Live–virtual–constructive simulation for testing and evaluation of air combat tactics, techniques, and procedures, Part 1: assessment framework

Mansikka, H. P., Virtanen, K., Harris, D. & Salomaki, J.

Original citation & hyperlink:
Mansikka, Heikki Petteri et al. "Live–virtual–constructive simulation for testing and evaluation of air combat tactics, techniques, and procedures, Part 1: assessment framework". Journal of Defense Modeling and Simulation: Applications, Methodology, Technology. 2019, (In-Press). (In-Press).
https://dx.doi.org/10.1177/1548512919886375

DOI 10.1177/1548512919886375
ISSN 1548-5129
ESSN 1557-380X

Publisher: SAGE Publications

LIVE-VIRTUAL-CONSTRUCTIVE SIMULATION FOR TESTING AND EVALUATION OF AIR COMBAT TACTICS, TECHNIQUES AND PROCEDURES, PART 1: ASSESSMENT FRAMEWORK

Heikki Mansikka

Department of Mathematics and Systems Analysis, Aalto University, Helsinki, Finland

Insta DefSec, Tampere, Finland

Kai Virtanen

Department of Mathematics and Systems Analysis, Aalto University, Helsinki, Finland Department of Military Technology, Finnish National Defence University, Helsinki, Finland

Don Harris

Faculty of Engineering and Computing, Coventry University, Coventry, United Kingdom

Jaakko Salomäki

Department of Military Technology, Finnish National Defence University, Helsinki, Finland

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors have no financial interest or benefit arising from applications of this research. Correspondence concerning this manuscript should be addressed to Heikki Mansikka, Department of Mathematics and Systems Analysis, Aalto University, P.O.Box 11100, FIN 00076 Aalto, Finland. E-mail: heikki.mansikka@aalto.fi

ABSTRACT

This paper advances live (L), virtual (V) and constructive (C) simulation methodologies by introducing a new L-V-C simulation framework for the development of air combat tactics, techniques and procedures (TTPs). In the framework, TTP is developed iteratively in separate C-, V- and L-simulation stages. This allows the utilization of the strengths of each simulation class while avoiding the challenges of pure LVC-simulations. C-stage provides the optimal TTP with respect to probabilities of survival (P_s) and kill (P_k) of aircraft without considering the human-machine interaction (HMI). In V-stage, the optimal TTP is modified by assessing its applicability with P_k and P_s, as well as HMI measures regarding pilots' situation awareness, mental workload and TTP adherence. In L-stage, real aircraft are used to evaluate whether the developed TTP leads to acceptable P_k, P_s, and HMI measures in a real-life environment. The iterative nature of the framework enables that V- or L-stages can reveal flaws of the TTP and an inadequate TTP can be returned to C- or V-stage for revision. This paper is Part 1 of a two-

part study. Part 2 demonstrates the use of the framework with operationally used C- and V-simulators as well as real F/A-18C aircraft and pilots.

Keywords: air combat, human factors, human-machine interaction, live-virtual-constructive, mental workload, performance, simulation, situation awareness, testing and evaluation

1. INTRODUCTION

A 'flight' in standard air force usage refers to a unit of four aircraft, which is composed of two 'elements', a lead element and a wing element. The elements have two aircraft in each, the leader and the wingman. A flight's primary goal in air combat is to keep itself in an offensive position that increases the probability of weapon intercept with the enemy, while simultaneously denying or lowering the enemy's probability of achieving the same.[1] For a flight to successfully achieve its primary goal, its members need to coordinate their actions. Pilots achieve this coordination by following tactics, techniques and procedures (TTPs), which, when followed, create discipline and provide some structure to a seemingly unpredictable and chaotic activity.[2] TTPs are comprised of a set of qualitative and quantitative rules. A quantitative rule has a variable and its value, or rule value. For example, 'Airspeed at missile launch must be Mach 1.0' is a quantitative rule, where Mach 1.0 is the rule value. A qualitative rule is a verbal description of activity. For example, 'Flight members must communicate their tactical status' is a qualitative rule. Application of inappropriate TTP makes it challenging for the flight to achieve its primary goal. Therefore, when TTPs are developed, a lot of effort is put on testing and evaluation (T&E) to increase the flight's likelihood of success. In this paper, live (L), virtual (V) and constructive (C) simulation methodologies are used for TTP T&E in a novel way by taking into account not just the aircraft's and weapon system's characteristics, but also the pilot's ability to interact with the aircraft and the whole air combat environment. By iterating and sequencing the separate C-, V- and L-simulations during TTP T&E, it is possible to utilize the strengths of each simulation class [3] in a cost-effective and safe manner

A pilot and a fighter aircraft form a complex human-machine process (HMP). The inputs to this process are twofold. Aircraft capabilities and limitations define the limits of the machine process (MP), while human capabilities and limitations define the human's ability to interact with the MP. The human-machine interaction (HMI) has three components; task adherence, situation awareness (SA) and mental workload (MWL).[4-5]

