Normative Performance Measurement in Simulated Air Combat

Heikki Mansikka; Kai Virtanen; Lauri Mäkinen; Don Harris

BACKGROUND: Normative performance (NP) describes the pilots' adherence to tactics, techniques, and procedures (TTPs). Until now, there has not been a global NP measurement technique for beyond visual range (BVR) air combat, and the methodology and technology related to the evaluation of NP have fallen behind the pace of the overall technical progress of distributed mission operations (DMO) training.

METHODS: Platform-independent core air combat tasks were identified. The execution of these tasks is directed with TTPs. BVR air combat missions were flown in a DMO simulator system and the design NP was varied between missions. Observers viewed debriefs of these missions and attempted to identify TTP-regulated air combat tasks. Once identified, they scored the pilots' NP in those tasks. The scoring was based on the level of TTP adherence and the impact a nonadherence had on the mission accomplishment.

RESULTS: All observers were able to identify most of the TTP-regulated air combat tasks. There was a strong positive correlation between the observed and design NP scores. The overall Kappa indicated a fair agreement between the observers. The percentage of observers' NP assessments which agreed with the design NP varied from 49.60 to 85.28% in different air combat missions. On average, 73.96% of the observers' NP scores agreed with the design NP scores.

CONCLUSIONS: Observers were able to accurately identify TTP-regulated tasks and score NP of these tasks during an air combat debrief. There was a moderate agreement between the observers' NP scores.

KEYWORDS: air combat, normative performance, simulation, tactics, techniques, and procedures.

Mansikka H, Virtanen K, Mäkinen L, Harris D. *Normative performance measurement in simulated air combat*. Aerosp Med Hum Perform. 2021; 92(11):908–912.

In air combat, both friendly and enemy aircraft have six fundamental offensive functions: 1) opposing aircraft must be found and identified; 2) their exact locations must be obtained; 3) they must be monitored until a decision is made whether to engage; 4) a weapon and/or a sensor must be assigned against the opposing aircraft once the decision to engage has been made; 5) the weapon and/or a sensor must be employed against the opposing aircraft; and 6) the postengagement state must be monitored to determine the required follow-on actions.[4] In addition, both friendly and enemy aircraft have six fundamental defensive functions aimed at denying the accomplishment of the opponent's offensive functions. Finally, there are general functions, such as fuel management, which must be completed in every mission.

While the air combat functions describe what must be achieved, the core air combat tasks describe what must be done to accomplish those achievements. For example, a radar search is a core task associated with the offensive function of finding and identifying the enemy. To enable fighter pilots to cope with the task demands of air combat, they usually operate in flights of four aircraft. For the flight to sequence and integrate the tasks of its members, their actions must be coordinated. Coordination can be either explicit or implicit.[3] Explicit coordination is achieved through communications, whereas implicit coordination is accomplished by adhering to tactics, techniques, and procedures (TTPs). TTPs are tactical contracts of how air

From Aalto University, Aalto, Finland; and the National Defence University, Helsinki, Finland.

This manuscript was received for review in April 2021. It was accepted for publication in September 2021.

Address correspondence to: Heikki Mansikka, Adjunct Professor, National Defence University, Kadettikouluntie, Helsinki FI-00860, Finland; heikki.mansikka@aalto.fi.

DOI: https://doi.org/10.3357/AMHP.5914.2021

combat tasks should be executed.[6] A single air combat task can have one or many TTP task variables associated to it. When the TTP task variables are given quantitative values, they can be written as a quantitative contract or rule, e.g., "During radar search, a wingman must search the enemy aircraft at altitudes of 10,000 ft and below, and at least 45° left and right from the aircraft centreline." A TTP rule can also be qualitative. For example, "Flight members must communicate their tactical status" is a qualitative rule.

During air combat, the flight continuously observes the air combat environment, recognizes potential TTPs, and decides which of them is expected to produce a satisfactory outcome. As the selected TTP is executed, the flight dynamically evaluates whether to continue the execution of the TTP or to select a more appropriate TTP. If the TTP execution is completed, the flight selects a new TTP. From the flight's perspective, air combat is about constantly making TTP selection decisions and about executing one TTP or another. With TTP, if adhered to, it is possible for the pilots to anticipate each other's behaviors, enabling an orchestration of the flight members' actions without having to communicate or plan the activity.[5]

