2013

Measuring Distribution of Attention as a Part of Situational Awareness - a Different Approach

Katrin Meierfrankenfeld

Werner Greß

Tina Vorbach

MEASURING DISTRIBUTION OF ATTENTION AS A PART OF SITUATIONAL AWARENESS - A DIFFERENT APPROACH

Meierfrankenfeld, Katrin
German Air Force Institute of Aviation Medicine
Fürstenfeldbruck, Germany
Greß, Werner
German Air Force Institute of Aviation Medicine
Fürstenfeldbruck, Germany
Vorbach, Tina
German Air Force Institute of Aviation Medicine
Fürstenfeldbruck, Germany

This paper outlines a different approach to measure pilot aptitudes during flight simulator missions. An algorithm was developed to assess a candidate's distribution of attention beyond observation technique, eye-tracking or multi-dimensional tracking (e.g. altitude, speed, heading), thus getting rid of typical measurement problems. The algorithm used to evaluate candidate's distribution of attention in Phase III, German Armed Forces' third phase of aircrew selection consisting of simulator flights in a typical training scenario, is a mere time measure. The following article describes its construction as well as advantages and disadvantages.

Situation awareness (SA) is a key concept in aviation psychology. Crashes are frequently explained by loss of SA (e.g. Endsley & Garland, 2000; Nullmeyer, Stella, Montijo, & Harden, 2005; Jones, D.G., & Endsley, M.R., 1996). Aircraft interface upgrades are justified by assumed increases in SA (Vidulich, 2003), and a lot of research deals with SA. Nevertheless, it is yet unclear what SA in fact is. A widely accepted definition is proposed by Endsley (2000): [SA is] "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future." (p. 5). Distribution of attention might be seen as first level of SA. It can also be called attentional flexibility or allocation of attention, which is often measured by visual scanning behavior (Bellenkes, Wickens & Kramer, 1997). Salmon, Stanton, Walker, and Green (2006) state: [Measurement of eye movements is] "recording the process that operators use in order to develop SA" (p.234). For use in personnel selection, it is inconvenient to wear eye trackers because of weight, costs, and usual drop-outs due to technical problems. Furthermore, "to see" does not necessarily mean "to perceive". Another approach to measure distribution of attention is multi-dimensional tracking (e.g. altitude, speed, heading) based on deviation measures, leading to the problem of units: How to build a score of e.g. 10° heading deviation, 5 knots speed deviation and 100 feet altitude deviation? In flight training and selection programs, expert ratings are frequently used (e.g. FAA flight test standards, FAA, 2002). As typical observer mistakes can occur, objective data might be helpful to ensure objectiveness and reliability. This paper will describe the development of an approach to measure distribution of attention objectively.

Measuring Distribution of Attention – a different approach

This section gives a brief overview over German Armed Forces' aircrew selection process, leading to a description of the simulator flight investigated here and steps towards measuring distribution of attention.

German Armed Forces' Aircrew Selection

German Armed Forces' (GAF) aircrew selection procedure consists of three phases. Phase I and II include the assessment of basic aptitudes and the aviation-medical examination. Phase III (fixed wing) is more complex. It consists of one week simulator-based screening in a typical training scenario: Candidates prove their skills both in 4 simulator-flight missions with increasing workload and in academic training. As in real flight training, a briefing, a demonstration and a practice phase and subsequent debriefings prepare candidates for their check phases. The aim is to evaluate aptitudes, to propose specific cockpit assignments (e.g. jet pilot, weapon system operator/ navigator, transport pilot), and to minimize attrition rate during basic flight training. The aircrew selection process works quite well, as long term evaluation shows: Attrition rates during flying training are very low (e.g. in ENJJPT: 2007 to 2012: 5,4% total and 3,8% due to flying deficiencies). Per year, approximately 200 applicants are tested at Phase III fixed wing.

Flight Simulator used in this study: The FPS/F

The FPS/F (Aviation Psychological Pilot Selection System/ Fixed Wing) is a flight simulator consisting of 4 cockpits with lockable canopies, a spherical projection dome with 200° horizontal and 45° vertical field of view, a 5-channel high resolution projection system, a multi-functional display with all basic flight instruments plus a master caution panel for malfunctions and a radio panel (Figure 1). The instructor's consoles enable monitoring the applicant's activities and performance. Video protocols as well as mission logs are used for debriefing purposes. Data can be analyzed at an evaluation station.