In air combat, task adherence describes how accurately a pilot follows the directed TTP. Therefore, task adherence is referred to as normative performance (NP). In contrast, MWL describes the imbalance between the demands of the HMI and the pilot's cognitive resources available to satisfy that demand.[6] Finally, according to Endsley[7], SA is a three-level, hierarchical state of pilot's comprehension of the elements in the environment within a volume of time and space (level 1 SA), the pilot's comprehension of their meaning (level 2 SA), and his/her projection of their status in the near future (level 3 SA). HMP output is affected by these components and dictates how well a flight achieves its primary goal, which is measured with the probability of kill (P_k) of enemy aircraft and the probability of survival (P_s) of friendly aircraft. To ensure TTP's efficacy, it is necessary to measure the components of HMI during TTP T&E. Adequate MWL, NP and SA during TTP T&E ensure the pilots are capable of maintaining the required level of HMP output even when the task complexity and stressors exceed those seen during TTP T&E. MWL, NP and SA are largely dissociated but inter-related components of HMI, and therefore it is not possible to determine any of their values by measuring the values of one of the others.[8] If one chooses to measure MP output with P_k and P_s, human limitations must be excluded from HMP by assuming full SA, perfectly balanced MWL and faultless NP. As outlined in Figure 1, the difference between MP and HMP outputs describes HMP loss caused by MWL, NP and SA of the pilots.

[insert Figure 1]

TTP T&E can be conducted in L-, V- and C-simulations. In a L-simulation real people operate real systems, in a V-simulation real people operate simulated systems or simulated people operate real systems, and in a C-simulation simulated people operate simulated systems.[9-10] Although there exist descriptive decision models imitating pilot's decision making (see, e.g., [11, 12]), NP, SA, MWL and qualitative TTP rules are difficult to take into account in C-simulations. In contrast, the ability to control precisely simulation entities, e.g., aircraft, systems, and behaviors, and to conduct batch-run simulations make C-simulations a cost-effective way to analyze P_k and P_s without the potentially confounding influence of HMI. Therefore, C-simulations are best suited for the early stages of TTP T&E, where the optimal values of TTP quantitative rules (hereinafter referred to as MP optimal values and MP optimal rules) are determined with respect to given optimization criteria and constraints related to P_k and P_s measuring MP output.

V-simulations provide a practical environment for TTP T&E, especially when safety, security, fiscal or resource limitations prevent T&E being conducted in L-simulations.[3] Also, as V-simulations enable the use of both qualitative and quantitative TTP rules by pilots, the values of MWL, NP, SA and HMP output can be measured for a given TTP in a simulated environment. While V-simulations add value to TTP T&E, they are time consuming compared to C-simulations when multiple simulation runs are required.

L-simulations are an expensive and resource heavy T&E environment. They are still needed in TTP T&E as they provide MWL, NP, SA and HMP output for the quantitative and qualitative TTP rules followed by pilots with real-life task complexity and stressors. However, if L-simulations are introduced to TTP T&E prematurely, they can be both inefficient and a risk to flight safety. While L-simulations are needed for TTP evaluation, they should be used for TTP testing only after the potential flight safety issues have been identified and mitigated in C- and V-simulations.

The division between the simulation classes is not clear-cut, as a simulation can include entities from more than one class.[13] This is particularly the case in LVC-simulations, where L-, V- and C-simulations are integrated into a large-scale, and often distributed, simulation activity. While such simulations have benefits for TTP T&E, e.g., reduced costs, access to simulations of limited availability assets and the ability to conduct multiservice test events [14], they have interoperability and architectural issues.[15] In short, LVC-simulations are most useful for testing and experimenting with large-scale systems and system of systems, technology demonstrations, and mission rehearsals.[10] In TTP T&E, however, it is more efficient to use the different simulation classes separately and in incremental stages, as this allows for a better utilization of the strengths of each simulation type.

This paper is Part 1 of a two-part study. In this Part 1, a Live-Constructive-Virtual (L-V-C) simulation assessment framework for TTP T&E is introduced. In Part 2[16], the use of this L-V-C framework is demonstrated with operationally used C- and V-simulators as well as real F/A-18C aircraft and qualified fighter pilots. Unlike LVC-simulations, the L-V-C framework does not attempt to mix the different simulation classes, and thus it avoids the challenges of LVC-simulations.[17] The proposed framework consists of separate C-, V- and L-stages in which TTP is iteratively developed in a given air combat scenario. In C-stage, qualitative TTP rules are not considered, but C-simulations are used to determine MP optimal values for TTP quantitative

rules. In V-stage, these optimal quantitative rules are employed, and verbal descriptions of qualitative rules are refined until HMP output measured using P_k and P_s is adequate and the scores for NP, SA, and MWL are acceptable. In this way, HMP optimal rules, consisting of MP optimal values of the quantitative rules and HMP optimal descriptions of the qualitative rules, in the simulated environment are obtained. Finally, TTP with these HMP optimal rules is evaluated in L-stage using L-simulations. If the real-life use of the HMP optimal rules results in an adequate HMP output, and the scores of NP, SA, and MWL are acceptable, the HMP optimal rules can be cleared for operational use. In other words, the resulting operationally HMP optimal rules ensure that the flight's primary goal is likely to be achieved in the light of P_k and P_s, while NP, SA, and MWL in the real-life environment remain acceptable. In this way, a desired safety margin with respect to human capabilities and limitations is maintained, even if the end-use task demands and complexity exceed those seen during TTP T&E.

