between moderate and low TL conditions, $p > .05$.

No main effects were found for Mental Demand, $F(2, 57) = 0.42$, $p = 0.66$, $\eta_p^2 = 0.14$, Physical Demand, $F(2, 57) = 0.524$, $p = 0.60$, $\eta_p^2 = .02$, Performance $F(2, 57) = 0.67$, $p = 0.52$, $\eta_p^2 = .02$, Effort $F(2,57) = 2.16$, $p = 0.13$, $\eta^2 p = .07$ or Frustration $F(2,57) = 0.60$, $p = .55$, $\eta_p^2 = .02$ between conditions of high, moderate and low TL. The findings therefore suggest that the main difference between the task conditions was the speed at which the task events occurred. Similarly, to examine the three dimensions of

SA as measured by SART (Demand, Supply, and Understanding) one way ANOVAs were carried out across each dimension and TL condition (see Table 5).

Table 5

Means and Standard Deviations for dimensions of SA according to Task Load

Task Load	Demand	Supply	Understanding
High	13.70	17.20	10.80
	(2.81)	(3.37)	(2.82)
Mod	13.65	20.10	12.85
	(4.57)	(3.11)	(3.21)
Low	10.05	18.25	13.10
	(3.73)	(2.12)	(3.84)
Total	12.47	18.52	12.25
	(4.10)	(3.11)	(3.42)

Note. Standard deviations are in parenthesis

Attentional demand includes measures which assess individual feelings of the instability of the situation, variability of the situation and complexity of the situation. Significant differences were found for demand of the task for the different task conditions, $F(2, 57) = 6.15$, $p = .001$, $\eta_p^2 = .18$. Demand was significantly higher for high than low ($p = .01$). Moderate was found to be significantly higher than low ($p = .01$). No differences were found between moderate and high ($p > .05$). Therefore, these findings suggest that high and moderate task conditions increased participant's feelings of attentional demand. Specifically, participants rated the likeliness of the situation to change suddenly, number of variables that require attention, and the degree of complication of the situation as higher in the high and moderate task conditions.

Attentional supply includes constructs of arousal, spare mental capacity, concentration and division of attention. A significant difference was also found for attentional supply, $F(2, 57) = 5.07$, $p = .009$, $\eta_p^2 = .51$. Comparisons showed significant differences between high and moderate ($p = .01$) with attentional supply being higher

for the moderate condition than high. No significant differences between low and high or moderate and low conditions were observed. Therefore, results show that participants rated a higher degree of readiness for the activity, amount of mental ability to apply to new tasks, degree to which individual thoughts are bought to bear on the situation, and the amount of division of attention on the situation in the moderate condition.

Understanding includes measures of information quantity, quality and familiarity. No significant differences for understanding was found, $F(2, 57) = 2.89$, $p = .06$, $\eta_p^2 = .09$. Hence, participants rated the amount of knowledge received and understood, degree of value of knowledge communicated, and the degree of acquaintance with the situation as the same across the conditions. These findings suggest that the task manipulation induced differences in demand and supply but not understanding and this would seem likely as all participants were provided with the same information.

An interesting finding was that the attentional supply was higher for moderate condition than high; this suggests that participants may have struggled more with the potential uncertainty that the moderate condition might have brought to bear. It appears participants were more able to deal with the ends of the spectrum where real differential could be identified (i.e., low and high conditions).

Relationships between WL, SA and Accuracy, Confidence and W-S C-A

To establish whether a relationship existed between WL, SA, accuracy, confidence, and W-S C-A a number of Pearson's correlations were carried out. Results revealed a significant negative relationship was found between overall WL and confidence, $r\ (58) = -.42$, $p = .001$. As subjective measures of workload increased, confidence in decisions decreased.

In addition, a significant strong positive relationship was found between overall SA and confidence, r (58) = .63, p = .001. Higher scores in subjective SA were related to higher scores of confidence in decisions.

