Towards the Learning Organisation

Frameworks, Methods, and Tools for Resource-Efficient and Effective Training

HENRIK ARTMAN, JONATHAN BORGVALL, MARTIN CASTOR,
HANS JANDER, SINNA LINDQUIST, ROBERT RAMBERG

FOI-R--3711--SE
ISSN 1650-1942 September 2013

Henrik Artman, Jonathan Borgvall, Martin Castor,
Hans Jander, Sinna Lindquist, Robert Ramberg

Towards the Learning Organisation

Frameworks, Methods, and Tools for Resource-Efficient and Effective Training

Titel	Mot den lärande organisationen
Title	Towards the Learning Organisation
Rapportnr/Report no	FOI-R--3711--SE
Månad/Month	September
Utgivningsår/Year	2013
Antal sidor/Pages	34 p
ISSN	1650-1942
Kund/Customer	FM HKV/Swedish Armed Forces HQ
Forskningsområde	3. Flygteknik och luftstridssimulering
FoT-område	Ledning och MSI
Projektnr/Project no	E36714
Godkänd av/Approved by	Lars Höstbeck
Ansvarig avdelning	Informations- och aerosystem

Sammanfattning

Syftet med denna rapport är att beskriva modeller för operativ träning av militär personal samt ramverk, metoder, och verktyg som stöder analys, planering, uppföljning, och utvärdering av träning. Som exempel används i huvudsak simulatorbaserad träning och övning, och i synnerhet verksamheten vid FOI/FLSC (Totalförsvarets Forskningsinstitut/ Flygvapnets Luftstridssimuleringscenter). Rapporten inkluderar även konkreta rekommendationer kring hur ramverk, metoder, och data från utvärderingsverktyg kan användas av en organisation för effektiv träning, och föreslår ett antal forsknings- och utvecklingsaktiviteter som strävar mot en lärande organisation.

Nyckelord: träning, simulering, pedagogik, träningseffektivitet, prestationsvärdering, verktyg, genomgång efter övning, FLSC

Summary

The purpose of this report is to describe models for operational training of military personnel and frameworks, methodologies, and tools that support the analysis, planning, monitoring, and evaluation of such training. The primary example used is simulator-based training and exercises, and in particular the operations at FOI/FLSC (Swedish Defence Research Agency/Swedish Air Force Combat Simulation Centre). The report also includes specific recommendations on how frameworks, methodologies, and data from assessment tools can be used by an organization for effective training, and suggests a number of research and development activities that strive towards a learning organization.

Keywords: training, simulation, pedagogy, training effectiveness, performance measurement, tools, after action reviews, AAR, FLSC

Table of contents

1 Introduction

This introduction provides a theoretical foundation regarding various frameworks, methods, and tools that can be utilized as the foundation of a learning organization.

1.1 The Need for Pedagogical Models and Supporting Tools in Training

Military operations are in many aspects uncertain and dynamic. Military organizations must strive towards each individual understanding both the intents of peers within the organization, implications of enemy actions, and how to respond effectively. To function successfully, as an organization under conditions of uncertainty, one must train and exercise not only for the expected but also for the unexpected. Training for knowledge and skills which can be adapted, at the individual and collective levels, demands both the structure of standard/doctrinal operations as well as training on disruptions and atypical situations.

The learning organization denotes a perspective that focuses on learning as a continuous and collective process, as much as an individual process. The idea is to stress that operational performance in complex and dynamic environments needs to occur in a fluent manner with as little disruptions or doubts as possible. It does not imply that all actions must be predictable nor that training and learning must be based on conducting the right actions - quite the opposite: the true learning organization admits elaboration and is prepared for change - to adapt to the unpredictable.

To commit mistakes in a training session is an opportunity for learning - as long as guidance and feedback is provided by peers. Organizations must be able to design training sessions to stimulate the elaboration and development of the individual as well as the team knowledge and skills. In order to train an organization one must communicate clearly what the plan is, the intent, the results, and ultimately what could have been done differently. Aquinis & Kraiger (2009), for example, stress the importance of allowing trainees to make mistakes during training that are later analysed and reflected, under guidance, during after action reviews (AAR). The rationale is that this develops a deeper understanding for the dynamics of the environment and exploration of a larger set of outcomes, beyond the preferred or expected ones. This perspective differs from traditional training that typically has had the purpose of showing “the right way”, and hence to minimize mistakes also during training.

The larger the scope of the learning is, such as team or organizational learning, the greater is the need for a structured approach as well as appropriate evaluation. Timely and appropriate performance feedback is crucial in order to enable a deliberate practice (Ericsson, Krampe, & Tesch-Römer, 1993) and develop expertise. The tools discussed in this report are means to that end as they support the debriefing as well as the instructors’ assessments.

The conclusions concerning the use of simulators provided by Salas, Bowers, & Rhodenizer (1998), summarized as “it is not how much you have but how you use it” is still valid. Their recommendation concerning future development is that simulation development has to focus on training instead of technology. This implies that more sophisticated and appropriate frameworks, methods, tools, and measures of effectiveness of simulation-based training systems must be developed. It is also necessary to abandon the notion that simulation equals training and that higher fidelity in the simulation alone means better training.