A major advantage of the L-V-C simulation framework is its iterative nature. That is, V-stage can be repeated after either L- or V-stage if a need to modify qualitative rules arises. Alternatively, if quantitative rules need modifying, TTP can be returned from V- or L-stages back to C-stage. In L- and V-stages, HMP output and particularly SA, NP, and MWL scores provide a powerful tool for detecting TTP's possible flaws and for identifying how TTP could be improved. Also, if needed, these scores in V- and L-stages support the generation of modified optimization criteria and constraints applied at C-stage, should TTP T&E require repeated C-simulations. The resulting TTP is well balanced between MWL, NP, SA - and eventually HMP output, which reflects the primary goal of the flight.

2. L-V-C SIMULATION FRAMEWORK

2.1 Initialization of TTP T&E

The L-V-C simulation framework comprising of C-, V- and L-stages is presented in Figure 2. Before the framework can be used, initial TTP and a scenario where it is meant to be used must be defined according to the overall objectives of TTP T&E. The scenario describes the friendly and enemy aircraft involved, and their primary goals. TTPs are a way to describe how the friendly aircraft can best achieve their goals in the given scenario. TTPs are typically briefed to pilots in the form of a timeline relative to enemy aircraft. The aircraft and the systems used in the scenario are modeled in C- and V-simulations. These models are

necessarily slightly incomplete abstractions of reality. This does not, however, impose a significant challenge for the L-V-C framework, as the L-stage of TTP T&E is conducted using real aircraft and systems.

The quantitative rule values and qualitative rule descriptions of the initial TTP are based on the best available assumptions and practises. The L-V-C simulation framework is used to identify the operationally HMP optimal values or descriptions for some, or all, of those rules. It can be used to identify operationally HMP optimal rules for a whole flight, an element or just for a single pilot.

[insert Figure 2]

2.2 Measures in V- and L-stages

2.2.1 NP Measure

The term 'normative performance' (NP) used in this paper is adopted from the field of decision analysis, where normative decision models explain how ideal people should make decisions under specific assumptions of rationality (see, e.g.,[18]). NP of such ideal people can be viewed as perfect. Real-life decisions, however, do not coincide with these ideal decisions as real people typically do not conform to all theoretical assumptions. Therefore, in general terms, NP represents the difference between the decisions of ideal and real people. The closer to the decisions of ideal people the real-life decisions are, the better NP.

In the context of the L-V-C simulation framework, NP describes how accurately a pilot adheres to directed TTP rules in a V- or L-simulation. The NP measure consists of a selected set of quantitative and qualitative TTP rules. The rules included in the measure are selected such that NP provides a representative picture of the pilot's rule adherence. The score for the NP measure is based on the pilot's accuracy of adhering to the values and verbal descriptions of the rules.

2.2.2 SA Measure

In the L-V-C framework, SA describes the level of agreement between the pilot's understanding of the state of the scenario and the actual state of the scenario. An array of SA measuring techniques have been developed, including both self-rating or subjective [19-20] and objective techniques.[21-22] Objective SA measurement techniques typically assess the level of

pilot's SA using questions, or probes, to capture the pilot's knowledge structures about the task situation. In existing objective techniques, the probes are introduced during the task. The probes are formulated such that the correct answers require the knowledge necessary to build and maintain SA. In an air combat task, the pilot's responses are compared to the real state of the air combat scenario, and the number of correct answers is used as a score of SA. However, administration of the probes during the air combat task is disruptive in V-simulations, especially if it is necessary to freeze the simulation for the duration that the data are collected, and is impossible in L-simulations. Another criticism of objective SA measurement techniques is that they are a test of working memory, rather than understanding of the evolving flight situation.[23]

Self-rating SA measurement techniques rely on the pilots' ability to rate their SA on a pre-defined scale. While the subjective rating can be administered post-task, causing little or no disruption to the air combat task, subjective measurement techniques suffer from a host of problems. These include, e.g., a possible correlation between NP and subjective SA ratings, and a more fundamental issue about the pilot's inability to be aware of what he or she is not aware – a phenomenon known as unknown unknowns.[24-25]

In the L-V-C simulation framework, a post-trial modification of objective SA measurement techniques is used to measure SA. The post-trial modification administers probes after L- and V-simulations, thereby avoiding disruptions during the air combat task. Furthermore, compared to the detailed questions used in existing objective techniques, the modification uses broad questions. By doing so, use of the technique attempts to unveil some of the complex knowledge structures pilot uses to describe, explain and predict the air combat scenario[26], as well as to reason and to make tactical decisions.[27] The breadth and depth of the probes is a balance between the array of the knowledge structures required in air combat and the time available to elicit them in a natural task environment.

When the post-trial SA measurement technique is used in the L-V-C simulation framework, pilots attend a normal debrief where the mission playback is paused at predetermined times and probes tapping the three levels of SA are introduced. The pilots answer the probes with reference to their cockpit recordings and recollection of the air combat scenario. The responses from the pilots are then compared with the real state of the scenario. The number of correct answers defines the SA score on its three levels. A few characteristics of this post-

trial measurement technique should be noted. While it is fundamentally a self-rating technique as it relies on pilots' ability to recall past events; fighter pilots are trained to compare their recollections of the scenario and the real state of the scenario. In fact, such comparisons form the foundation of every fighter pilot training.[20,28] However, the technique is suited only for non-punitive SA assessment, such as TTP T&E, where possible pilot bias is minimal.