TTPs are trained in live and virtual simulations. Distributed mission operations (DMO) training is used to satisfy the need to train TTP execution in individuals and teams in increasingly realistic and complex scenarios. DMO combines live, virtual, and constructive simulators to facilitate training of many remotely located participants. However, the methodology and technology related to the evaluation of air combat performance have fallen behind the pace of the overall technical progress of DMO training. Any performance assessment should capture three aspects: 1) what was the performance output relative to a set of criteria; 2) were the systems used as they were supposed to be used; and 3) was the TTP executed as it was supposed to be executed? It is still a common practice to use subject matter experts' (SMEs') subjective assessments for the evaluation of air combat performance despite the benefits of objective, automated assessments.[7]

Normative performance (NP) describes the level of pilots' TTP adherence, i.e., how accurately the TTP is followed during its execution.[4] In addition, it also considers the impact nonadherence has on mission accomplishment. Measurement of NP is critical when the utility of TTPs, the competence of flights, or the applicability of aircraft systems is evaluated and compared. Similarly, an evaluation of an aircraft system is worthless if the system is not used during the TTP execution as it is supposed to be used. If NP measurement is ignored or overlooked, it is possible that ineffective TTPs end up in operational use, potentially effective aircraft systems are abandoned during operational testing and evaluation, and nonadherent behavior is not effectively identified during air combat training.

The objective of this paper is to introduce a global NP measurement technique for beyond visual range (BVR) air combat. The development of the NP measurement technique is discussed and demonstrated in a simulated air combat. However, while this paper focuses on air combat, the principles of developing the NP measurement technique and using it can be applied to almost any military or civilian tasks where the execution of those tasks is regulated and postactivity observations are possible.

METHOD

In Finland, ethical review of nonmedical research involving human participants is based on a set of guidelines drawn up by the Finnish National Board on Research Integrity, TENK. According to the guidelines of TENK, the research configuration of this paper was such that it did not require an ethical review statement from a human sciences ethics committee.

Materials

The objective of the proposed NP measurement technique is to determine how well pilots comply with TTPs and the underlying TTP task rules. To enable NP measurement, it was necessary to identify the tasks a typical air combat mission is comprised of. For this purpose, air combat related research articles, technical reports, and military manuals were reviewed: 131 candidate air combat tasks were identified. Of these candidates, a group of SMEs selected 25 tasks for further refinement, during which closely related tasks were combined into more meaningful units, duplicates were removed, and the remaining tasks were organized under offensive and general air combat functions. As a result, the number of core air combat tasks was reduced to 17. Next, 20 qualified fighter pilots evaluated the importance of whether these core tasks are correctly executed for the flight's mission accomplishment. A decision-tree based on a Cooper-Harper format[8] was developed to aid the pilots with the rating exercise. The rating scale ranged from 1 (low importance) to 5 (high importance). Each rating was associated with a verbal description (**Fig. 1**). Tasks which had little or no impact on mission accomplishment were discarded. As a result, 14 core air combat tasks were shortlisted. Each shortlisted task is typically regulated by TTP.

The NP measurement technique is based on postflight observer ratings about the pilots' TTP adherence. To support this, TTP adherence questions tapping each shortlisted core air combat task were prepared (**Table I**). The questions were formulated in a generic form such that they did not refer to any specific TTP rule, thus avoiding disclosure of classified TTP information.

The TTP adherence questions enabled NP measurement of any BVR air combat TTPs, regardless of the rules used. For example, the core air combat task 'electronic protection' was written as a probe 'Did the flight member conduct electronic protection (e.g., chaff, flare, and self-protection jamming) as directed by TTP?'. Then, a platform specific TTP associated with that task could dictate, e.g., a certain jamming program to be used at a certain range from the threat.

Statistical Analysis

When NP is assessed, an observer monitors the pilot's cockpit recordings and mission reconstruction. When the observer identifies a TTP-regulated air combat task, the mission

