The Development, and Day-to-Day Variation, of a Military-Specific Auditory N-Back Task and Shoot-/Don't-Shoot Task

Vine C.A.J [1*], Coakley S.L. [1,2], Myers S.D. [1], Blacker S.D. [1], Runswick O.R.[1,3]

[1]Occupational Performance Research Group, Institute of Sport, Nursing and Allied Health, University of Chichester, Chichester, UK, [2] Faculty of Sport, Allied Health and Performance Science, St Mary's University, Twickenham, [3] Institute of Psychiatry, Psychology & Neuroscience, Kings College London, * corresponding author.

ORCID:

Christopher Vine – 0000-0002-3592-9894
Sarah Coakley – 0000-0002-9314-1392
Stephen Myers - 0000-0002-7855-4033
Sam Blacker – 0000-0003-3862-3572
Oliver Runswick – 0000-0002-0291-9059

Address for correspondence:

Christopher Vine,
Institute of Sport, Nursing and Allied Health
University of Chichester,
Chichester,
PO19 6PE.
Tel: +44 (0) 1243 796231
Email: c.vine@chi.ac.uk

The Development, and Day-to-Day Variation, of a Military-specific Auditory N-Back Task and Shoot-/Don’t-Shoot Task

During military operations, soldiers are required to successfully complete numerous physical and cognitive tasks concurrently. Understanding the typical variance in research tools that may be used to provide insight into the interrelationship between physical and cognitive performance is therefore highly important. This study assessed the inter-day variability of two military-specific cognitive assessments; a Military-Specific Auditory N-Back Task (MSANT) and a Shoot-/Don’t-Shoot Task (SDST) in 28 participants. Limits of agreement ± 95% Confidence Intervals, Standard Error of the Mean, and Smallest Detectable Change were calculated to quantify the typical variance in task performance. All parameters within the MSANT and SDST demonstrated no mean difference for trial visit in either the seated or walking condition, with equivalency demonstrated for the majority of comparisons. Collectively, these data provided an indication of the typical variance in MSANT and SDST performance, whilst demonstrating that both assessments can be used during seated and walking conditions.

Keywords: Occupational, Performance, External Validity, Decision Making

Funding details: The authors declare no funding was received for this research.

Disclosure statement: The authors report there are no competing interests to declare.

Data availability statement: The raw data for this project can be viewed at: https://osf.io/jekv8/.

Acknowledgements: The authors would like to thank the participants for their participation in the current project.

Contribution Statement: CV, SC, SM, SB, OR - Research concept and study design. CV - data collection, data and statistical analyses, writing of the initial manuscript. CV, SC, SM, SB, OR - reviewing/editing final manuscript.

Word Count: 1992

Introduction

During military operations, personnel are required to maintain performance in both their role-specific physical tasks (e.g. load carriage), and in corresponding cognitive tasks (e.g. decision making, and communication) (Crawford et al., 2017; Scribner, 2016). Failure to maintain performance, in either domain, can result in reduced combat readiness and decreased operational performance (Crawford et al., 2017; Vrijkotte et al., 2016). Consequently, there is growing interest in the relationship between military-specific physical activity and cognitive performance within military operators (Armstrong et al., 2022; Bhattacharyya et al., 2017; Eddy et al., 2015; Giles et al., 2019; Kobus et al., 2010; Nibbeling et al., 2014; Son et al., 2019, 2022; Vine et al., 2021). Despite this interest, the methodologies and approaches used to investigate this relationship have differed considerably, particularly concerning the assessment of cognitive performance.

Based on the assessment tools used to date, and the visual and auditory requirements of soldiers, two assessment tools were developed; A Military Specific Auditory N-Back Task (MSANT), and a Shoot-/Don't-Shoot (SDST). The former, used phonetically described pairs of letters, and represented aspects of military radio communications, whilst the latter represented aspects of any military scenario where visual search and inhibition are required (e.g. assaulting an enemy position or operations in built up areas). The current study, therefore, aimed to detail the methodology of the MSANT and SDST, along with quantifying the typical day-to-day variability of both assessment tools under seated and walking condition. The investigation did not seek to investigate the influence of physical fatigue or dual-tasking on the performance of these assessment tools.

Methods

The full methods for this project are available in the supplementary material, with the raw data available at: https://osf.io/jekv8/. Briefly, the study comprised of two elements. First, the day-to-day variability of the MSANT and SDST was assessed in a seated condition on three separate occasions (Part 1). This was chosen due to the large variability in potential application of these assessment tools in future projects. Second, within a sub-sample of the study population, the day-to-day variability of the MSANT and SDST was assessed during a 10-minute walking activity, on three separate occasions (Part 2). Whilst a matched study population, for this part of the study would have been optimal, given the time required for this portion of the study (a result of the necessity to reach a physiological steady-state before conducting the test, and the recovery period required between each walking bout to prevent the onset of fatigue), a sub-sample approach was instead chosen. Physiological steady state refers to the stabilisation in the physiological responses to exercise (e.g. increases in heart rate). Without this stability, variability in cognitive performance could be induced as a consequence of adapting the exercise stimulus opposed to just reflecting the typical variation in test performance.