2.2.3 MWL Measure

A variety of techniques are available to assess MWL. Most techniques can be categorized either as behavioral, subjective or physiological techniques.[29] Subjective techniques utilize operators' subjectively experienced MWL, i.e., how they feel when doing a task.[30] Behavioral techniques assume that an operator's ability to conduct cognitive tasks diminishes as MWL increases.[31] Finally, when the physiological techniques are used, variations in MWL are identified by measuring MWL induced physiological changes on an operator.[32-33]

Behavioral techniques are disruptive and cannot be safely used in L-simulations.[34] It is also argued that behavioral measures should not even be used as measures of MWL.[29] In comparison, physiological measures seldom take place in isolation. The second and third order physiological effects and bodily interactions may generate physiological responses, which can be falsely interpreted as MWL responses.[35] The non-intrusiveness, ease of use, and low-cost implementation of the subjective MWL measures are some of the features that motivate their usage in V- and L-simulations. The subjective measures have been successfully employed to assess pilots' MWL on many occasions.[5, 36-37]

A non-weighted NASA-TLX[38] is used as the MWL measure in the L-V-C simulation framework. NASA-TLX is practical and easy to administer even in L-simulations, and the non-punitive context of TTP T&E greatly reduces potential pilot biases typically related to subjective MWL measures. NASA-TLX uses six different dimensions to assess MWL: mental demand, physical demand, temporal demand, frustration, effort, and performance. NASA-TLX considers performance as the pilot's subjective opinion about the success or failure in meeting the task requirements. As such, it should not be confused with NP or HMP output used in this paper. The multi-dimensional approach increases the MWL measure's diagnosticity, as it helps

identifying why MWL is high or low. Unlike NP and SA scores, a single MWL score for each dimension is determined for the complete V- or L-simulation run.

2.3 C-stage

Quantitative rules are only analyzed in C-stages. In the first C-stage, the quantitative rules of the initial TTP are implemented into C-simulation. In C-stage, the enemy aircraft follow the scenario determined in the initialization of TTP T&E. The rule values of interest - determined by the overall objective of TTP T&E - are adjusted. A required number of simulation runs are conducted until MP optimal values maximizing the optimization criterion P_k and fulfilling the constraint P_s=1 are found. Since some of the quantitative rule values are related to weapon employment parameters, e.g., missile launch ranges, the estimation of P_k is based on a ratio of friendly weapon launches resulting in a kill and a total number of friendly weapon launches. Similarly, the estimation of P_s is based on a ratio of enemy weapon launches resulting in a miss and a total number of enemy weapon launches. P_k and P_s are estimated separately for each missile launch and their averages are used as the measures of MP output reflecting the primary goal of the flight. While these measures are suitable for addressing the flight's goal achievement, the L-V-C simulation framework itself does not limit the forms of the optimization criterion and constraint used. It should be noted that the rule adjustments are made not just for the constructive entities corresponding to the pilots of the flight whose rules are of interest, but for all friendly aircraft for which the rules affect the rules of interest. For example, if TTP assumes a certain formation within an element, rules for both the element leader and the wingman are adjusted to maintain that formation - even if only the rules of the wingman are of interest.

C-stage is repeated after V- or L-stages if SA, MWL, NP or HMP output are subsequently found to be unacceptable (see the dashed lines from V- and L-stages to C-stage in Figure 2). In this case, the results of the V- or L-simulations reveal those quantitative rules whose values should be adjusted to promote better SA, MWL, NP or HMP output. When C-stage is repeated, the original optimization criterion, i.e., the maximization of P_k, is relaxed by the minimization of $(P_k\text{-}P_{kref})^2$, where the reference probability of kill, denoted by P_{kref}, is selected based on the results of V- or L-stage as well as on the optimal values of P_k obtained in earlier C-stages. If a need for modifying the quantitative rule values to improve NP, increase SA or lower MWL is recognized in V- or L-stage, a lower value of P_{kref} than the optimal P_k in the previous C-stage

must be selected. As a result, the solution of C-stage is likely to promote the required changes of NP, SA and MWL, and to decrease the optimal value of P_k – while still meeting the constraint P_s=1. The L-V-C simulation framework does not restrict relaxing or even removing the constraint P_s=1, if friendly losses are accepted. It is also possible to introduce new optimization criteria that do not explicitly depend on P_k. As the optimization criterion and constraint of the repeated C-stage are based on the analysis of the V- or L-simulations, the resulting MP optimal values of the quantitative rules are now implicitly affected by SA, MWL and HMP output, whereas in the first C-stage these are ignored.

2.4 V-stage

V-stage considers both the qualitative rules and the MP optimal quantitative rules originating from C-stage. The qualitative rules of the first V-stage are the ones defined in the initial TTP. When V-stage is repeated, the qualitative rules originate from the preceding V- or L-stage (see the dashed line from L-stage to V-stage, and the dotted line from V-stage to V-stage in Figure 2). By iterating V-stage with the fixed quantitative rules, it is possible to refine the qualitative rules' verbal descriptions within V-stage.