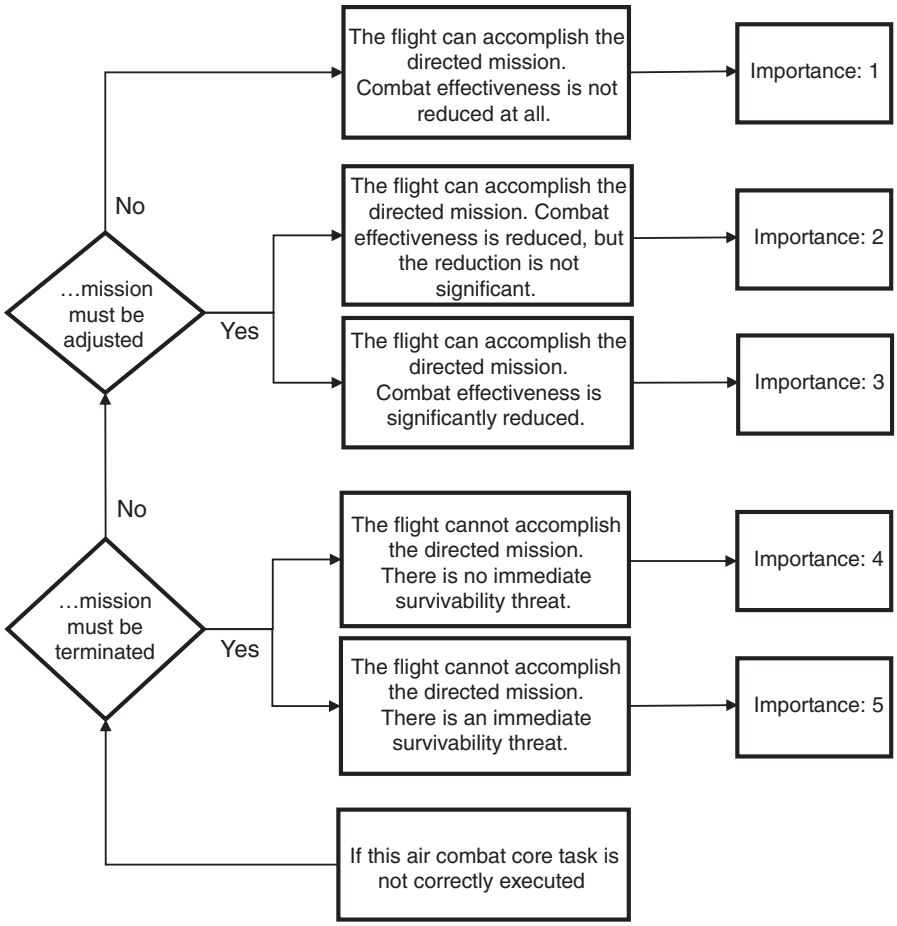

Fig. 1. Rating tool for the core air combat tasks.

playback is paused, the TTP adherence question is introduced and NP for an associated TTP rule is rated. The NP score is based on the level of TTP adherence and the impact a nonadherence has on the mission accomplishment. The NP scores range from 0 (low NP) to 3 (high NP). Each score is associated to a verbal description as follows: 0 = Did not adhere to TTP, negative impact on mission accomplishment, impact was significant; 1 = Did not adhere to TTP, negative impact on mission accomplishment, impact was not significant; 2 = Did not adhere to TTP, no impact on mission accomplishment; 3 = Did adhere to TTP. Once the NP score is assigned, the mission reconstruction and cockpit recordings are again played until the next TTP-regulated air combat task is identified. The procedure is repeated until all relevant TTP-regulated air combat tasks are assigned NP scores. In the next section, the use of the NP measurement technique is demonstrated in a simulated air combat.

Table I. TTP Adherence Questions.

AIR COMBAT FUNCTIONS	TTP ADHERENCE QUESTIONS
Find, Fix, Track	Did the flight member comply with his search responsibilities as directed by TTP?
	Did the flight member comply with his targeting and sorting responsibilities as directed by TTP?
	Did the flight member comply with his identification and rules of engagement responsibilities as directed by TTP?
Engage, Assess	Did the flight member employ weapons as directed by TTP (e.g., shot doctrine, weapon engagement zone management)?
	Did the flight member conduct weapon support as directed by TTP (e.g., datalink support, datalink support termination timing)?
	Did the flight member assess weapon probability of kill as directed by TTP?
General	Did the flight member maintain mutual support unless otherwise approved as directed by TTP?
	Did the flight member follow the intercept geometry and timeline as directed by TTP?
	Did the flight member manage the enemy weapon engagement zones as directed by TTP?
	Did the flight member manage his fuel as directed by TTP?
	Did the flight member conduct electronic attack as directed by TTP?
	Did the flight member conduct electronic protection as directed by TTP?
	Did the flight member use tactical radios and brevity as directed by TTP?
	Did the flight member use the datalink as directed by TTP?

TTP: tactics, techniques, and procedures.

Table II. Descriptive Statistics of NP Scores

	DESIGN NP SCORES		OBSERVED NP SCORES		SCORE AGREEMENT	
NP MISSION	M	SD	M	SD	*N*	%
High	2.93	0.25	2.82	0.42	21	84
Medium	2.45	0.75	2.39	0.84	18	72
Low	1.86	1.16	1.82	1.18	16	64

Means (M) and standard deviations (SD) of the design and observed normative performance (NP) scores in high, medium and low NP missions. In addition, the number and percentage of the observers' NP scores which agreed with the design NP scores are summarized.

