Validating IMPACT: A new cognitive test battery for defence

Erinn Sturgess[1], Victoria Steane[1] & Mark Chattington[1]

[1]Research Technology & Innovation, Thales, Reading, United Kingdom

SUMMARY

This paper introduces the Interactive Measures of Performance and Assessment of Cognitive Tasks (IMPACT) tool, a new cognitive test battery for defence human sciences practitioners. The paper describes a comprehensive research study adopting a multi-method approach using a range of subjective and objective measures of human performance and cognitive states (including functional Near Infrared Spectroscopy; electrodermal activity; heart rate variability; gaze metrics; subjective workload; task performance and subjective situational awareness) to validate the assumptions made about IMPACT's ability to elicit a number of cognitive properties. A brief overview of the findings is presented demonstrating the potential of the tool to elicit different behavioural outcomes based on task load manipulation – an important first step in understanding the construct validity of the IMPACT tool for human sciences practitioners moving forward.

KEYWORDS

Construct validity, defence, IMPACT

Introduction

Over the past four years, the Defence Science and Technology Laboratory (Dstl) have sought to develop a new generic, military orientated, cognitive test battery called the Interactive Measures of Performance and Assessment of Cognitive Tasks (IMPACT) tool. They recognised that human sciences practitioners within the defence community have lacked a tool capable of capturing human performance data for computer based military themed cognitive tasks, despite tools existing for other domains (e.g., the National Aeronautical and Space Administration (NASA) Multi-Attribute Task Battery II for aircrew; Comstock & Arnegard, 1992). The newly developed IMPACT tool has been intentionally designed to provide coverage of all military domains and cover several "key cognitive abilities necessary for military system users" (Tatlock et al., 2015). IMPACT consists of six tasks, an overview of which can be found in Sabine & Thompson (2024). However, for ease, a brief summary is provided in Table 1.

Thales UK were tasked to perform an independent assessment of the IMPACT tool and generate an evidence base that could be used to provide insight into whether the tool, and the tasks of which it is comprised, elicits the desired properties, and hence provide a starting point on the validation of the tool. Validity refers to how accurately an assessment method, technique and/or tool measures something. Within the academic literature, there are many definitions and types of validity but Trochim (2001) proposes that all types fall under the broad heading of 'construct validity'. This overarching term encompasses all forms of validity which refers to the extent to which a measure adequately assesses the construct it aims to assess (Nunnally & Bernstein, 1994).

Table 1. IMPACT task descriptions

Task	Description
Communications (Comms)	Tests a participant's ability to track and report a fictional convoy of vehicles' location.
Target Acquisition (TA)	Tests the ability of participants to identify targets from simulated Unmanned Aerial Vehicle footage and then classify as friendly or hostile.
Formation Maintenance (FM)	Tests the ability of a participant to navigate a squadron of unmanned vehicles through a simulated maritime environment.
Identify Friend or Foe (IFF)	Tests a participant's ability to correctly classify contacts on a screen when given auditory instructions.
Resource Management (RM)	Tests the ability of participants to prioritize the demands of three units while depleting their resources and time deliveries to ensure all three remain well supplied.
Immediate Action (IA)	Tests the ability of participants to respond in a timely manner to fictional events.

Physical sciences have many specific measures, such as weight and length, which are concrete and agreed. This is unlike the human sciences domain where many measures are less tangible (Smith, 2005). Construct Validity is at the centre of any study in which researcher use a measure as an index that is not directly observable (e.g. working memory; Westen & Rosenthal, 2003). Therefore, when it comes to measuring the validity of an assessment method, technique and/or tool, there are a number of approaches that can be taken. For example, the inference-based approach argues that validity is not a property of a test; instead, it is a property of the interpretations and inferences made about the resulting data (Kane, 2013; Borsboom et al., 2004). In contrast, Messick (1987) outlined a validity framework identifying alternative sources of validity evidence (e.g. content-orientated, response process, internal structure, relations to other variables and consequences of assessment). Smith (2005) defined a five-step model for construct validity. This partly referred to the ability for researcher to develop sound hypotheses from theory of the construct that you are looking into. Regardless of the approach taken, assessing the reliability and validity of a tool should be viewed as an ongoing process that requires the accumulation of evidence over time, settings, and samples to build a scientifically sound foundation (Cronbach & Meehl, 1995; Rickards et al., 2012).