Performance measurements tools such as those described later in this current report are important enablers for this. The report is intended to provide the foundation for further directions of research and development along several related activity threads, that include both technology and methodology development. Many of the examples in the report are based on the authors' personal experiences from primarily air combat training, but they are in essence equally valid for training and human performance assessment in most other military sectors and many civilian domains such as police & fire services, first-responders, civil aviation and maritime shipping, to mention a few.

1.2 Pedagogical Models of Learning

Organized education and training occurs in a context, an educational environment, according to a determined structure - a pedagogical model. However, in many cases there has been no active and informed decision of the actual pedagogical model to build upon. The tendency is rather to do what others do or go about doing things the way you have always done them. There are several different theoretical perspectives to use as the foundation when planning and implementing training programs and environments.

One approach that has been very influential is based on the theoretical perspective of behaviourism. In behaviourism, there is a focus on observable and discernible behaviour, that is, actions that can be directly observed (Henley, 2003). The behaviourist approach to learning is that we learn by receiving stimuli from the outside world and change our behaviour until we react with an appropriate response. Behaviourism thus regards learning as a systematic process consisting of receiving stimuli and reacting with a response, where the goal of an instructor is to generate observable change in behaviour by either rewarding or punishing certain responses. When learning, humans (re)shape behaviour as a result of being subjected to various stimuli, consciously or unconsciously (Lefrancois, 1997). In this perspective, the behaviourist reason that there is no need to focus on the cognitive and mental processes.

Unlike behaviourism, cognitively oriented perspectives have a strong focus on individual and mental processes. A basic assumption in strong cognitive oriented approaches, cognitivism, is that people always have conscious, or unconscious, intentions with their actions. There is thus an intentional stance in learning assuming that the learner is an active learner, consciously or unconsciously. In this perspective, learning is based on more than what is directly perceived through our senses, and experiences, interpretations and reflections of what is perceived affects the learning process. Motivation is in this context an important aspect and considered to be an internal property of the learner. Another cognitively oriented approach is constructivism. There are similarities between these two approaches, but one difference is that cognitivism has a distinct focus on individual and mental processes, whereas constructivism to a larger degree emphasize social aspects and interaction with "the outside world" during learning.

Other perspectives, such as e.g. Kolb's learning styles (Kolb, 1983), build more strongly on the "uniqueness" of individuals in terms of circumstances, personality, and experience. This means that learners have different opportunities to embrace, interpret, understand, and benefit from different learning situations. Here the importance of fulfilling the needs of the learner is emphasized strongly. The instructors' role in this context is to facilitate the learning process. Yet other perspectives, often referred to as socio-cultural perspectives on learning, emphasize more strongly the social aspects of learning and the specific context in which learning takes place, e.g., situated learning (e.g., Lave & Wenger, 1991). Learning is an everlasting process that takes place everywhere and all the time, not just in contexts and environments that have been denoted as a learning environments (Henley, 2003).

The perspectives presented above obviously emphasize different and to some extent complementary aspects important to learning, for example in terms of what can be expected by the learner and the role of an instructor. Building on the strengths of complementary perspectives on learning accounted for above, Ramberg, Nählinder, Borgvall, & Castor (2012) describes a number of central tenets of learning that needs to be considered for any efficient training setup. Students must actively build their own understanding and support for this need to be provided to the students. The student must actively process information, for example through discussions with peers and/or instructors under “safe” conditions. I.e. the strengths of collaborative and social aspects of learning need to be taken into account. The student needs sufficient time in order to assimilate new knowledge. Learning contexts and environments thus need to be dynamic to take this aspect into account. The student should be faced with contextualized challenges/tasks that while being on an adequate level of difficulty still challenges and stimulates elaboration/experimentation of the new knowledge. Compare for example with the concept of Zone of Proximal Development (Vygotsky, 1976). This puts strong demands on planning of training and the role of the instructor to successfully scaffold the student (Lave & Wenger, 1991). Detailed, timely and constructive feedback supports both learning and the students own appraisal of performance. Adhering to these aspects support student motivation and individual interest - deliberate practice - that are important for learning and maintaining an interest to develop skills and competencies. Finally but not the least, the assessment and aspects that are decided to be assessed of course affect what, when, and how students learn. Adhering to these aspects when planning and carrying out of training of course in turn puts demands on use of different methodologies tailoring different important and complementing measures. There is a need for a mixed-method approach to empirically evaluate training (Jander, Borgvall, & Castor, 2011; Stanton, Hedge, Brookhuis, Salas, & Hendrick, 2005).

The importance of feedback and guidance has been identified as an essential part of any training effort since many years (e.g., Holding, 1987), and emphasis on the importance has accelerated over the past two decades following the increased use of simulation and game-based training (e.g., Cannon-Bowers & Bell, 1997; Wickens & Hollands, 2000; Freeman, Salter, & Hoch 2004). Several studies have shown that allowing trainees to practice without specific feedback and guidance of good and bad performance can be detrimental (Wickens & Hollands, 2000), and may generate sub-optimal decision-making skills (Cannon-Bowers & Bell, 1997). In that sense, performance measurement tools is essential for a trainees learning process since it provides observable evidence that facilitates critical thinking of the own performance and that of others (Freeman et al., 2004). It may also support the development of shared mental models and common ground (e.g., Clark, 1996) among the team members, for example while discussing audio-visual replays of a training event (based on performance measurement).