The pilots whose operationally HMP optimal rules are of interest fly the V-simulator as participants. Their NP, SA, MWL, and HMP output are recorded, and HMP output is measured using P_k and P_s. Since HMP output should reflect the achievement of the flight's primary goal as a whole, the estimation of P_k is based on the ratio of enemy aircraft alive at the beginning and at the end of the simulation, whereas the estimation of P_s is based a ratio of friendly aircraft alive at the beginning and at the end of the simulation. All but the aircraft flown by the participants are in a supporting role and are implemented in the V-simulation as constructive simulation entities. The friendly constructive entities are set to follow the qualitative and MP optimal quantitative rules derived in the preceding stages of TTP T&E - effectively making the V-simulation mirror the C-simulation and forcing it to evolve similarly for all participants flying in it. The enemy aircraft follow the same scenario as in C-stage. The participants are tasked to follow the directed qualitative rules and MP optimal quantitative rules, but they are not told how the scenario unfolds. As the participants fly the V-simulator, their SA, MWL and NP, together with P_k and P_s, are measured.

NP, SA and MWL scores are analyzed at the completion of V-stage. In data analysis, the probability level for a significance result can be chosen before the data collection. For example, in Part 2^{16} of this study, $p<0.05$ is used. If P_k and P_s are unsatisfactory, the analysis aims to identify the rules that could be revised to improve the overall HMP output. If, however, P_k and P_s are satisfactory, the objective is to identify the rules that could improve NP, SA or MWL. As NP describes adherence of rules, low NP scores can be associated with specific rules. The rules with the lowest NP scores are candidates for revision. In contrast, neither post-trial SA probes nor NASA-TLX are associated with specific rules. The post-trial SA measurement technique provides a separate SA score for every probe within the air combat task, and NASA-TLX provides a MWL score (across six of its dimensions) for the whole flying task. If SA or MWL scores are low, subject matter experts (SMEs) must analyze the progression of the scenario and the implementation of TTP, and identify which rules are most likely to cause such undesired scores. The identified rules are candidates for revision and the ones with the most potential to improve NP, SA, MWL, P_k and P_s are modified. If the quantitative rules are modified, TTP is returned to C-stage without modifying the qualitative rules (see the dashed line from V- to C-stage in Figure 2). If the qualitative rules are modified, V-stage is repeated with refined verbal descriptions of the participants' qualitative rules (see the dotted line from V- to V-stage in Figure 2). The constructive entities' qualitative rules are adjusted only if they affect the participants' ability to adhere to their rules.

Each time V-stage is repeated, NP, SA, MWL, P_k and P_s are compared to those of the preceding V-stage. The comparison is made both qualitatively using non-statistical methods and SME judgement, and statistically when the aim is to identify significant differences between the simulations' P_k, P_s, and NP, SA, and MWL scores. If the results have improved or remained unchanged, TTP T&E can progress to L-stage. Alternatively, both type of comparisons can reveal further TTP improvement opportunities. Finally, deteriorating results indicate a need to re-analyze TTP and to recognize some new ways of modifying it. Once the outcome of V-stage is satisfactory, HMP optimal rules - both qualitative and quantitative - in the simulated environment are obtained.

2.5 L-stage

In L-stage, the HMP optimal rules obtained in V-stage are evaluated in a real-life environment. Due to the real-life task complexity and stressors, L-simulations often result in lower SA,

higher MWL and lower NP than those seen in V-simulations. While L-stage provides scores for SA, NP and MWL in the real-life environment, the decision about their acceptability is a military judgement call taken by the SMEs.

Real aircraft and pilots are used in L-stage. The participants are the pilots whose operationally HMP optimal rules are of interest. The participants are tasked to follow the HMP optimal rules defined in V-stage. They are given a standard flight briefing but are not told how the scenario unfolds. In addition to the participants, supporting pilots are needed to operate the other aircraft required in L-simulation. First, supporting friendly pilots are needed to fly the friendly aircraft not flown by the participants. The supporting friendly pilots are briefed to follow the same rules as the constructive entities did when the participants' HMP optimal rules were identified in V-stage. Second, supporting enemy pilots are needed to fly the enemy aircraft. The supporting enemy pilots are briefed to follow the same scenario used in C- and V-stages.

After each L-simulation, HMP output, measured by P_k and P_s, and the scores of NP, SA and MWL are recorded. Here, P_k and P_s are estimated in a same way as in the V-simulations. Due to limited availability of fighter aircraft and operational pilots, the number of simulation runs and hence the sample size may be restricted at L-stage. Therefore, the results of V- and L-stages must be comprehensively compared to establish whether the results from the two stages are balanced or not. A balanced result means that L-stage's P_k and P_s are acceptable, and based on a military judgement, the scores of NP, MWL and SA are not significantly worse than those obtained at V-stage. If this is not the case, the SMEs should use the NP, SA and MWL scores to identify potential rules to be revised in a same fashion as at V-stage. Then, TTP is returned to C- or V-stage depending on the need for either qualitative or quantitative rule adjustments (see the dashed lines from L-stage to C- and V-stages in Figure 2).

If V- and L-stages' results are balanced, TTP T&E is complete and HMP optimal rules evaluated at L-stage can be cleared for operational use. In other words, the resulting operationally HMP optimal rules ensure that a flight having decent NP, SA and MWL can achieve its primary goal in a real-life environment.