Approach

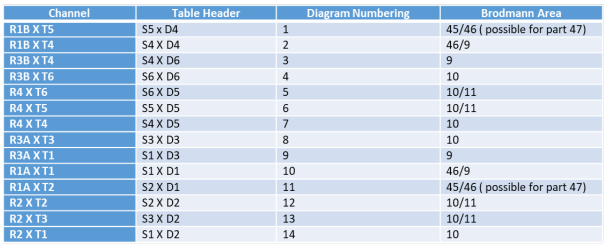
To explore the construct validity of the IMPACT tool and its ability to elicit a number of cognitive properties, a comprehensive study protocol was designed. This received favourable opinion by the Ministry of Defence Research Ethics Committee (MODREC) prior to the study (application number: 2225/MODREC/23). A multi-methods approach, utilising a range of subjective and objective measures of human performance and cognitive state was adopted to show how the manipulation of task load impacts on individual's performance and cognitive state when using the IMPACT tool and all of its associated tasks. Measures included the use of functional Near Infrared Spectroscopy (fNIRS); electrodermal activity (EDA); heart rate variability (HRV); gaze metrics including pupil dilation, fixation duration and saccadic eye movements; a subjective workload questionnaire (NASA-TLX; Hart & Staveland, 1988); task performance measures; and a subjective situational awareness questionnaire (3D SART; Taylor, 1990). fNIRS, a relatively novel neuroimaging technique, was used to assess cortical activity during the completion of each IMPACT task, thus providing insight into different brain regions that are implicated during task completion and allowing the perceived mapping by Sabine (2022) and Sabine & Thompson (2024) to be explored in greater detail. EDA was chosen because it is deemed to be a sensitive psychophysiological index of changes in Autonomic Nervous System arousal, which can be used to infer emotional and cognitive state (Yu & Sun, 2020). Similarly, HRV was chosen because it reflects the fluctuation in the time intervals between adjacent heartbeats and therefore the regulation of the Autonomic Nervous System. HRV is often used as a sensitive measure of workload (Shakouri et al., 2018), cognitive pressure (Mulder & Mulder-van der Meulen, 1973), and stress (Kim et al., 2018). Finally, the gaze metrics used in this study represent standard measures that can be used to help understand the distribution of visual attention across a user interface (Chandra et al., 2015). This multi-methods approach was chosen because it provides a more holistic view of human performance and cognitive state, thus facilitating discussion between methods (Izzetoglu & Richards, 2019). For each of these measures, hypotheses were derived based on the surrounding scientific research.

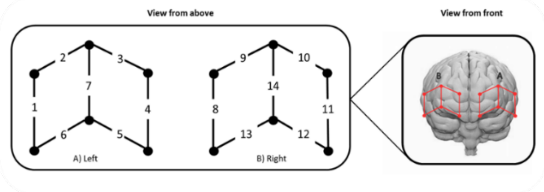
Thirty-five participants took part in the study, all of which were serving UK military personnel. Upon providing informed consent, participants were invited to complete questionnaires to collect basic demographic information. They were then introduced to IMPACT and given the opportunity to practice using the tool and all the tasks. Following a short break, they were fitted with physiological recording equipment and went through a calibration process. A resting baseline was completed prior to the main experiment commencing. Participants completed two formats of each task: high and low task load. This in recognition that cognitive workload is deemed a determinant of cognitive performance (Hancock & Parasuraman, 1992) and therefore differences in task loading is likely to lead to observable differences in subjective, performance and psychophysiological behaviours. Once all conditions had been completed, participants were invited to complete supplementary questionnaires and a debrief interview. In total, the study lasted no longer than 3 hours, 20 minutes.

Data Handling and Processing

Given the volume of data collected in this study, a brief overview of the data handling and processing approaches for each measure is described in Table 2.

Table 2. Description of data handling and processing

<table>
<tr><th>Measure</th><th>Description</th></tr>
<tr><td>fNIRS</td><td>An Artinis Oxymon MKIII device was used to measure brain activity with data recorded at 50hz across 8 channels in a 2x7 channel split optode configuration. Data was processed using bespoke software called Oxysoft and exported in a Shared Near Infrared Spectroscopy Format) for onward filtering and processing in Python. In order to help identify activity within different areas of the prefrontal cortex, the following template was used;