All laboratory visits were separated by a minimum of 48 hours, and participants were required to arrive in a fed and hydrated state having avoided caffeine for a minimum of three hours. Study visits were completed at approximately the same time of day (± 2 hours) to control for the potential effect of circadian rhythm on test performance. All participants were recruited from the university population (all were students or from academic positions), spoke fluent English, and had self-declared normal, or corrected to normal vision.

Twenty-eight participants volunteered for Part 1 of the study (14 male, 14 female, age [mean ± SD] 27.3 ± 4.3 y) and 12 participants for Part 2 (6 male, 6 female, age 28.4

± 3.5 y). Sample size for Part 1 was calculated using an A Priori power calculation (G Power; version 3.1.9.4) (Prajapati et al., 2010). For the seated portion of the investigation, 28 participants were required to a moderate effect size (*f* = 0.25), with a statistical power of 80%, and an alpha level of 0.05, based upon a correlation coefficient of *r* = 0.5 (identified from initial pilot testing). A moderate effect size (Cohen, 1988) was selected based on the combination of effect sizes reported in previous investigations, utilising similar cognitive assessment tools (Eddy et al., 2015), and the anticipated smallest effect size of interest to military policymakers. The sub-sample size was designed to represent the typical size (and therefore likely variation) of study populations within this research area (e.g. Bhattacharyya et al., 2017; Crowell et al., 1999; Eddy et al., 2015). Ethical approval was provided by the Institution's Research Ethics Committee, with written consent obtained from all participants.

Cognitive Assessments

The MSANT involved identifying a pair of phonetically described letters two previous to an auditory tone (i.e. 2-back). During the seated condition, participants recorded their answers, whilst during walking trials, participants were required to relay their answers verbally which were recorded on their behalf. The SDST was designed to be a visual search and inhibition task similar to those tasks previously employed within the literature (Armstrong et al., 2022; Eddy et al., 2015; Kobus et al., 2010). The assessment involved responding appropriately to targets and non-targets. Participants were instructed to place equal importance on both response time and accuracy. For the SDST there was a 2:1 ratio between targets and non-targets.

For Part 1, during the first visit, participants were familiarised (two full trial completions of each assessment) with the MSANT and SDST, in a randomised counterbalanced order. For the second, and third visits, participants completed the MSANT and SDST in the same randomised counterbalanced order. For Part 2, a sub-sample of 12 participants completed three additional laboratory visits completing the SDST and MSANT whilst walking on a treadmill. Again the MSANT and SDST were completed in a randomised order. All tests were completed with 10 minutes of seated rest between trials to negate the influence of physical fatigue. To enable a physiological steady-state to occur, participants completed five minutes of walking before the commencement of the cognitive assessments. For all walking trials, participants walked on a motorised treadmill (6.5 km·h^{-1}, 1% gradient) at a load carriage speed representing a typical 'enemy contact' speed (Armstrong, Ward, Lomax, Tipton, & House, 2019).

Statistical Analysis

Data were principally analysed using JASP (JASP, 2020; v0.14.1). For normally distributed data, a one-way ANOVA was employed to identify whether a likely main effect of assessment time point was apparent. Effect sizes are presented as Omega squared (ω^2) (Levine & Hullett, 2002). For non-normally distributed data a Friedman's test was employed with effect sizes presented using Kendall's W. Holm-Bonferroni adjusted pairwise comparisons, and pairwise comparisons using Conover's test were made *post-hoc* as appropriate. For key assessment variables, equivalency between trials was calculated using the Two One-Sided Test approach (Lakens et al., 2018). Based upon the A Priori sample size calculation, *d* = 0.5 was employed as the smallest effect size of interest. To describe the typical variation in assessment parameters between trials, Limits of Agreement (LoA) ± 95% Confidence Intervals (CI), Standard Error of the Mean (SEM), and Smallest Detectable Change (SDC) values were calculated (Hopkins, 2000; Ludbrook, 2010; van Kampen et al., 2013).

Results

Descriptive statistics are presented in Table 1, with day-to-day variation descriptors reported in Table 2. One participant was removed from the analysis, due to being more than two SDs outside the remainder of the data set.

Seated Performance

MSANT

There was no likely main effect for time for total correct response ($\chi^2_{(4)} = 4.531$, $p = 0.361$, Kendall's W = 0.492), or combined correct responses ($\chi^2_{(4)} = 3.856$, $p = 0.426$, Kendall's W = 0.488), however, a likely main effect for time was evident for partial correct responses ($\chi^2_{(4)} = 11.846$, $p = 0.019$, Kendall's W = 0.426). For the key variable of combined correct responses, the comparison between trial 1 vs trial 2 was both statistically equivalent ($W_{(25)} = 64$, $p = 0.002$) and not statistically different ($W_{(25)} = 70$, $p = 0.938$). Similarly, trial 2 vs trial 3 were both statistically equivalent ($W_{(25)} = 20$, $p = 0.06$) and not statistically different. Likewise trial 1 vs trial 3, were also both statistically equivalent ($W_{(25)} = 50$, $p = 0.032$) and not statistically different.