However, no significant relationships were found between SA, WL, and W-S C-A, or between SA and accuracy or WL and accuracy in decisions; all comparisons, p > .05. Furthermore, no significant relationship was found between-subjects confidence and accuracy, p > .05. The findings suggest that decision confidence influences both WL and SA. In this study accuracy was found to be unrelated to WL and SA.

Personality Constructs

This study was also interested in establishing whether accuracy, confidence, and W-S C-A were related to the psychometric scores. For this, Pearson's correlations were conducted. A significant negative relationship was found between tolerance to ambiguity and accuracy, r (58) = -.34, p = .008. Those who scored higher on the tolerance to ambiguity scale (i.e., less tolerant) were less accurate.

In addition, a significant negative relationship was also found between Decision Style and Accuracy r (58) = -.35, p = .005. High scorers on the decision style scale were less accurate. Decision style explicitly probes the need for quick and unambiguous answers.

To investigate whether there were individual differences in participants experiences of WL and SA correlations were conducted on each measure of the NEO-PI-R (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism). The results showed a significant relationship was found between Openness to Experience and WL, r (58) = -.28, p = .03. High scorers on the Openness to Experience scale reported lower levels of WL during the task. No other relationships

were found to be significant, $p > .05$. These findings therefore suggest that some cognitive constructs are involved in decision accuracy and individual differences in participants' feelings of WL.

Discussion

A novel method to measure metacognitive ability to assess the impact of Decision Criticality (DC) and Task Load (TL) on measures of confidence, accuracy, and W-S C-A was used with mock air defence operators. Personality constructs, workload (WL), and situational awareness (SA) were also assessed. DC impacted on both decision confidence and decision accuracy. Low DC was found to increase confidence in decisions and high DC increased decision accuracy. Cognitive constructs were also found to be related to decision accuracy.

The findings suggest that accuracy increases with DC. Participants made more accurate decisions in high DC than both low DC and medium DC. This outcome supports previous literature that criticality influences performance (Hanson et al., 2014; Wheatcroft et al., 2017). Research has also shown that task performance increases when participants find the task more important (Kliegel, et al., 2004) perhaps indicating that individuals believed high DC decisions to be important within the task context. This could relate to participants applying more attention and effort to decisions with greater consequences for an incorrect decision. Future research could examine decision processes and which mechanisms lead to increased accuracy.

The study also demonstrates that DC also influenced confidence. Individuals were significantly more confident in low DC decisions than medium DC decisions, lending support to previous literature that confidence decreases as difficulty increases (Chung & Monroe, 2000; Kebbell, Wagstaff & Covey, 1996). No significant differences were

found between high and low DC or medium and high DC. This could be due to increased uncertainty as condition criticality increased in the conditions relative to low DC, but high DC was perceptually transparent. Nevertheless, it is the corresponding confidence relative to an individual's awareness of the accuracy of decisions that is most important. W-S C-A remained unaffected, with no significant differences evident in W-S C-A across TL and DC. Some research has shown that training and experience improves calibration (Lichstenstein, Fischhoff & Phillips, 1977). It would therefore be beneficial to conduct further studies using naval participants with appropriate experience.

The high TL condition did not impact on decision confidence, accuracy, or W-S C-A. However, the manipulation check was not significant; this could be one reason why no differences were found for some variables. The results support previous research which demonstrates that confidence is a relatively robust and general trait (Stankov & Lee, 2008). Confidence remained high, irrespective of accuracy. No relationship was found between decision confidence and accuracy. The accuracy scores were just below chance but individuals displayed elevated confidence - the means of both correct and incorrect scores were around 70%, suggesting that individuals are unaware of incorrect responses and overconfident in some decisions taken. This is consistent with previous literature that has demonstrated a general tendency for overconfidence (Lichtenstein et al., 1977). Importantly, none of the participants had any prior knowledge of air defence decision-making. Consequently, an additional explanation for elevated confidence levels can be explained by the Dunning-Kruger effect (Kruger & Dunning, 1999) where unskilled/novice individuals assess their ability to be too high.