3. DISCUSSION

This paper is Part 1 of the two-part study that introduced the L-V-C simulation framework, where the flight's initial TTP is developed into operational TTP with acceptable HMP output,

NP, SA and MWL. In the framework, MP optimal quantitative rules are obtained at C-stage, HMP optimal qualitative rules are identified in V-stage, and HMP optimal qualitative and quantitative rules are evaluated in L-stage using real aircraft and pilots. The iterative and multipart nature of the framework enables that V- or L-stages can reveal flaws of the TTP and an inadequate TTP can be returned to C- or V-stage for revision.

C-, V- and L-simulations have been widely used separately in earlier TTP T&E studies (see, e.g., [39-42]). However, if only C-simulations are utilized, the ideal decision making of pilots can only be assumed. Such TTP T&E results in MP optimal quantitative rules but without considering how HMI affects HMP output in V- and L-simulations, not to mention in real combat operations. The optimal selection of qualitative rules' verbal descriptions is not possible in C-simulations. Therefore, V-simulations are needed to evaluate the impact of pilots' NP, SA and MWL in a safe environment. Finally, before TTPs can be released into operational use, they must be evaluated in L-simulations. Alternatively, if only V-simulations are conducted, it would be challenging to determine MP optimal quantitative rules. The time and the number of participants required for statistically relevant simulation results would soon exceed the cost constraints related to time and manpower. Even if TTP T&E could be completed in V-simulations, it could not become established whether HMP output, NP, SA and MWL would remain acceptable in a real flying environment. Hence, both C- and L-simulations are needed to complement the V-simulations. The L-V-C simulation framework provides a safe and cost-effective way of utilising the strengths of each simulation class by sequencing and iterating their use during TTP T&E.

Pilots' SA, MWL and performance have been extensively studied - both separately and together (see e.g., [8, 42]). Unlike the previous studies, the L-V-C simulation framework explicitly illustrates how NP, SA and MWL results, together with HMP output of V- and L-simulations, can be used to identify the qualitative and quantitative rule candidates for revision, and to support the formulation of optimization criteria and constrains for C-simulations.

In summary, the L-V-C simulation framework presents a multifaceted approach for TTP T&E. The framework uses multiple models to investigate TTP, to increase the transparency and validity of the simulation study, and to ease the interpretation of simulation results – an approach often recommended in simulation and the modeling literature. By using diverse simulation classes and models, the framework minimizes the impact of the individual

simulation models' and sub-models' incompleteness on the outcome of TTP T&E and thus improves the reliability of this outcome.

In Part 2^{16} of this two-part study, the use of the L-V-C simulation framework is demonstrated. In the demonstration, C-simulations are conducted using a constructive Air Combat Evaluation Model (ACEM). ACEM is a Raytheon built air combat simulation, typically used for studying operational-level requirements, preliminary designs and tactical utility of TTPs at the engagement level. V-simulations in Part 2^{16} are run in a Boeing built Weapon Tactics and Situational Awareness Trainer (WTSAT). WTSAT is a non-motion, high-fidelity flying simulator used for basic and advanced F/A-18 pilot training. L-simulations in Part 2^{16} utilize real F/A-18C aircraft and operational F/A-18 pilots. While both parts of the study concentrate on air combat, the principles of the L-V-C simulation framework are domain independent. As long as there are suitable C-, V- and L-simulation models, the same methodology can be applied to any civil or military task where HMI is of concern.

REFERENCES

1. Schreiber B, Schroeder M and Bennett W Jr. Distributed mission operations within simulator training effectiveness. Int J Aviat Psychol 2011; 21: 254-268.
2. Rajabally E, Valiusaityte M and Kalawsky R. Aircrew performance measurement during simulated military aircrew training: a review. In: AIAA Modeling and Simulation Technologies Conference, Chicago, IL, 10-13 August 2009, pp. 5829-5838. Reston, VA: AIAA.
3. Haase C, Hill R and Hodson D. Planning for LVC simulation experiments. Appl Math 2014; 5: 2153-2162.
4. Bolstad C and Cuevas H. Integrating situation awareness assessment into test and evaluation. ITEA Journal 2010; 31: 240-246.
5. Mansikka H, Virtanen K and Harris D. Comparison of NASA-TLX scale, modified Cooper-Harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks. Ergonomics 2019; 62: 246-254.
6. Wickens C. Multiple resources and performance prediction. Theor Issues Ergon Sci 2002; 3: 159-177.
7. Endsley M. Toward a theory of situation awareness in dynamic systems. Hum Factors 1995; 37: 32-64.
8. Mansikka H, Virtanen K and Harris D. Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. IEEE Trans Hum-Mach Syst 2019; 49: 1-9.
9. Diallo S, Padilla J, Papelis Y, Gore R and Lynch C. Content analysis to classify and compare live, virtual, constructive simulations and system of systems. JDMS 2016; 13: 367-380.
10. Hodson D and Hill R. The art and science of live, virtual, and constructive simulation for test and analysis. JDMS 2014; 11: 77-89.
11. Virtanen K, Raivio T, Hämäläinen R. Decision theoretical approach to pilot simulation. J Aircr 1999; 36: 632-641.