Table III. Design and Observed NP Scores.

DESIGN		OBSERVERS' SCORES				PERCENTAGE OF AGREEMENT	
NP SCORE	*N*	M	SD	κ	*P*	M	SD
0	4	0.50	0.44	0.454	<0.001	60.00	26.73
1	15	1.05	0.50	0.323	<0.001	52.27	18.39
2	15	2.09	0.26	0.189	<0.001	49.60	13.02
3	66	2.83	0.21	0.552	<0.001	85.28	16.27
All	100	1.62	0.35	0.398	<0.001	73.96	22.88

Kappa values, means (M), and standard deviations (SD) of the observers' NP scores, and the percentage of their NP scores which agreed with the design NP scores 0, 1, 2, and 3 ($N = 100$). The second column denotes the number of the core air combat tasks associated with each design NP score.

Subjects

Twenty-five active-duty F/A-18 fighter pilots volunteered to support the demonstration as observers. The mean age of participants was 32.12 yr (SD = 3.55) and their average flying experience with F/A-18 was 628.80 flight hours (SD = 317.10). All pilots were familiar with the TTP-regulated air combat tasks they were directed to observe. Before the observation session, written informed consent was obtained from each participant. An additional four pilots were recruited to fly the BVR air combat simulator missions used in the demonstration.

Procedure

Three BVR air combat missions were programmed into high fidelity virtual simulators. All simulators were linked via a distributed interactive simulation (DIS) networking protocol. The pilots flew the missions as a flight with lead and wing elements. Both elements had two pilots, a leader and a wingman. NP of the lead element's wingman was varied between missions. In a high NP mission, the level of the wingman's TTP adherence was high, in a medium NP mission the TTP adherence was moderate, and in the low NP mission it was low. For example, in the low NP mission the wingman was briefed to engage a wrong target. The enemy aircraft's behaviors were programmed for each mission. Based on those behaviors, the flight's TTP-regulated air combat tasks were scripted and factual NP or design NP was predetermined for each of them. The missions were practiced until the wingman's NP scores matched the design NP scores of the TTP-regulated air combat tasks. Reconstructions of the missions where the two NPs matched were saved for observers' NP scoring.

The saved mission reconstructions were shown to observers in random order. The observers could view, replay, zoom, and rewind the mission reconstruction and cockpit recordings at will. The observers' task was to identify the TTP-regulated air combat tasks and to determine the NP scores of the lead element's wingman using the TTP adherence questions shown in Table I. The accuracy with which the observers were able to identify the TTP-regulated air combat tasks and the agreement between the design and observers' NP scores were analyzed. In addition, the level of agreement between the observers' NP scores was examined.

RESULTS

The missions included a total of 153 TTP-regulated air combat tasks. In the high NP mission, 31 of the 44 tasks were identified by all observers. In the medium NP mission, there were 58 TTP-regulated air combat tasks, of which 40 were identified by all observers. In the low NP mission, 29 of the 51 tasks were identified by all observers. Overall, the observers were able to identify 65.36% ($N = 100$) of the TTP-regulated tasks. Only the 100 tasks identified by all 25 observers were used for further analysis. **Table II** summarizes the means and SDs of design NP scores and observed NP scores in the high, medium, and low NP missions. Table II also presents the number and percentage of observers' NP scores which agreed with the design NP scores.

A Pearson product-moment correlation was run to determine the relationship between the design scores and the observed scores. There was a strong, positive correlation between the two, which was statistically significant ($r = 0.933$, $N = 100$, $P < 0.001$).

Fleiss' Kappa was run to determine if there was an agreement between the observers' NP scores. **Table III** summarizes the observers' NP scores and Kappa values, and the percentage of their NP scores which agreed with the design NP scores. As shown in Table III, the percentage of the observers' scores which agree with the design ones varied from 49.60% (design NP score of 2) to 85.28% (design NP score of 3). On average, 73.96% of the observers' NP scores agreed with the design ones.

DISCUSSION

NP is a critical, yet an under-used, measure. It describes the pilots' TTP adherence, i.e., how accurately TTPs are followed during their execution. In NP measurement, both the level of TTP adherence and the impact nonadherence had on the mission accomplishment are considered when assigning scores. Unless NP is measured, it is possible that dangerously misleading conclusions are drawn from the human-machine performance evaluations. In air combat, the pilots may arrive at (seemingly) poor decisions, which do not coincide with those dictated by TTP. In addition, the pilots can arrive at rational decisions but fail in their response execution. As a result, to capture the quality of pilots' response execution and the effectiveness of the directed TTP, NP must be measured.[4]

In this paper, a generic NP measurement technique for BVR air combat was developed and demonstrated. Based on a literature review and SME evaluations, the core air combat tasks were identified and TTP adherence questions tapping each

shortlisted task were prepared (see Table I). The generalized and unclassified TTP adherence questions were developed to enable reliable postmission observation-based NP measurement of any BVR air combat mission.