The study also examined how SA and WL relate to decision confidence, accuracy, and W-S C-A. It was found that SA was related to decision confidence. Individuals who reported higher levels of SA were also more confident in their decisions. However, these findings should be taken with caution. SA was measured subjectively and a confidence bias has previously been found in SA reporting (Sulistyawati & Chui, 2009). Thus, it might be that individuals are generally confident in their assessments of SA performance. Importantly, SA was not related to accuracy in decisions taken. Individuals may privately believe they had a better understanding of the situation than they accepted. Conversely, WL was found to be negatively related to overall decision confidence. Higher reported levels of WL resulted in lower levels of decision confidence. This is important for decision-making; reduced confidence in decisions taken could lead to increased WL. Individuals may seek out more information to support/contradict decision certainty.

Further analysis of the WL dimensions showed significant differences in individuals' feelings of temporal demand in the high TL condition. Participants felt more time pressured and reported the speed at which the task events occurred to be higher. Time pressure has been linked to individuals using different decision-making strategies (Maule, Hockey & Bdzola, 2000); thus, speed of decision events may play an important role in critical environments. Further research with experts may demonstrate a mitigated effect.

The investigations into broad personality constructs were found to be unrelated to confidence, accuracy, W-S-C-A, or SA. This suggests the constructs may not be related to these measures or sufficiently salient to decision-making processes. Relationships did exist with other measured constructs. Individuals less tolerant to ambiguity were less accurate in their decisions and high scorers on the decision style scale were also less

accurate. Budner (1961) argues that individuals who are less tolerant find ambiguous situations threatening. It is probable that a lack of tolerance hindered individuals' ability to make accurate decisions. Tolerance to ambiguity is a likely desirable trait for accurate air defence decision-making. Further, although not replicated in this study, Wheatcroft et al. (2017) found an intolerance of ambiguity was negatively related to W-S C-A (i.e., a greater tolerance of ambiguous conditions was related to increased W-S C-A). Further research is warranted to investigate the relationship between decision-making and ambiguity tolerance in critical environments. Outcomes also showed WL to be negatively related to openness to experience, providing support for the findings that some aspects of personality impact on the perception of WL (Chiorri, Garbarino, Bracco & Magnavita, 2015).

Although the study was initially interested in the OR role and novice capacity to the task, one limitation was the use of novice participants rather than experts. However, as suggested by Hoffman & Klein (2017) it may be beneficial to the NDM paradigm to understand how expertise is developed (Sala & Gobert, 2016). For instance, Klein, Hintze & Saab (2013) developed the shadow box technique which helps novices understand the decision-making processes of experts. Therefore, the use of novices in this study does allow for a baseline comparison. The use of novices may also help to understand the training needs of less experienced decision-makers. The authors will use experts to further validate the work in future research. Research has found that practice can degrade certain aspects of metacognitive performance (Jackson, Kleitman & Aldman, 2014). However, this study minimised this effect by reducing potential fatigue.

It has been argued that NDM research should use a mixture of measures to reduce the limitations of using a single methodology (Lipshiz et al., 2001). This paper aimed to introduce the W-S C-A measure to assess an element of metacognitive ability in air

defence operators using a realistic decision-making scenario together with combined objective and subjective measures. The proposed method and outcomes will provide a wider view of metacognition in critical decision-making environments. The broader implications include the potential for the approach to be used to prioritize training and selection, with the aim of improving effective air-defence decision-making.

Conclusion

This study found DC has significant impact on both decision accuracy and confidence. Future work should consider the impact of decision criticality and tolerance ambiguity on accurate air defence decision-making.

Author contributions

All named authors were involved in the conception and design of the study, critical revision of the article, and final approval of the published version. The first author was responsible for data collection and data analysis, interpretation, and in drafting the final article.

Acknowledgements

The research is supported by the Defence Science Technology Laboratory (contract no. DSTLX-1000092181). As the work is funded by DSTL no data can be made available to other researchers. However, sufficient information is available in the paper for researchers to use the analytical methods should they wish to do so. The authors would also like to thank Systems Engineering and Assessment Ltd for their valuable assistance during the expert decision development stage of the study. No conflicts of interest are declared.