12. Virtanen K, Raivio T and Hämäläinen R. Modeling pilot's sequential maneuvering decisions by a multistage influence diagram. J Guidance Control Dyn 2004; 27: 665- 677.
13. Hodson D and Baldwin R. Characterizing, measuring, and validating the temporal consistency of live-virtual-constructive environments. Simulation 2009; 85: 671-82.
14. Millar J, Hodson D, Peterson G and Ahner D. Data quality challenges in distributed live-virtual-constructive test environments. ACM J Data Inf Qual 2016; 7: 1-3.
15. Zhihua D, Yuanchang Z, Yanqiang D and Xianguo M. Constructing LVC simulation environments based on legacy simulations. In: Proceedings of the 2013 international conference on virtual reality and visualization (ICVRV), Xian, 14-15 September 2013, pp. 325-328. Washington, DC: IEEE Computer Society Press, 2014.
16. Mansikka H, Virtanen K, Harris D and Salomäki J. Live-virtual-constructive simulation for testing and evaluation of air combat tactics, techniques and procedures, Part 2: Demonstration of framework. JDMS (submitted for publication).
17. Hodson D, Esken B, Gutman A and Hill R. Quantifying radar measurement errors in a live-virtual-constructive environment to determine system viability: a case study. JDMS 2014; 11: 115-124.
18. Bell D, Raiffa H and Tversky A. Descriptive, normative, and prescriptive interactions in decision making. In: Bell D, Raiffa H and Tversky A (eds) Decision making: Descriptive, normative, and prescriptive interactions. Cambridge: Cambridge University Press, 1988; pp. 9-32.
19. Taylor R. Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. In: Salas E (ed) Situation Awareness. London: Routledge, 2017, pp. 111-128.
20. Waag W and Houck M. Tools for assessing situational awareness in an operational fighter environment. Aviat Space Environ Med 1994; 65: A13-A19.
21. Endsley M. Situation awareness global assessment technique (SAGAT). In: Proceedings of the IEEE 1988 national aerospace and electronics conference, Dayton, OH, 23-27 May 1988, pp. 789-795. New York, NY: IEEE, 1988.
22. Pritchett A and Hansman R. Use of testable responses for performance-based measurement of situation awareness. In: Endsley M and Garland D (eds) Situation awareness analysis and measurement. Mahwah, NJ: CRC Press, 2000, pp. 170-189.
23. Durso F, Bleckley M and Dattel A. Does situation awareness add to the validity of cognitive tests? Hum Factors 2006; 48: 721-733.
24. Mitzen J and Schweller R. Knowing the unknown unknowns: misplaced certainty and the onset of war. Security Studies 2011; 20: 2-35.
25. Pawson R, Wong G and Owen L. Known knowns, known unknowns, unknown unknowns: the predicament of evidence-based policy. Am J Eval 2011; 32: 518-546.
26. Rouse W and Morris N. On looking into the black box: prospects and limits in the search for mental models. Psychol Bull 1986; 100: 349-363.
27. Jones N, Ross H, Lynam T, Perez P and Leitch A. Mental models: an interdisciplinary synthesis of theory and methods. Ecology and Society 2011; 16: 46-60.
28. Aronsson S, Artman H, Lindquist S, Mitchell M, Persson T, Ramberg R, Romero M and ter Vehn P. Supporting after action review in simulator mission training: co-creating visualization concepts for training of fast-jet fighter pilots. JDMS. Epub ahead of print 14 January 2019. DOI: 10.1177/1548512918823296.
29. O'Donnell R and Eggemeier F. Workload assessment methodology. In: Boff K, Kaufman L and Thomas J (eds) Handbook of Perception and Human Performance. New York, NY: John Wiley and Sons, 1986, pp. 2-43.

30. Johanssen G, Moray N, Pew R, Rasmussen J, Sanders A and Wickens C. Final report of experimental psychology group. In: Moray N (ed) Mental Workload. NATO Conference Series, vol 8. Boston, MA: Springer, 1979, pp. 101-14.
31. Muse L, Harris S and Feild H. Has the inverted-U theory of stress and job performance had a fair test? Human Performance 2003; 16: 349-364.
32. Mansikka H, Simola P, Virtanen K, Harris D and Oksama L. Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. Ergonomics 2016; 59: 1344-1352.
33. Mansikka H, Virtanen K, Harris D and Simola P. Fighter pilots' heart rate, heart rate variation and performance during an instrument flight rules proficiency test. Appl Ergon 2016; 56: 213-219.
34. Casali J and Wierwille W. On the measurement of pilot perceptual workload: a comparison of assessment techniques addressing sensitivity and intrusion issues. Ergonomics 1984; 27: 1033-1050. 50. Hart S, McPherson D and Simpson C. Airline pilot time estimation during concurrent activity including simulated flight. In: 47th Annual meeting of the aerospace medical association, Bal Harbour, FL, May 1976.
35. Lutfi M and Sukkar M. Effect of blood pressure on heart rate variability. Khartoum Medical Journal 2011; 4: 548-553.
36. Casner S. Perceived vs. measured effects of advanced cockpit systems on pilot workload and error: are pilots' beliefs misaligned with reality? Appl Ergon 2009; 40: 448-456.
37. Lee Y-H and Liu B-S. Inflight workload assessment: comparison of subjective and physiological measurements. Aviat Space Environ Med 2003; 74: 1078-1084.
38. Hart S and Staveland L. Development of NASA-TLX (task load index): results of empirical and theoretical research. Advances in Psychology 1988; 52: 139-183.
39. Poropudas J and Virtanen K. Game-theoretic validation and analysis of air combat simulation models. IEEE Trans Syst Man Cybern Part A Syst Humans 2010; 40: 1057- 1070.
40. Poropudas J and Virtanen K. Simulation metamodeling with dynamic Bayesian networks. Eur J Oper Res 2011; 214: 644-655.
41. Hill R, Miller J and McIntyre G. Simulation analysis: applications of discrete event simulation modeling to military problems. In: Proceedings of the 33nd conference on Winter simulation, Arlington, VA, 9-12 December 2001, pp. 780-788. Washington, DC: IEEE Computer Society Press.
42. Endsley M. A survey of situation awareness requirements in air-to-air combat fighters. Int J Aviat Psychol 1993; 3: 157-168.