<table>
<tr><th>Channel</th><th>Table Header</th><th>Diagram Numbering</th><th>Brodmann Area</th></tr>
<tr><td>R1B X T5</td><td>S5 x D4</td><td>1</td><td>45/46 (possible for part 47)</td></tr>
<tr><td>R1B X T4</td><td>S4 X D4</td><td>2</td><td>46/9</td></tr>
<tr><td>R3B X T4</td><td>S4 X D6</td><td>3</td><td>9</td></tr>
<tr><td>R3B X T6</td><td>S6 X D6</td><td>4</td><td>10</td></tr>
<tr><td>R4 X T6</td><td>S6 X D5</td><td>5</td><td>10/11</td></tr>
<tr><td>R4 X T5</td><td>S5 X D5</td><td>6</td><td>10/11</td></tr>
<tr><td>R4 X T4</td><td>S4 X D5</td><td>7</td><td>10</td></tr>
<tr><td>R3A X T3</td><td>S3 X D3</td><td>8</td><td>10</td></tr>
<tr><td>R3A X T1</td><td>S1 X D3</td><td>9</td><td>9</td></tr>
<tr><td>R1A X T1</td><td>S1 X D1</td><td>10</td><td>46/9</td></tr>
<tr><td>R1A X T2</td><td>S2 X D1</td><td>11</td><td>45/46 (possible for part 47)</td></tr>
<tr><td>R2 X T2</td><td>S2 X D2</td><td>12</td><td>10/11</td></tr>
<tr><td>R2 X T3</td><td>S3 X D2</td><td>13</td><td>10/11</td></tr>
<tr><td>R2 X T1</td><td>S1 X D2</td><td>14</td><td>10</td></tr>
</table>
</td></tr>
</table>

Measure	Description
	 View from above View from front 2 3 7 1 4 6 5 A) Left 9 10 14 8 11 13 12 B) Right B A
EDA	EDA was measured using the Shimmer3 GSR+ device. A series of 2x7 repeated measures ANOVA's were performed in SPSS for each task to determine whether there was a statistically significant difference in average EDA between the baseline, low task load and high task load conditions.
HRV	A Polar H10 Chest Strap was used to measure HRV. A series of repeated measures ANOVA's were performed in SPSS for each task to determine whether there was a statistically significant difference in average HRV between the baseline, low task load and high task load conditions.
Gaze metrics	Data collected from the eye-tracking unit was processed using Tobii Lab Pro software. All data was analysed in SPSS using paired samples t-tests.
Workload	All NASA-TLX questionnaires were completed on screen using a PyCharm script to enable ease of transfer in to SPSS. All data was then analysed using Wilcoxon signed ranks tests.
Performance	Data was recorded automatically by the IMPACT software and outputted in xlsx format for each participant. For each IMPACT task, associated Key Performance Indicators (KPIs) were calculated. All data was analysed in SPSS using Wilcoxon signed ranks tests.
Situational Awareness	All 3D SART questionnaires were completed on screen using a PyCharm script to enable ease of transfer in to SPSS. All data was then analysed using Wilcoxon signed ranks tests.

Results

Given the volume of data collected in this study, this section provides a high-level summary of the key findings for each measure.

fNIRS

Analysis of data suggests that there are significant changes in cortical activity across different areas of the brain for all IMPACT tasks as a consequence of manipulating task load. Figure 1 identifies which areas elicited significant changes in cortical activity during interaction with the individual IMPACT tasks. Broadly speaking, interaction with the IMPACT tasks activated brain regions typically associated with working memory, decision-making, language, intelligence, perception and attention (Carlén, 2017).