References

Barnes-Yallowley, J. (2015). Subject matter expert personal communication. Somerset: SEA.

Brehmer, B., & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior*, *9(2),* 171-184.

Bingi, P., Turnipseed, D., & Kasper, G. (2001). The best laid plans of mice and men: The role of decision confidence in outcome success. *North American Journal of Psychology*, *3(1),* 91-108.

Budner, S. (1961). An Investigation of Intolerance to Ambiguity. *Nursing Research*, *10(3),* 179.

Buratti, S., Allwood, C. M., & Kleitman, S. (2013). First- and second-order metacognitive judgments of semantic memory reports: The influence of personality traits and cognitive styles. *Metacognition and Learning*, *8(1),* 79–102.

Cannon-Bowers, J. A., & Salas, E. E. (1998). *Making decisions under stress: Implications for individual and team training*. Washington DC: American Psychological Association.

Chiorri, C., Garbarino, S., Bracco, F., & Magnavita, N. (2015). Personality traits moderate the effect of workload sources on perceived workload in flying column police officers. *Frontiers in Psychology*, *6*.

Chung, J., & Monroe, G. (2000). The effects of experience and task difficulty on accuracy and confidence assessments of auditors. *Accounting & Finance*, *40(2)*, 135-151.

Clarke, F. R., Birdsall, T. G., & Tanner*, W. P. Jr. (*1959*). Two types of ROC curves and definitions of parameters. Journal of the Acoustical Society of America, 31*, 629-630.

Cohen, M. S., Freeman, J. T., & Wolf, S. (1996). Metarecognition in time-stressed decision-making: Recognizing, critiquing, and correcting. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *38*(2), 206-219.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Elliott, T., Welsh, M., Nettelbeck, T., & Mills, V. (2007). Investigating naturalistic decision making in a simulated microworld: What questions should we ask? *Behavior Research Methods*, *39(4)*, 901-910.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39(2)*, 175-191.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443.

Fogarty, W. M. (1988). Formal investigation into the circumstances surrounding the downing of a commercial airliner by the U. S. S. Vincennes (CG 49) on 3 July 1988 [Unclassified Letter Ser. 1320 of 28 July 1988, to Commander in Chief, U.S. Central Command]. Washington, DC: U.S. Department of the Navy.

Hanson, J. A., Bliss, J. P., Harden, J. W., & Papelis, Y. (2014). The effects of reliability and criticality on an IED interrogation task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58(1),* pp. 2340-2344. Chicago, USA:. Sage Publications.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: Elsevier.

Hoffman, R. R., & Klein, G. L. (2017). Challenges and prospects for the paradigm of naturalistic decision making. *Journal of Cognitive Engineering and Decision Making*, *11(1)*, 97-104.

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: capturing decision tendencies in a fictitious medical test. *Metacognition and Learning*, *9(1)*, 25-49.

Jackson, S. A., Kleitman, S. & Aldman, E. (2015). Low cognitive load and reduced arousal impede practice effects on executive functioning, metacognitive confidence and decision making. *PLoS ONE, 9(12).*

Jøsok, Ø., Knox, B. J., Helkala, K., Lugo, R. G., Sütterlin, S., & Ward, P. Exploring the hybrid space. *International Conference on Augmented Cognition, July 2016,* (pp. 178-188). Chicago, USA: Springer International Publishing.

Kaempf, G. L., Klein, G., Thordsen, M. L., & Wolf, S. (1996). Decision-making in complex naval command-and-control environments. *Human Factors*, 38(2), 220-231.

Kebbell, M. R., Wagstaff, G. F., & Covey, J. A. (1996). The influence of item difficulty on the relationship between eyewitness confidence and accuracy. *British Journal of Psychology*, *87(4)*, 653-662.

Kliegel, M., Martin, M., McDaniel, M., & Einstein, G. (2004). Importance effects on performance in event-based prospective memory tasks. *Memory*, *12(5)*, 553-561.