When the technique is used, substantial subject matter expertise is required from the observer for two reasons. As the TTP adherence questions do not refer to any specific TTP rule, the observer must be familiar with the TTP rules used in the observed mission. The observer must also be capable of evaluating how much the observed level of TTP adherence affects the flight's mission accomplishment. As shown in Tables II and III, qualified fighter pilots can identify TTP-regulated air combat tasks and score NP from the mission reconstruction.

According to Landis and Koch,[2] the computed overall Kappa indicated fair agreement between observers. For design NP scores 0 and 3, the Kappas indicated moderate agreement, and for design NP scores 1 and 2, the agreements were fair and slight, respectively (see Table III). However, as discussed by Gwet,[1] Kappa values can be misleading: a low frequency in one category and a high frequency in another can distort Kappa value (see Tables II and III). In a conventional interrater reliability analysis where the design NP scores would not be known, the agreement of the observers' NP scores would be of concern. In this paper's demonstration, however, the design NP scores were known a priori. Therefore, in addition to interrater reliability, it was relevant to analyze the percentage of observers whose NP scores agreed with the design NP scores.

In DMO air combat training, with potentially hundreds of participants, there is a need for automated and objective performance assessments.[7] Assessment of air combat performance is straightforward to automate as it is ultimately about reporting the number of kills and losses. This information is typically easily extracted from the DIS traffic. Automated assessment of NP is more challenging, as in addition to information about what happened, information what was supposed to happen in terms of TTP execution is needed. It is not likely that directed TTPs, TTP-regulated air combat tasks, and pilots' NP can be automatically determined from the DIS traffic any time soon. However, the TTP adherence questions (see Table I) can be helpful in the development of such automated NP measurement algorithms. Meanwhile, observations are likely to remain as the most relevant technique to measure NP in air combat training. The developmental phases of the measuring technique described in this paper can be used to develop NP measures for all air combat missions. Moreover, the principles of the proposed NP measurement technique can be applied to any regulated civil or military activity where postactivity observations about the adherence of regulated tasks are possible.

ACKNOWLEDGMENTS

Financial Disclosure Statement: The authors have no competing interests to declare.

Authors and Affiliations: Heikki Mansikka, Ph.D., M.A., and Kai Virtanen, D.Sc., M.Sc., Department of Military Technology, National Defence University, Helsinki, Finland; Heikki Mansikka, Insta DefSec, Tampere, Finland; Kai Virtanen, Department of Mathematics and Systems Analysis, Aalto University, Aalto, Finland; Lauri Mäkinen, M.Sc.(Mil.), Senior Staff Officer, Finnish Air Force Academy, Tikkakoski, Finland; and Don Harris, Ph.D., B.Sc., Faculty of Engineering, Environment and Computing, Coventry University, Coventry, United Kingdom.

REFERENCES

1. Gwet K. Handbook of inter-rater reliability. Gaithersburg (MD): Advanced Analytics; 2014.
2. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174.
3. Lim B, Klein K. Team mental models and team performance: a field study of the effects of team mental model similarity and accuracy. J Organ Behav. 2006; 27(4):403–418.
4. Mansikka H, Virtanen K, Harris D, Jalava M. Measurement of team performance in air combat–have we been underperforming? Theor Issues Ergon Sci. 2021; 22(3):338–359.
5. Mansikka H, Virtanen K, Harris D, Salomäki J.Live-virtual-constructive simulation for testing and evaluation of air combat tactics, techniques, and procedures, Part I: assessment framework. Journal of Defense Modeling and Simulation. November 2019.
6. Rajabally E, Valiusaityte I, Kalawsky R. Aircrew performance measurement during simulated military aircrew training: a review. In: AIAA Modeling and Simulation Technologies Conference proceedings. Reston (VA): AIAA; 2009:5829–5838.
7. Schreiber B, Bennet W Jr, Colerove C, Portrey A, Greschke D, Bell H. Evaluating pilot performance. In: Ericsson K, editor. Development of expert performance and design of optimal learning environment. Cambridge: Cambridge University Press; 2009:247–270.
8. Wierwille W, Casali J. A validated rating scale for global mental workload measurement applications. Proc Hum Factors Ergon Soc Annu Meet. 1983; 27(2):129–133.