First steps on the way to measure distribution of attention objectively

First results from Mission 2 (Figure 2) are reported. Mission 2 consists of traffic patterns with full stop landing. As the required X-check varies with the demands of the tasks, sets of variables necessary for a proper X-check for every maneuver in Mission 2 were defined. Acceptable deviations around ideal values were defined in a second step. Examples are shown in Table 1. Third, an algorithm was defined that computes the proportion of time where distribution of attention fails. This means the candidate violates one or more of the defined ranges in heading, altitude, speed, vertical speed and/or angle of bank **and** is **not** correcting. Afterwards, composite scores were calculated. These composite scores for distribution of attention reflect distribution of attention during the whole mission. As performance during Mission 2 changes due to concentration, practice effects or difficulty, composite scores for each pattern were calculated, too. In spite of grouping by time (pattern 1, 2 or 3), we also grouped by maneuver type. This means composite scores were calculated for turns, legs and other maneuvers.

Figure 1. The flight simulator used in Phase III/ fixed wing (FPS/F) consists of cockpits with a high quality screen comprising the field of view (200° horizontal, 45° vertical) (left), and the multifunctional display showing expanded instrumentation as well as touchscreen and radio (right).

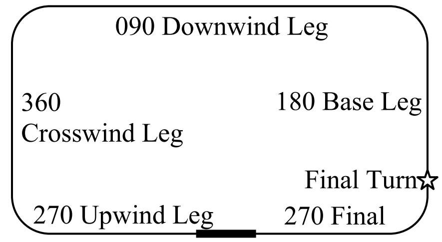

Figure 2. Mission 2 consists of 3 pattern flights starting on runway 27. In pattern 3, a gear emergency occurs on downwind leg and on base, new R/T is required.

Expert ratings in Phase III. Experts grade each maneuver, each pattern, the whole Mission 2, as well as distribution of attention on scales from 1, "excellent", to 7, "unsatisfactory". Further aptitudes assessed during Phase III are not reported in this article. Success in Phase III as reported here is based on performance and progress from Mission 1 to Mission 4. It ranges from 1, "excellent", to 7, "unsatisfactory", with grades 6 and 7 indicating no proposal for any cockpit position.

Hypothesis. An objective time-based score for distribution of attention should correlate with the experts' ratings of distribution of attention, with performance in Mission 2 and with overall success at Phase III. The last mentioned correlation is expected to be only small to medium, as Mission 2 is only one of four missions plus theoretical tests that lead to the decision if someone is proposed to become a GAF's crew member.

Results

Data from all candidates who passed Phase III from January to April 2012 were used. From 52 applicants, 1 passed with "A/excellent", 4 with „B/good", 11 with „C/average", 8 with „D/marginal", and 28 with "U/unsatisfactory". There was 1 female candidate. Applicants were young adults, aged from 18 to 24 (M = 20, SD = 1,96). 83% have passed A-levels, 7 had a secondary school level certificate (13%) and 2 (4%) had an advanced technical college entrance qualification.

Experts graded distribution of attention ranged from 2 to 7 (M = 4.8, SD = 1.42). Performance in Mission 2 ranged from 1 to 7, with a mean of 4.9 (SD = 1.57). Average grade in pattern 1 was 4,8 (SD = 1.31), 4,7 in pattern 2 (SD = 1.32) and 5,1 in pattern 3 (SD = 1.6). The score for distribution of attention in Mission 2 could theoretically range between 0 (all relevant parameters within the acceptable ranges **or** correcting towards the desired values) and 57 (always exceeding the acceptable ranges **and** not correcting). In this sample, the score ranged from 9.85 to 31.56 (M = 20.6, SD = 4.59), with lower scores indicating a better distribution of attention. The mean pattern-wise scores are 7.60 for pattern 1 (SD = 2.16); 7.30 for pattern 2 (SD = 1.77) and 7.44 for pattern 3 (SD = 2.06).