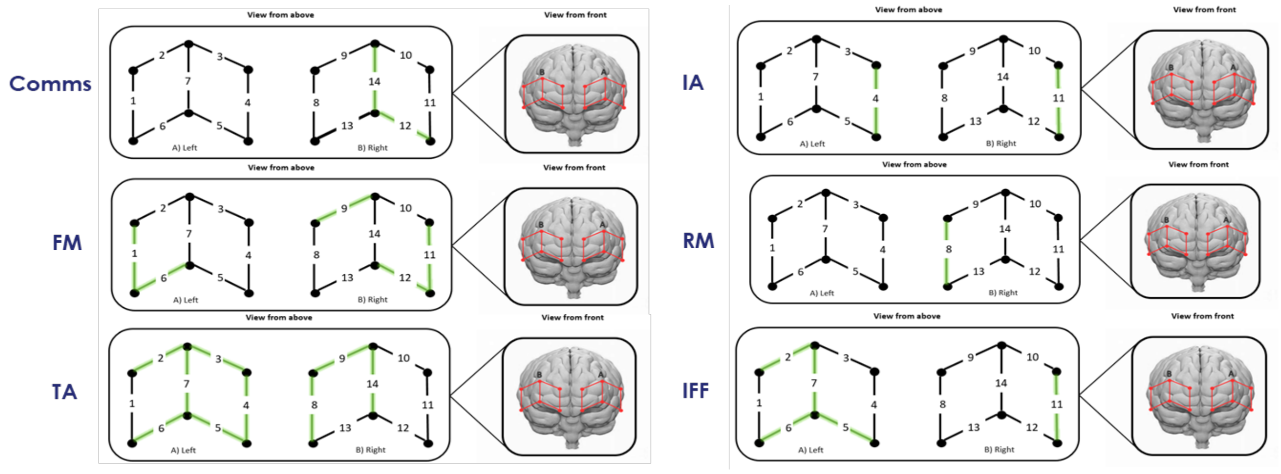


Figure 1: Areas showing significant differences in cortical activity for each task (shown by green lines).

Subjective workload

Analysis of the data relating to perceived workload suggests that increasing task load is coupled with a significant increase in perceived workload for all tasks with one exception – Immediate Action. This means that perceptually, participants did not notice any difference in their perceived workload during the two task load conditions for Immediate Action but the general trend indicates perceived workload is impacted by task load manipulation. Possible reasons for this may be attributable to the simplicity of the Immediate Action task and manner of presentation in the current study. Participants were required to respond to an alert via a single button press.

Subjective situational awareness

Analysis of the data relating to perceived situational awareness suggests that as task load increases, there is a significant reduction in perceived situational awareness for all tasks except for Immediate Action. This suggests that the IMPACT tool has task load variability sufficient to produce varying levels of perceived situational awareness.

Performance Metrics

Each IMPACT task had slightly different Key Performance Indicators (KPIs) but these broadly centred on accuracy of response, success rate and response time. Analysis of the data suggests that increasing task load led to significant decreases in performance across all tasks except for Immediate Action. This suggests that broadly speaking, the IMPACT tool has task load variability sufficient to induce different levels of performance.

Gaze Metrics

Pupil diameter was found to significantly differ between the high and low task load conditions for 4 out of the 6 IMPACT tasks (Formation Maintenance, Target Acquisition, Resource Management and Identify Friend or Foe). The general trend was that pupil diameter was significantly greater in the high task load condition. Given the strength of evidence within the academic literature suggesting that pupil diameter is linked to the level of cognitive load, a tentative conclusion is that this study appears to support the claim that IMPACT is capable of producing observable differences within eye movement data under different task load manipulations.

Findings for the other metrics used in this study were less conclusive with task load variability leading to significant differences in fixation duration in 2 out of the 6 tasks (Communications and

Formation Maintenance) and saccade duration in 2 out of 6 tasks (Formation Maintenance and Target Acquisition).

EDA

There was no significant difference in EDA between the high and low task load conditions. One possible reason for this is that the processing of data using 30-second epochs was too long to identify spikes in activity. An alternative explanation may be that the chosen task parameters to develop high and low task load where not sufficient. Given that the task parameters did lead to significant differences in perceived workload and situational awareness, this finding may suggest participants could have been pushed further particularly as the literature indicates EDA is a sensitive psychophysiological index of changes in ANS arousal, which can be used to infer emotional and cognitive state.