FIGURE 1

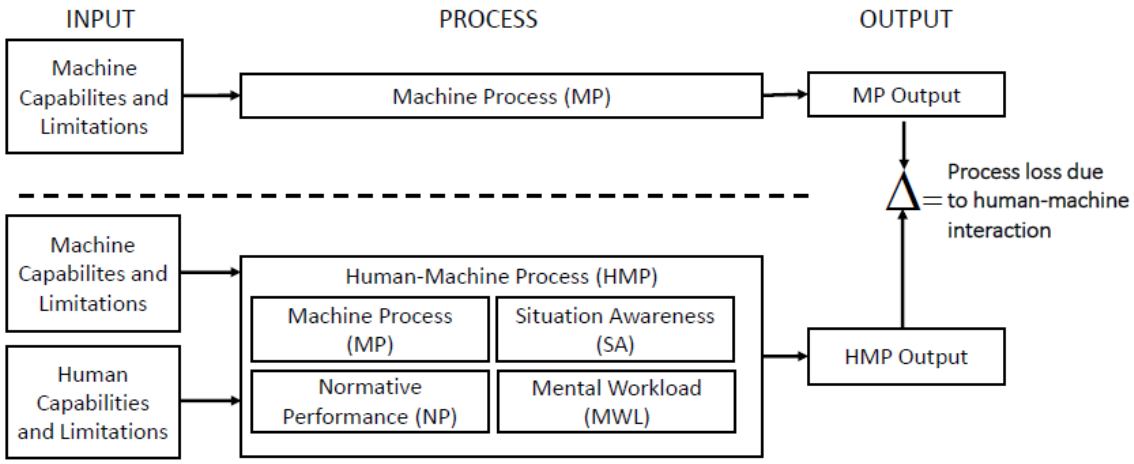
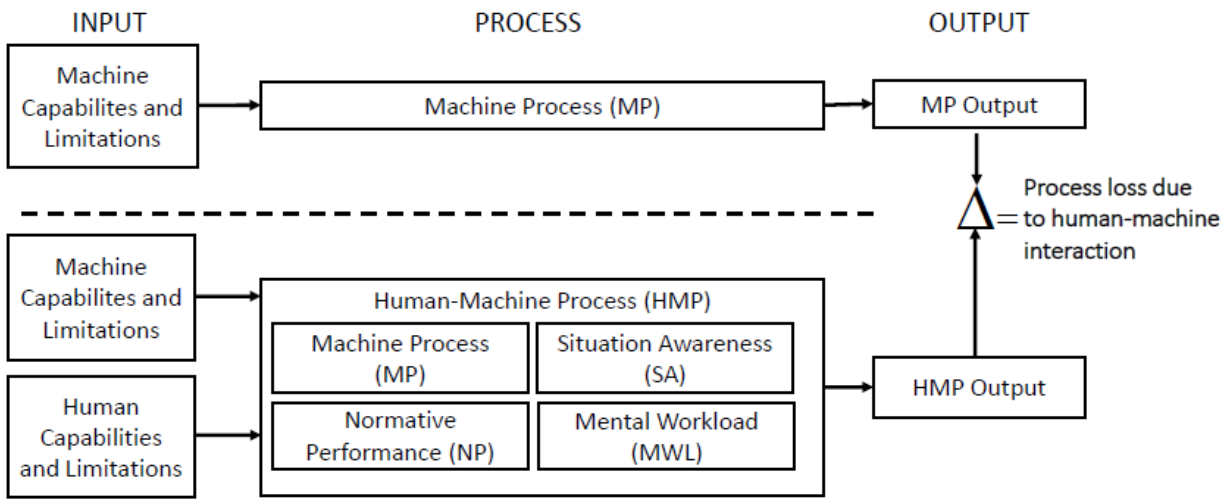

INPUT
PROCESS
OUTPUT
Machine Capabilites and Limitations
Machine Process (MP)
MP Output
Process loss due to human-machine interaction
Machine Capabilites and Limitations
Human Capabilities and Limitations
Human-Machine Process (HMP)
Machine Process (MP)
Situation Awareness (SA)
Normative Performance (NP)
Mental Workload (MWL)
HMP Output

FIGURE 2


INPUT
PROCESS
OUTPUT
Machine Capabilites and Limitations
Machine Process (MP)
MP Output
Process loss due to human-machine interaction
Machine Capabilites and Limitations
Human Capabilities and Limitations
Human-Machine Process (HMP)
Machine Process (MP)
Situation Awareness (SA)
Normative Performance (NP)
Mental Workload (MWL)
HMP Output