Correlations between reported scores are shown in Table 2. The computed score and the experts' ratings correlate .70, which is significant with $\alpha < .01$ and large (Cohen, 1988): The smaller the score that is the better the distribution of attention, the better the expert's rating of distribution of attention. Furthermore, the computed composite score and success in Phase III correlates .53 ($\alpha < .01$): As expected, a low composite score (meaning good distribution of attention) and good grades (indicating good performance in Phase III) are associated. Anyway, the expert's grades for distribution of attention correlate .72 with results in Phase III ($\alpha < .01$).

Correlations between expert's ratings of distribution of attention and pattern-wise and maneuver-wise composite scores were medium to high and all significant ($\alpha < .05$; see Table 2). Among those correlations, the correlation with the composite score of pattern 2 and with the composite score consisting of legs in pattern 3 are the most high (> .60, $\alpha < .01$).

Conclusion and Discussion

First results concerning the development of a time-based measurement of distribution of attention are reported. Correlations between the computed scores and expert ratings were – as expected – large (Cohen, 1988) and significant. Thus, it seems to be a promising approach. Yet, some questions remain. At first, the scores presented here are mere composite scores without any weighting. Weighting maneuvers depending on their difficulty would be reasonable: Turns are more dynamic than straight and level legs, thus a more fluent and quick distribution of attention might be needed. Legs in pattern 3 are more difficult than in pattern 1 and 2 because of the occurrence of an emergency and unexpected R/T, thus indicating distribution of attention although distractors are calling for attention. Correlations (Tab. 2) might point in that direction: Pattern 2 and 3 and legs in pattern 3 seem to be most important. Further investigation is needed to test these hypotheses. Second, the candidate's progress in distribution of attention might be interesting. As can be seen in the results section, applicants perform best at pattern 2 and the least at pattern 3. Does distribution of attention differ during patterns, too? And how is its progress during whole Phase III, from Mission 1 to Mission 4? The third issue to be discussed is about limits and chances of performance measures. The score is performance-based – time of (non-) performance is measured. In flight training, performance is main criteria, too. But performance measures are always contaminated: Psychomotor skill, decision making, speed of information processing, concentration, speed of automation, stress level, aggressiveness and other aptitudes influence performance in such a complex scenario as Mission 2, too. This problem cannot be solved in this article; for discussion see e.g. Pew (2000). Anyway, relationships between aptitudes assessed in Phase III and the composite score should be analyzed, giving hints concerning the score's construct validity. Furthermore, the relationship between expert ratings and composite scores should be examined in more detail: While objective measures are supposed to be more reliable than subjective measures because the latter might be contaminated by the human observer, objective measures reduce information, thus they might be failing to explain complex decision processes and/ or scenarios (e.g. Bell & Lyon, 2001). How is the computed score working, does it explain variance in applicant's performance beyond observer's ratings, is there a gain in incremental validity? Is the higher correlation between success in Phase III and experts' ratings of distribution of attention due to implicit weighting, to methodology effects – observer's scales are similar, and both rely on observation method- , is observation biased, or is its predictive validity as to later training results in fact higher? A promising approach for further improvements of the selection process.

Table 1.

Sets of Criteria (Examples) and Acceptable Deviations Around Them as Basis for the Distribution of Attention Score.

Maneuver	Set of variables	Range
Level Flight	Altitude 1000 feet	*+/- 20 feet*
	Indicated Air Speed 130 kts	*+/- 3 kts*
	Required HDG	*+/- 1°*
Level Turn	Altitude 1000 feet	*+/- 20 feet*
	Indicated Air Speed 130 kts	*+/- 3 kts*
	AOB 30°	*+/- 2°*
	R/O: Required HDG	*+/- 1°*

Note. $N = 52$. HDG = Heading. AOB = Angle of Bank. R/O = Roll out.

Table 2.