SDST

There was no likely main effect for time on either shoot correct ($\chi^2_{(4)} = 4.00$, $p = 0.406$, Kendall's W = 0.175), don't-shoot correct ($\chi^2_{(4)} = 3.069$, $p = 0.546$, Kendall's W = 0.482), total correct ($\chi^2_{(4)} = 3.375$, $p = 0.497$, Kendall's W = 0.471), and average response time ($F_{(2.981,77.515)} = 1.035$, $p = 0.382$, Ꝏ^2= 0.001). There was however, a main effect for time in the ASTO parameter ($F_{(4,104)} = 7.037$, $p < 0.001$, Ꝏ^2= 0.089). Importantly, the sole noteworthy difference, occurred between familiarisation 1 and trial

3 ($t_{(26)}$ = 4.855, p < 0.001, d = 0.756) suggesting no discernible difference were likely between performances in the three experimental trials, following two familiarisation trials. For the key variable of total correct responses trial 1 vs trial 2, trial 1 vs trial 3, and trial 2 vs trial 3, were both statistically equivalent (1 vs 2: $W_{(26)}$ = 93, p = 0.011; 1 vs 3: $W_{(26)}$ = 61, p = 0.047; 2 vs 3: $W_{(26)}$ = 41, p = 0.040) and not statistically different. For the other key variable of ASTO all comparisons were likely neither statistically equivalent (1 vs 2: $t_{(26)}$ = -1.701, p = 0.050; 2 vs 3: $t_{(26)}$ = -0.127, p = 0.45; 1 vs 3: $t_{(26)}$ = 0.287, p = 0.612), nor statistically different.

Walking Performance

MSANT

As with seated MSANT performance, there was no likely effect of time on total correct responses ($\chi^2_{(2)}$ = 1.000, p = 0.607, Kendall's W = 0.568), partial correct responses ($\chi^2_{(2)}$ = 1.280, p = 0.527, Kendall's W = 0.541) and combined correct responses ($\chi^2_{(2)}$ = 1.000, p = 0.607, Kendall's W = 0.582). For the key variable of combined correct responses trials 1 vs 2, and trials 2 vs 3 were statistically equivalent (1 vs 2: $W_{(11)}$ = 12, p = 0.017; 2 vs 3: $W_{(11)}$ = 13, p = 0.020) and not statistically different. Conversely, trial 1 vs 3 was neither statistically equivalent, nor statistically different.

SDST

Again there were no likely effects of time, on shoot correct responses ($\chi^2_{(2)}$ = 4.800, p = 0.091, Kendall's W = 0.449), don't-shoot correct responses ($\chi^2_{(2)}$ = 2.480, p = 0.289, Kendall's W = 0.672), total correct responses ($\chi^2_{(2)}$ = 3.161, p = 0.206, Kendall's W = 0.741), response times ($F_{(2,22)}$ = 2.880, p = 0.077, η_G^2= 0.018), and ASTO ($F_{(2,22)}$ = 2.713, p = 0.088, η_G^2= 0.042). For the key variable of total correct responses all comparisons were neither statistically equivalent (1 vs 2: $W_{(11)}$ = 6, p = 0.096; 1 vs 3:

$W_{(11)} = 0$, $p = 0.093$; 2 vs 3: $W_{(11)} = 14$, $p = 0.084$), nor statistically different. Similarly, For the other key variable of ASTO all comparisons all comparisons were likely neither statistically equivalent (1 vs 2: $t_{(11)} = 0.127$, $p = 0.549$; 2 vs 3: $t_{(11)} = -1.205$, $p = 0.127$; 1 vs 3: $t_{(11)} = 0.787$, $p = 0.776$), nor statistically different.

Discussion

This study has described the methods of two military-specific cognitive assessment tools (MSANT and SDST), and quantified their typical day-to-day variability. These data provide typical magnitudes of variance for the key assessment parameters. Whilst no likely performance differences were observed across the experimental measurement points, not all walking comparisons were statistically equivalent; suggesting additional data are required before this assertion is made, for the given equivalency bounds. It should however be noted that borderline statistically significant results may become non-significant were correction for multiple testing is utilised. The current investigation has also demonstrated the suitability of these assessment tools for use during military-specific physical activity within a laboratory setting.

Before this investigation, the day-to-day performance variation in any military-specific cognitive assessments had not been quantified. This is an issue for several reasons, including the translational ability of research findings to the ‘real world’ (Close et al., 2019), and also for methodological decision making (e.g. sample size calculations). Moreover, with military operations rarely conducted in isolation, information on inter-test performance is highly relevant to research investigating sequential or repeated bout performance. The comparison between seated and walking performance was not a research question of interest in the current study; particularly given that deficits in cognitive performance are typically observed after ~30 minutes of military activity (e.g.

Eddy et al., 2015; Giles et al., 2019). However, observationally, the typical variation in performance between trials appears similar between seated and walking conditions.

Familiarising participants with assessment tools is critical for research, particularly when time limitations may inhibit access to study participants (e.g. military populations). Collectively, the current study's data demonstrates that beyond two full seated trials, a continued improvement in performance was not likely apparent, suggesting this familiarisation length is sufficient to minimise possible learning effects.

Several limitations exist with the current investigation, including the use of a civilian population, and the limited walking sub-sample size. As acknowledged previously the smaller sub-sample size was chosen for largely practical reasons, although it does match many studies within this area; highlighting issues with underpowered investigations. Future research should attempt to pair reliable and applied cognitive tasks (such as those described herein) with operationally relevant and appropriate physical activity. This in turn will support enhanced applied research as well as enabling a greater focus to be placed on developing mitigation strategies where the greatest mission impact can be obtained.