HRV

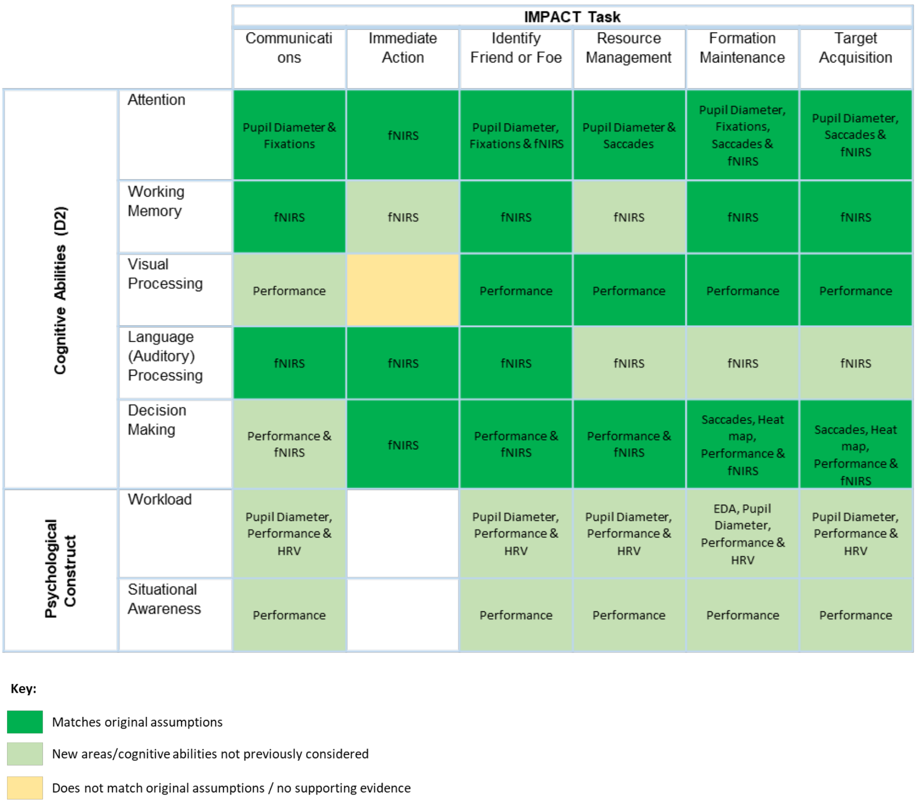
Analysis of HRV data suggests that whilst the manipulation of task load did cause an underlying change in cardiac activity, differences between the high and low task load conditions were not significant. Again, this may have been attributable to chosen task load parameters but more research is needed to help identify human limits whilst interacting with the tool.

Discussion

When selecting measures to use in human science experimentation, it is important that they represent the constructs they claim to measure. Thus, validity and reliability are critical considerations when selecting and interpreting results (Salmon et al., 2009). This paper provides an interesting insight in to the potential of the IMPACT tool to elicit different behavioural outcomes based on task load manipulation using a range of subjective, psychophysiological and performance measures. The inclusion of fNIRS has been particularly powerful in confirming what areas of the brain are active during interaction with the tool. Whilst this goes some way in validating the assumptions made about the tool in previous work (Sabine, 2022; Sabine & Thompson, 2024), it is recognised that determining psychometric validity is an ongoing process involving the accumulation of evidence over time, settings, and samples to build a scientifically sound foundation (Cronbach & Meehl, 1995; Rickards et al., 2012). Further testing using IMPACT is required to (i) fully understand its capabilities and (ii) build upon this evidence base to both build and maintain user confidence in the ability of the tool to accurately measure what it claims to measure.

Cronbach & Meehl (1995) recognise the importance to appreciate that the construct validation process involves an ongoing, iterative process in which new findings and new theories can clarify and alter the existing understanding of existing theories. This too is supported by Vitoratou & Pickles (2017) who argue that validity assessments are always subject to new findings and understanding. This study represents an important first step in understanding the potential of the tool itself to stimulate alternative cognitive areas in the ways anticipated. More research is needed to add to this body of evidence. Table 3 presents a summary of the cognitive areas implicated for each IMPACT task based on the evidence collected so far. When compared to the mapping of tasks conducted in previous work (Sabine, 2022; Sabine & Thompson, 2024), it is clear that IMPACT performs in many of the ways designed and anticipated for. Even so, this should be viewed as a starting point in understanding the capabilities of the tool to elicit different behavioural outcomes. The table should also be updated as new data becomes available and is used more widely by different participant samples and populations. It does appear however that the IMPACT tool has real potential to become a valuable tool for human sciences practitioners within the defence community.

Table 3. Evidence-based mapping of IMPACT tasks to cognitive areas and psychological constructs.