Klein, G., Hintze, N., & Saab, D. Thinking inside the box: The Shadow Box method for cognitive skill development. *Proceedings of the 11th International Conference on*

Naturalistic Decision Making, Marseille, France, 24 May 2013, pp. 121-124. Paris: Arpege Science Publishing.

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, *140(1)*, 117-139.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77(6)*, 1121-1134.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In H. Jungermann & G. de Zeeuw (Eds.) *Decision Making and Change in Human Affairs* (pp. 275-324). Dordrecht, Netherlands: Reidel.

Lipshitz, R., & Strauss, O. (1997). Coping with uncertainty: A naturalistic decision-making analysis. *Organizational Behavior and Human Decision Processes*, *69(2)*, 149-163.

Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, *14(5)*, 331-352.

Lovallo, D., & Kahneman, D. (2003). Delusions of success. *Harvard Business Review*, *81(7)*, 56-63.

Maule, A. J., Hockey, G. R. J., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta Psychologica*, *104(3)*, 283-301.

Maniscalo, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21,* 422-430.

NATO (2008). Department of defence interface standard. Received from http://www.dtic.mil/doctrine//doctrine/other/ms_2525c.pdf.

Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. In G. A. Klein, J. Orasanu, R. Calderwood & C.E. Zsambok (Eds.), *Decision Making in Action: Models and Methods* (pp. 3-20). Norwood, NJ: Ablex.

Parker, A. M., & Stone, E. R. (2014). Identifying the effects of unjustified confidence versus overconfidence: Lessons learned from two analytic methods. *Journal of Behavioral Decision Making*, *27(2)*, 134-145.

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality*, *43(2)*, 137-145.

Roets, A., & Van Hiel, A. (2007). Separating ability from need: Clarifying the dimensional structure of the need for closure scale. *Personality and Social Psychology Bulletin*, *33(2)*, 266-280.

Rousseau, R., Tremblay, S., Banbury, S., Breton, R., & Guitouni, A. (2010). The role of metacognition in the relationship between objective and subjective measures of situation awareness. *Theoretical Issues in Ergonomics Science*, *11(1-2)*, 119-130.

Sala, G., & Gobet, F. (2016). Do the benefits of chess instruction transfer to academic and cognitive skills? A meta-analysis. *Educational Research Review*, *18*, 46-57.

Stichman, E. P. (1967). Relationship of expressed confidence to accuracy of transcription of operational communication personnel. U.S. Army BESRL Technical Research Note No.192.

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology*, *100(4)*, 961.

Sulistyawati, K., & Chui, Y. P. Confidence bias in situation awareness. *International Conference on Engineering Psychology and Cognitive Ergonomics, July 2009,* (pp. 317-325). Heidelberg, Berlin: Springer.

Taylor, R. M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. *Situational Awareness in Aerospace Operations* (AGARD-CP-478) (pp. 3/1–3/17). Neuilly Sur Seine, France: NATO - AGARD.

Wheatcroft, J. M., Jump, M., Breckell, A. L., & Adams-White, J. (2017). Unmanned aerial systems (UAS) operators' accuracy and confidence of decisions: Professional pilots or video game players? *Cogent Psychology*, *4(1)*.

Wheatcroft, J. M., & Woods, S. (2010). Effectiveness of Witness preparation and cross-examination non-directive and directive leading question styles on witness accuracy and confidence. *The International Journal of Evidence & Proof*, *14(3)*, 187-207.

Wheatcroft, J.M., Wagstaff, G.F., & Kebbell, M.R. (2004). The influence of courtroom questioning style on actual and perceived eyewitness confidence and accuracy. *Legal & Criminological Psychology, 9*, 83-101.

Wheatcroft, J.M., Wagstaff, G.F., & Manarin, B. (2015). The influence of delay and item difficulty on eyewitness confidence and accuracy. *International Journal of Humanities and Social Science Research, 1,* 1-9.

Zsambok, C. E., & Klein, G. A. (Eds.) (1997). *Naturalistic decision making*. Mahwah, N.J: Lawrence Erlbaum Associates.