References

Armstrong, N., Smith, S., Risius, D., … D. D.-B. M., & 2022, U. (2022). Cognitive performance of military men and women during prolonged load carriage. *BMJ Military Health.*

Armstrong, N., Ward, A., Lomax, M., Tipton, M. J., & House, J. R. (2019). Wearing body armour and backpack loads increase the likelihood of expiratory flow limitation and respiratory muscle fatigue during marching. *Ergonomics*, *62*(9), 1181–1192.

Bhattacharyya, D., Pal, M., Chatterjee, T., & Majumdar, D. (2017). Effect of load carriage and natural terrain conditions on cognitive performance in desert environments. *Physiology & Behavior, 179*, 253–261.

Close, G. L., Kasper, A. M., & Morton, J. P. (2019). From paper to podium: quantifying the translational potential of performance nutrition research. *Sports Medicine*, *49*(1), 25–37.

Cohen, J. (1988). The effect size index: d. In *Statistical power analysis for the behavioral sciences*.

Crawford, C., Teo, L., Lafferty, L., Drake, A., Bingham, J. J., Gallon, M. D., O'connell, M. L., Chittum, H. K., Arzola, S. M., & Berry, K. (2017). Caffeine to optimize cognitive function for military mission-readiness: a systematic review and recommendations for the field. *Nutrition Reviews*, *75*(suppl_2), 17–35.

Crowell, H. P., Krausman, A. S., Harper, W. H., Faughn, J. A., & Sharp, M. A. (1999). *Cognitive and Physiological Performance of Soldiers While They Carry Loads Over Various Terrains*. Army research lab aberdeen proving ground md.

Eddy, M. D., Hasselquist, L., Giles, G., Hayes, J. F., Howe, J., Rourke, J., Coyne, M., O'Donovan, M., Batty, J., & Brunyé, T. T. (2015). The effects of load carriage and physical fatigue on cognitive performance. *PloS One*, *10*(7), e0130817.

Giles, G. E., Hasselquist, L., Caruso, C., & Eddy, M. D. (2019). Load Carriage and Physical Exertion Influence Cognitive Control in Military Scenarios. *Medicine and Science in Sports and Exercise*.

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, *30*(1), 1–15.

JASP. (2020). *JASP (Version 0.14.1)[Computer software]*.

Kobus, D. A., Brown, C. M., Wu, L., Robusto, K., & Bartlett, J. (2010). *Cognitive Performance and Physiological Changes under Heavy Load Carriage*. Pacific Science and Engineering Group Inc, San Diego, CA.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*(4), 612–625.

Ludbrook, J. (2010). Confidence in Altman-Bland plots: A critical review of the method of differences. *Clinical and Experimental Pharmacology and Physiology*, *37*(2), 143–149.

Nibbeling, N., Oudejans, R. R. D., Ubink, E. M., & Daanen, H. A. M. (2014). The effects of anxiety and exercise-induced fatigue on shooting accuracy and cognitive performance in infantry soldiers. *Ergonomics*, *57*(9), 1366–1379.

Prajapati, B., Dunne, M., & Armstrong, R. (2010). Sample size estimation and power analysis. *Optometry Today*, *16*, 123–132.

Scribner, D. R. (2016). Predictors of shoot–don't shoot decision-making performance:

An examination of cognitive and emotional factors. *Journal of Cognitive Engineering and Decision Making*, *10*(1), 3–13.

Son, M., Hyun, S., Beck, D., Jung, J., & Park, W. (2019). Effects of backpack weight on the performance of basic short-term/working memory tasks during flat-surface standing. *Ergonomics*, *62*(4), 548–564.

Son, M., Jung, J., Hwang, D., Beck, D., & Park, W. (2022). The effect of backpack weight on the performance of basic short-term/working memory tasks while walking along a pre-determined route. *Ergonomics*, 1–23.

van Kampen, D. A., Willems, W. J., van Beers, L. W. A. H., Castelein, R. M., Scholtes, V. A. B., & Terwee, C. B. (2013). Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *Journal of Orthopaedic Surgery and Research*, *8*(1), 40.

Vine, C. A. J., Myers, S. D., Coakley, S. L., Blacker, S. D., & Runswick, O. R. (2021). Transferability of Military-Specific Cognitive Research to Military Training and Operations. *Frontiers in Psychology*, *12*, 386.

Vrijkotte, S., Roelands, B., Meeusen, R., & Pattyn, N. (2016). Sustained military operations and cognitive performance. *Aerospace Medicine and Human Performance*, *87*(8), 718–727.

Table 1. Descriptive statistics for cognitive assessments (mean ± SD [range]) during seated (S) and walking (W) conditions.