		IMPACT Task					
		Communications	Immediate Action	Identify Friend or Foe	Resource Management	Formation Maintenance	Target Acquisition
Cognitive Abilities (D2)	Attention	🟩 Pupil Diameter & Fixations	🟩 fNIRS	🟩 Pupil Diameter, Fixations & fNIRS	🟩 Pupil Diameter & Saccades	🟩 Pupil Diameter, Fixations, Saccades & fNIRS	🟩 Pupil Diameter, Saccades & fNIRS
	Working Memory	🟩 fNIRS	▨ fNIRS	🟩 fNIRS	▨ fNIRS	🟩 fNIRS	🟩 fNIRS
	Visual Processing	▨ Performance	🟨	🟩 Performance	🟩 Performance	🟩 Performance	🟩 Performance
	Language (Auditory) Processing	🟩 fNIRS	🟩 fNIRS	🟩 fNIRS	▨ fNIRS	▨ fNIRS	▨ fNIRS
	Decision Making	▨ Performance & fNIRS	🟩 fNIRS	🟩 Performance & fNIRS	🟩 Performance & fNIRS	🟩 Saccades, Heat map, Performance & fNIRS	🟩 Saccades, Heat map, Performance & fNIRS
Psychological Construct	Workload	▨ Pupil Diameter, Performance & HRV		▨ Pupil Diameter, Performance & HRV	▨ Pupil Diameter, Performance & HRV	▨ EDA, Pupil Diameter, Performance & HRV	▨ Pupil Diameter, Performance & HRV
	Situational Awareness	▨ Performance		▨ Performance	▨ Performance	▨ Performance	▨ Performance

Key:

🟩 Matches original assumptions

▨ New areas/cognitive abilities not previously considered

🟨 Does not match original assumptions / no supporting evidence

Acknowledgements

Thanks to Dstl for providing support and for access to the IMPACT tool during the period of research. IMPACT is MOD asset and is subject to crown Copyright.

References

Carlén, M. (2017). What constitutes the prefrontal cortex? *Science, 358*(6362), 478-482.

Comstock, J., & Arnegard, R. (1992). The multi-attribute task battery for human operator workload and strategic behavior research: 115.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Hancock, P. A., & Parasuraman, R. (1992). Human factors and safety in the design of intelligent vehicle-highway systems (IVHS). *Journal of Safety Research, 23*(4), 181-198.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139-183.

Izzetoglu, K., & Richards, D. (2019). Human performance assessment: Evaluation of wearable sensors for monitoring brain activity. In *Improving Aviation Performance through Applying Engineering Psychology* (pp. 163-180). CRC Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement,* 50(1), 1-73.

Kim, H. G., Cheon, E. J., Bai, D. S., Lee, Y. H., & Koo, B. H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation, 15*(3), 235-245.

Messick, S. (1987). Validity. *ETS Research Report Series, 1987*(2), i-208.

Mulder, G., & Mulder-Hajonides van der Meulen, W. R. E. H. (1973). Mental load and the measurement of heart rate variability. *Ergonomics, 16*(1), 69-83.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill

Rickards, G., Magee, C., & Artino Jr, A. R. (2012). You can't fix by analysis what you've spoiled by design: developing survey instruments and collecting validity evidence. *Journal of graduate medical education*, 4(4), 407-410.

Sabine, G. (2022). Design and Development: Interactive Measures of Performance and Assessment of Cognitive Tasks (IMPACT) Software Tool. Reference: *DSTL/TR139745*, Defence Science Technology Laboratories.

Sabine, G., & Thompson, D. J. (2024). Introducing IMPACT; Design and Development of a Military Orientated Cognitive Task Battery. *Ergonomics and Human Factors 2024*. Kenilworth. 24/04/2024.

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring situation awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics, 39*(3), 490-500.

Shakouri, M., Ikuma, L. H., Aghazadeh, F., & Nahmens, I. (2018). Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: the case of highway work zones. *International Journal of Industrial Ergonomics, 66*, 136-145.

Smith, G. T. (2005). On construct validity: issues of method and measurement. *Psychological assessment*, *17*(4), 396.

Tatlock, K., Delaney, S., Leahy, D., Croft, D., Brennen, S., Doherty, V., Fisher, N., McNamara, H., & Bowyer, S. (2015). “TIN 2.035 – Challenge 1: Specifying the Cognitive Requirements of People in Systems”. Reference: *UH-DHCSTC_I2_H_T2_035_1/003*, Defence Human Capability Science and Technology Centre.

Taylor, M. (1990). Situation awareness rating technique (SART): The development of a tool for aircrew systems design. France: Neuilly sur-Seine, NATO-AGARD-CP-478.

Trochim, W. M. K. (2001). *The Research Methods Knowledge Base*. (2nd ed.), Cincinnati, OH: Atomic Dog Publishing.

Vitoratou, S., & Pickles, A. (2017). A note on contemporary psychometrics. *Journal of Mental Health, 26*(6), 486-488.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: two simple measures. *Journal of personality and social psychology*, *84*(3), 608.

Yu, D., & Sun, S. (2020). A systematic exploration of deep neural networks for EDA-based emotion recognition. *Information*, *11*(4), 212.