Correlations between Computed Scores of Distribution of Attention, Expert Ratings and Success at Phase III

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 Expert Rating	1														
2 Score Total	.70**	1													
3 Score Pattern 1	.50**	.78**	1												
4 Score Pattern 2	.63**	.85**	.51**	1											
5 Score Pattern 3	.58**	.81**	.37**	.62**	1										
6 Score Turns 1	.41**	.66**	.80**	.40**	.37**	1									
7 Score Turns 2	.51**	.68**	.46**	.62**	.58**	.44**	1								
8 Score Turns 3	.43**	.66**	.26	.54**	.8**	.32*	.54**	1							
9 Score Legs 1	.43**	.65**	.71**	.36**	.47**	.63**	.54**	.37**	1						
10 Score Legs 2	.53**	.69**	.42**	.78**	.51**	.36**	.40**	.44**	.28*	1					
11 Score Legs 3	.61**	.81**	.50**	.59**	.89**	.39**	.57**	.61**	.6**	.52**	1				
12 Score Dyn 1	.31*	.52**	.76**	.39**	.10	.34*	.16	.00	.18	.31*	.22	1			
13 Score Dyn 2	.35*	.53**	.27	.76**	.32*	.14	.07	.25	.04	.40**	.27	.36*	1		
14 Score Dyn 3	.35*	.74**	.14	.36**	.73**	.17	.27	.29*	.16	.27	.53**	.02	.25	1	
15 Success_PhaseIII	.72**	.53**	.30*	.49**	.51**	.23	.33*	.45**	.21	.39**	.40**	.23	.35*	.4**	1

Notes. $N = 52$. Pearson product-moment correlation coefficients are reported. The computed score for distribution of attention at Mission 2 is called Score Total. Correlations between scores for legs, turns, and other maneuvers (dyn = dynamic) as well as pattern-wise scores are also reported. Expert Rating refers to expert rating of distribution of attention. * $p \leq .05$ ** $p \leq .001$.

References

Bell, H.H., & Lyon, D.R. (2001). Using observer ratings to assess situation awareness. In Mica R. Endsley, Danlie J. Garland (Eds.). Situation Awareness Analysis and Measurement (pp. 129-146). Mahwah, NJ: Lawrence Erlbaum Associates.

Bellenkes, A.W., Wickens, C.D., & Kramer, A.F. (1997). Visual scanning and pilot experience: The role of attentional flexibility and mental model development. *Aviation, Space, and Environmental Medicine*, *68*, 569-579.

Cohen, J.W. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn). Hillsdale, NJ: Lawrence Erlbaum Associates.

Endsley, M.R. (2000). Theoretical Underpinnings of Situation Awareness: A critical Review. In Mica R. Endsley, Danlie J. Garland (Eds.). *Situation Awareness Analysis and Measurement* (pp. 3-32). Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M.R., & Garland, D.G. (2000). Pilot situation awareness training in general aviation. In *Proceedings of the 14th triennial congress of the International Ergonomics Association and the 44th annual meeting of the Human Factors and Ergonomics Society* (pp. 2/357-2/360). Santa Monica, CA: Human Factors and Ergonomics Society.

Federal Aviation Administration. (2002). *Private pilot practical test standards for airplane* (FAA Pub. No. FAA-S-8081-14A). Retrieved from http://www.faa.gov/training_testing/testing/airmen/test_standards/

Jones, D.G., & Endsley, M.R. (1996), Sources of Situation Awareness Errors in *Aviation. Aviation, space and environmental Medicine,* 67 (6), 507-512.

Nullmeyer, R.T., Stella, D., Montijo, G.A., & Harden, S. W. (2005). Human Factors in Air Force flight mishaps: Implications for change. In *Proceedings of the 27th annual Interservice/Industry Training, Simulation, and Education Conference* (Paper No. 2260). Airlington, VA: National Training Systems Association.

Pew, R.W. (2001). The state of situation awareness measurement: Heading toward the next century. In Mica R. Endsley, Danlie J. Garland (Eds.). Situation Awareness Analysis and Measurement (pp.33-47). Mahwah, NJ: Lawrence Erlbaum Associates.

Salmon, P.M., Stanton, N.A., Walker, .H., & Green, D. (2006). Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics*, *37*(2), 225-238.

Vidulich, M.A. (2003). Mental Workload and Situation Awareness. In Pamela S. Tsang, Michael Vidulich (Eds.). Principles and Practice of Aviation Psychology (pp. 115-146). Mahwah, NJ: Lawrence Erlbaum Associates.