Task (condition)	**Parameter**	**Experimental Trial**				
		FAM 1	*FAM 2*	*Trial 1*	*Trial 2*	*Trial 3*
SDST (S)	*Total Correct (%)*	96.4 ± 3.3 [86.1-100.0]	97.3 ± 2.7 [91.7-100.0]	96.5 ± 3.0 [88.9-100.0]	96.3 ± 3.2 [88.9-100.0]	97.1 ± 3.2 [83.3-100.0]
	RT (ms)	579 ± 58 [490-684]	574 ± 57 [478-683]	562 ± 57 [472-704]	550 ± 51 [450-639]	528 ± 43 [433-655]
	ASTO (ms·cr^{-1})	16.7 ± 1.6 [14.1-20.4]	16.4 ± 1.7 [13.7-19.2]	16.2 ± 1.8 [13.7-19.9]	15.9 ± 1.4 [13.2-18.3]	15.1 ± 1.4 [12-18.7]
SDST (W)	*Total Correct (%)*			94.9 ± 5.3 [80.6-100]	96.1 ± 3.8 [88.9-100]	96.5 ± 5.4 [80.6-100]
	RT (ms)			594 ± 70 [496-678]	575 ± 69 [457-661]	566 ± 69 [451-666]
	ASTO (ms·CR^{-1})			17.4 ± 1.4 [15-19.4]	16.6 ± 1.6 [13.9-18.4]	16.3 ± 1.9 [13.3-19]
MSANT (S)	*Total Correct (%)*	87.7 ± 15 [50-100]	88.5 ± 16.7 [30-100]	90.4 ± 14.6 [40-100]	90.8 ± 16 [30-100]	94.2 ± 9.5 [60-100]
	Combined Score (%)	91 ± 11.3 [60-100]	91.4 ± 12.8 [46.7-100]	92.9 ± 10.8 [56.7-100]	92.7 ± 12.3 [46.7-100]	95.1 ± 7.7 [70-100]
MSANT (W)	*Total Correct (%)*			93.3 ± 8.9 [70-100]	95 ± 10 [70-100]	94.2 ± 9 [80-100]
	Combined Score (%)			95.3 ± 6.7 [76.7-100]	96.1 ± 7.9 [76.7-100]	95.8 ± 6.5 [83.3-100]

Where: S, seated; W, walking; FAM, familiarisation; RT, response time; ASTO, accuracy-speed trade-off; CR, correct response; SDST, shoot/don't-shoot task; MSANT, military-specific auditory n-back task. Greyed areas denote data that wasn't collected due to the seated condition acting as the familiarisation for the walking condition.

Table 2. Descriptive statistics for cognitive assessments (mean ± SD) during seated (S) and walking (W) conditions.

		Trial 1 vs 2			**Trial 2 vs 3**			**Trial 1 vs 3**		
Task (condition)	**Parameter**	*Mean Bias ± 95% CI*	*SEM*	*SDC*	*Mean Bias ± 95% CI*	*SEM*	*SDC*	*Mean Bias ± 95% CI*	*SEM*	*SDC*
SDST (S)	*Total Correct (%)*	0.2 ± 5.1	2.4	6.7	-0.8 ± 4.8	2.3	6.4	-0.6 ± 6.1	2.9	8.0
	RT (ms)	12 ± 88	42	116	22 ± 65	31	87	34 ± 84	40	111
	ASTO (ms·CR^{-1})	0.3 ± 2.8	1.3	3.7	0.7 ± 2.3	1.1	3.1	1.1 ± 2.9	1.4	3.8
SDST (W)	*Total Correct (%)*	-1.2 ± 6.2	2.7	7.5	-0.5 ± 8.3	3.6	10.1	-1.6 ± 4.1	1.8	4.9
	RT (ms)	19 ± 60	26	73	9 ± 66	29	80	28 ± 75	33	91
	ASTO (ms·CR^{-1})	0.8 ± 2.3	1.0	2.8	0.3 ± 3.1	1.4	3.8	1.0 ± 2.3	1.0	2.8
MSANT (S)	*Total Correct (%)*	-0.4 ± 18.5	8.8	24.5	-3.5 ± 13.2	6.3	17.5	-3.8 ± 15.2	7.2	20.0
	Combined Score (%)	0.3 ± 14.5	6.9	19.1	-2.4 ± 11.2	5.4	14.8	-2.2 ± 11.9	5.7	15.7
MSANT (W)	*Total Correct (%)*	-1.7 ± 13.5	5.9	16.4	0.8 ± 16.1	7.0	19.5	-0.8 ± 21.2	9.3	25.7
	Combined Score (%)	-0.8 ± 10.3	4.5	12.5	0.3 ± 11.8	5.2	14.3	-0.6 ± 15.8	6.9	19.1

Where: S, seated; W, walking; RT, response time; ASTO, accuracy-speed trade-off; CR, correct response; SDST, shoot/don't-shoot task; MSANT, military-specific auditory n-back task; SEM, standard error of the mean; SDC, smallest detectable change; CI, confidence intervals.

Supplementary Material – Detailed Methodology

Study Overview

The study comprised of two distinct elements. First, the reliability of the MSANT and SDST was assessed in a seated condition on three separate occasions (Part 1). This was due to the large variability in the way in which the assessment tools may be utilised in future projects. Second, within a sub-sample of the study population, the reliability of the MSANT and SDST was assessed during a 10-minute military-specific walking activity (6.5 km·h^{-1}; 1% gradient), on three separate occasions (Part 2). Whilst a matched study population, for this part of the study would have been optimal, given the time required for this portion of the study (a result of the necessity to reach a physiological steady-state before conducting the test, and the recovery period required between each walking bout to prevent the onset of fatigue), a sub-sample approach was instead chosen. For the subsample of the study population that completed Part 2, the familiarisation in Part 1 acted as the familiarisation for this part as well. For all study visits, participants were required to arrive in a fed and hydrated state having avoided caffeine for a minimum of three hours before their laboratory visit. Study visits were completed at approximately the same time of day (± 2 hours) to control for the potential effect of circadian rhythm on test performance.

Participants

Twenty-eight participants volunteered for Part 1 of the study (14 men, 14 women, age [mean ± SD] 27.3 ± 4.3 y) and 12 participants for Part 2 (6 men, 6 women, age 28.4 ± 3.5 y). The study sample size for part 1 was calculated using an a priori power calculation (G Power; version 3.1.9.4) as detailed by Prajapati, Dunne, & Armstrong (2010). A sample size, of 28 was identified, for the seated portion of the investigation, as the number required to identify a moderate effect size (f = 0.25), with a statistical power of 80%, and an alpha level of 0.05, based upon a correlation coefficient between repeated test performances of r = 0.5 (identified from initial pilot testing). This effect size was chosen as the smallest effect size of interest to military policymakers, based on data from both pilot testing and previous investigations utilising similar cognitive assessment tools (Eddy et al., 2015). The sub-sample population size was designed to represent the typical size (and therefore likely variation) of study populations within this research area (e.g.

Bhattacharyya et al., 2017; Crowell et al., 1999; Eddy et al., 2015). On the participants' first visit to the laboratory written consent was provided following a written and verbal brief of the study requirements and methodologies. Ethical approval was provided by the Institution's Research Ethics Committee. All procedures were conducted in accordance with the declaration of Helsinki.

Cognitive Assessments

Design decisions were always made in favour of military relevance over 'typical' cognitive research norms. For example, stimuli in the SDST were not of equal spatial frequency, as employed in previous research (Kobus et al., 2010), but instead, were of individuals adopting realistic stances that would require specific responses, as used previously within the literature (Armstrong et al., 2022; Eddy et al., 2015; Nibbeling et al., 2014).

MSANT

The MSANT was developed to mimic aspects of coded military radio traffic, with stimuli comprising of letter pairs, described phonetically, using the International Radiotelephony Spelling Alphabet (International Civil Aviation Organization, 2019). Each letter within a pair was separated by 0.4 s, and each pair was separated by 2 s. After a random number of letter pairs (3-7 pairs), an auditory tone (0.25 s, 1000 Hz) was sounded and the participant was required to identify the pair of letters described two previous to the auditory tone (i.e. 2-back). The auditory tone occurred 1 s after the last pairing of that stimuli string. In line with previously employed n-back assessments, each test lasted approximately 5 minutes; depending on letter combinations. Each MSANT contained 100 letter stimuli (Kazemi et al., 2018) and required 10 responses.

Letter stimuli were generated using online speech generation software (www.fromtexttospeech.com) and compiled into a single audio track using an open-source digital audio editing software (Audacity® *v2.3*, Audacity®, USA). Speech generation variables were set to 'British English', male voice, and medium for the speech speed. All letter stimuli were randomly selected using an online random number generator (Research Randomiser; https://www.randomizer.org/). The letter 'F' was excluded due to the lack of clarity in generated audio stimuli. For both seated and walking conditions participants received the auditory information via headphones at a standardised volume.

During the seated condition, participants recorded their answers, whilst during walking trials, participants were required to relay their answers verbally to be recorded on their behalf. Whilst this approach may have been less than optimal for action fidelity, due to a difference in response modes, as the study was not designed to compare walking and seated conditions, the practicalities of this approach made this approach preferential.

SDST

The SDST was designed to be a visual search and inhibition task similar to those tasks previously employed within the literature (Armstrong et al., 2022; Eddy et al., 2015; Kobus et al., 2010). The urban scene depicted a derelict warehouse (Figure 1), with 12 possible target locations (windows); comprising of 6 on the ground floor and 6 on the first floor. There were no stimuli on the gantry level. Using a calibration stick, the warehouse within the scene was measured to be 9.18 m high and 18.42 m in width. Windows containing the target stimuli were standardised, to a size of 1.60 m x 0.87m (555 x 300 pixels) and coloured using a dark grey (RGB 58,50,48) in order that they provided a uniform background for the target stimuli. At random time intervals (0.5 - 3 s), either a target (persons adopting a shooting stance) or non-target (persons with hands up above their head) would appear at a random window. For a target stimulus, a mouse click was required as quickly as possible (no locational movement required), whereas no response was required for a non-target. The two stimuli were not of the same spatial frequency due to this not being representative of real-world scenarios, however stimuli size was standardised. Participants were instructed to place equal importance on both response time and accuracy. For the SDST there was a 2:1 ratio between targets and non-targets, with two targets and one non-target appearing in each location during each SDST. The SDST was created and recorded using SuperLab 5 software (version 5.05; Cedrus®, San Pedro, USA), with response times recorded to the nearest millisecond for all target stimuli.

For both conditions, a gaming mouse (Logitech G203, Logitech, Lausanne, Switzerland) with 1 ms latency was employed. In the walking condition, the mouse was attached to the side of a replica SA80 rifle, of correct mass, with a mouse button adjacent to the trigger location. During the seated condition, the tests were displayed on a laptop screen (30.9 x 17.4 cm; Toshiba, Tokyo, Japan), whilst for the walking condition, the task was projected ~2.6 m in front of the individual walking on the treadmill (0.97 x 0.79 m). A marker was placed on the side of the treadmill, and participants were instructed to stay in line with it, so that the stimuli size was consistent for all participants. Again, whilst this approach may have been less than optimal, with respect to action fidelity, due to different response modes, as the study was not designed to compare walking and seated conditions, this approach was chosen principally due to the impracticalities of utilising physiological laboratories to collect seated data.

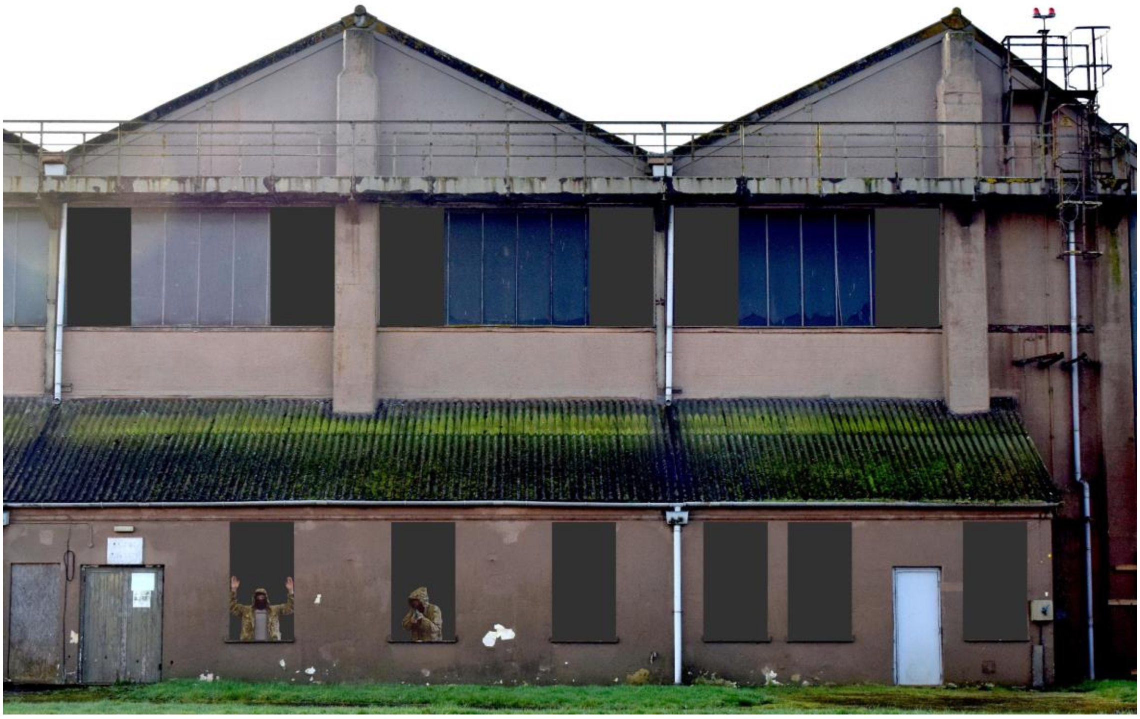

Figure 1. Example Shoot-/Don't-Shoot Stimuli.

Seated Reliability of Military Specific Cognitive Assessments

Part 1 comprised of three laboratory visits. During the first visit, participants were familiarised (two full trial completions of each assessment) with the MSANT and SDST, in a randomised counterbalanced order. The number of familiarisation trials was based on initial pilot testing, and time expediency. For the second, and third visits, participants completed the MSANT and SDST in the same randomised counterbalanced order as they had during visit one. Trials were separated by a minimum of 24 hours.

Walking Reliability of Military Specific Cognitive Assessments

For part 2, a subsample of 12 participants completed three additional laboratory visits and completed the SDST and MSANT whilst walking on a treadmill. The MSANT and SDST were again completed in a randomised counterbalanced order. All tests were completed with 10 minutes of seated rest between trials to negate the influence of physical fatigue. To enable a physiological steady-state to occur, participants completed five minutes of walking before the commencement of the cognitive assessments. For all walking trials, participants walked on a motorised treadmill (6.5 km·h^{-1}, 1% gradient) based on the fastest load carriage walking speed within the literature (Blacker et al., 2013) and represents a typical ‘enemy contact’ speed (Armstrong, Ward, Lomax, Tipton, & House, 2019).

Analysis of Cognitive Assessment Parameters

MSANT

The number of correct responses and partially correct responses was collected and compared for each trial. A correct response was when both letters were correctly identified and in the correct order. A partial response was when an individual letter in a pair was identified, in the correct location (i.e. first or second letter), but the other letter given was incorrect. To give an additional level of fidelity and sensitivity between individuals, the parameter of total combined correct responses was calculated (Equation 1). Within this equation, a weighting was added to total correct responses to differentiate from partial correct responses and also highlight the importance of correct responses compared with partial correct responses within the context of military operations.

Total combined correct responses =

(3 x Total correct responses) + Partial correct responses (1)

SDST

The number of shoot correct, don’t-shoot correct, total correct (Equation 2), and response times was compared across trials. A response time greater than 1 second, was classified as a non-response. To determine whether changes in the aforementioned SDST parameters were operationally relevant the accuracy-speed trade-off ASTO) variable was calculated (Equation 3).

$$\text{Total correct responses} = (\sum \text{shoot correct} + \sum \text{don't-shoot correct}) \qquad (2)$$

$$\text{ASTO} = (\text{Average response time}) \div (\text{Total correct responses}) \qquad (3)$$

Statistical Analysis

Data were principally analysed using JASP (JASP, 2020; Version 0.14.1) and are presented as mean ± standard deviation unless otherwise stated. For comparative purposes, scores were converted to percentages. Data were assessed for normality using skewness and kurtosis ratios (Fallowfield et al., 2005), and sphericity; with the Greenhouse-Geisser correction applied if sphericity assumptions were violated. For normally distributed data, a one-way ANOVA was employed to identify whether a likely main effect of assessment time point (including familiarisation trials) was apparent. Effect sizes are presented as Omega squared (ω^2) (Levine & Hullett, 2002), with 0.01, 0.06, and 0.14 classed as small, medium and large, respectively (Field, 2013). Where F-statistics, *p*-values, and effect sizes, likely indicate an incompatibility with the null model, Holm-Bonferroni adjusted pairwise comparisons were made *post-hoc*. Whilst this approach was utilised to investigate incompatibility with the null model across assessment periods, a lack of incompatibility does not imply equity between time points; thus two one-sided tests were employed between trials 1, 2, and 3, to assess whether differences in scores between trials were at least as extreme as the smallest effect size of interest (Lakens et al., 2018). For non-parametric data a Wilcoxon Two One Sided Tests was conducted in R Studio (version 2021.09.1), and the TOSTR package (version 0.4.1). Based upon the a priori sample size calculation, d = 0.5 was employed as the smallest effect size of interest. To describe the typical variation in assessment parameters between trials, Limits of Agreement (LoA) ± 95% Confidence Intervals (CI), Standard Error of the Mean (SEM), and Smallest Detectable Change (SDC) values were calculated. The SEM was calculated by dividing the standard deviation of the difference between trials by $\sqrt{2}$ (Hopkins, 2000; Ludbrook, 2010). Using SEM the SDC was also calculated; SEM x 1.96 x $\sqrt{2}$ (van Kampen et al., 2013). For non-parametric data a Friedman's test was employed with effect sizes presented using Kendall's W. Where the combination of χ^2-

statistics, *p*-values, and effect sizes, indicate a likely incompatibility with the null model, *post hoc* pairwise comparisons were made using Conover's test.

References

Armstrong, N., Smith, S., Risius, D., … D. D.-B. M., & 2022, U. (2022). Cognitive performance of military men and women during prolonged load carriage. *BMJ Military Health.*

Armstrong, N., Ward, A., Lomax, M., Tipton, M. J., & House, J. R. (2019). Wearing body armour and backpack loads increase the likelihood of expiratory flow limitation and respiratory muscle fatigue during marching. *Ergonomics*, *62*(9), 1181–1192.

Bhattacharyya, D., Pal, M., Chatterjee, T., & Majumdar, D. (2017). Effect of load carriage and natural terrain conditions on cognitive performance in desert environments. *Physiology & Behavior*, *179*, 253–261.

Blacker, S. ., Fallowfield, J. L., Bilzon, J. L. J., & Willems, M. E. T. (2013). Neuromuscular impairment following backpack load carriage. *Journal of Human Kinetics*, *37*(1), 91–98.

Crowell, H. P., Krausman, A. S., Harper, W. H., Faughn, J. A., & Sharp, M. A. (1999). *Cognitive and Physiological Performance of Soldiers While They Carry Loads Over Various Terrains*. Army Research Laboratory, Aberdeen Proving Ground, MD.

Eddy, M. D., Hasselquist, L., Giles, G., Hayes, J. F., Howe, J., Rourke, J., Coyne, M., O'Donovan, M., Batty, J., & Brunyé, T. T. (2015). The effects of load carriage and physical fatigue on cognitive performance. *PloS One*, *10*(7), e0130817.

Fallowfield, J. L., Hale, B. J., & Wilkinson, D. M. (2005). *Using statistics in sport and exercise science research*. Lotus Publishing.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. sage.

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, *30*(1), 1–15.

International Civil Aviation Organization. (2019). *Alphabet - Radiotelephony* (Vol. 2019). https://www.icao.int/Pages/AlphabetRadiotelephony.aspx

JASP. (2020). *JASP (Version 0.14.1)[Computer software]*.

Kazemi, R., Motamedzade, M., Golmohammadi, R., Mokarami, H., Hemmatjo, R., & Heidarimoghadam, R. (2018). Field study of effects of night shifts on cognitive performance, salivary melatonin, and sleep. *Safety and Health at Work*, *9*(2), 203–209.

Kobus, D. A., Brown, C. M., Wu, L., Robusto, K., & Bartlett, J. (2010). *Cognitive Performance and Physiological Changes under Heavy Load Carriage*. Pacific Science and Engineering Group Inc., San Diego, CA.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*(4), 612–625.

Ludbrook, J. (2010). Confidence in Altman-Bland plots: A critical review of the method of differences. *Clinical and Experimental Pharmacology and Physiology*, *37*(2), 143–149.

Nibbeling, N., Oudejans, R. R. D., Ubink, E. M., & Daanen, H. A. M. (2014). The effects of anxiety and exercise-induced fatigue on shooting accuracy and cognitive performance in infantry soldiers. *Ergonomics*, *57*(9), 1366–1379.

van Kampen, D. A., Willems, W. J., van Beers, L. W. A. H., Castelein, R. M., Scholtes, V. A. B., & Terwee, C. B. (2013). Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *Journal of Orthopaedic Surgery and Research*, *8*(1), 40.