Freeman, Salter, & Hoch (2004) have declared that feedback in debriefs is particularly important in team training since the teamwork itself does not necessarily produce immediate feedback of individual and team performance from which team members can learn during task execution. Further on, feedback offered during task execution is normally processed less well than feedback received immediately after completion since attention must be divided in the former case (Wickens & Hollands, 2000). One commonly adopted insight is that it is important to consider, and if possible reduce, the temporal distance between the training and the after action review (AAR).

1.3 Models for Structuring & Evaluating Training

As a support to organize training there are several different guiding models. Below we present two models. ADDIE is specifically designed for training, while DECIDE is for deciding what and how to evaluate.

1.3.1 ADDIE

ADDIE is a generic, underlying process for structuring a training program, and in that regard independent of factors such as training environment, content, and purpose. ADDIE refers to five different phases: Analyse, Design, Develop, Implement, and Evaluate. ADDIE can be thought of as a supporting, underlying structure for developing and providing effective training. There are several alternative models with the same purpose, such as Instructional Systems Design (ISD) and Successive Approximation Model (SAM). In essence ADDIE, ISD, and SAM are merely variations with the same purpose and structure at large - to guide resource-efficient development in support of effective training and ultimately performance improvement.

ADDIE was initially developed by Branson, Rayner, Cox, Furman, King, & Hannum (1975), with the purpose of formalizing the process of developing military inter-service training. Whereas the initial process was considered hierarchical and sequential, the current applications of the process are typically dynamical and iterative between and across phases.

The overarching purpose of each phase of the ADDIE process could be summarized as:

Analyse

This phase aims to clarify the objectives, needs, and requirements for the training, such as identifying which experiences that are to be provided, and what knowledge and skills that are to be developed. In this phase, questions such as the ones listed below should be addressed:

- Are there any prerequisite knowledge and skill requirements?
- What are the characteristics of the training audience?
- What are the new, desired knowledge and skills requirements?
- What are the pedagogical and learning theory considerations?
- What are the alternative learning environments?

Design

The purpose of the Design phase is to systematically identify and specify for example objectives, instruments, and media for the training. During the design phase, activities such as the examples below should be undertaken:

- design instructional strategies to meet the desired training objectives
- design prototypes (conceptual/physical/digital) of training media and materials
- document the instructional and technical design strategies

Develop

Building on the results of the analysis and design phases, the purpose of the development phase is to generate the training content and infrastructure, for example to develop and integrate training materials with supporting technologies (e.g., simulation). Throughout the development phase the training solution should be tested against the results of the analysis and design phases, for feedback and iterative revision where necessary.

Implement

During the implementation phase, required procedures for instructors and trainees are developed and established, methods and processes for managing and delivering training are finalized, and any required pre-training is defined and structured (e.g., training on a new software that is utilized during the training event). During implementation the conformity and effectiveness of previous phases - analysis, design, and development - is evaluated.

Evaluate

There are two purposes of the evaluate phase: to conduct evaluations of each phase of the process as they unfold, and to evaluate the overall success of a new training program. The former could be considered as an inner evaluation loop offering rapid feedback throughout iterations across the various phases (e.g., evaluating the design during development, or evaluating individual/team training effectiveness and performance improvement during and after implementation), while the latter could be considered as an outer loop evaluating the success of a complete ADDIE process for a new training program (e.g., evaluating a complete ADDIE-based training development project or evaluating organizational effects of a new training program).

1.3.2 DECIDE

DECIDE is a general framework useful for guiding the design of training evaluation. As the name suggests it is supporting decisions of how, what, why, and for what purpose evaluation should pertain. Evaluations must be driven by clear, concise, and appropriate goals that perpetuate both skilled and practical handling of more or less routinized tasks as well as objectives of higher level order. The framework was originally proposed for evaluation of how users interact with computer systems (Preece, Roger, & Sharp, 2002). Training in technology dense environments, such as a cockpit, together with its generality, makes the DECIDE framework relevant also for evaluating training in aviation. The framework proposes six stages described below:

1. **D**etermine the overall goals that the evaluation addresses.
2. **E**xplore the specific questions to be answered.
3. **C**hoose the evaluation paradigm and techniques to answer the questions.
4. **I**dentify the practical issues that must be addressed, such as selecting participants.
5. **D**ecide how to deal with the ethical issues.
6. **E**valuate, interpret, and present the data.

Determine the goals for the evaluation and training

What are the general goals of the training? Who should be able to do what, when and why? This is often a stage which must be prepared at a very beginning when designing the training sessions that are to be evaluated. The goals of the training for example improved skill in handling the system or collaborative tactical performance should then direct the measurements as well as approach to be used, i.e., if one is interested if one training condition gives quicker decisions than another condition an experiment with time measures might be best. In another study the focus might be on how and what subjects are communicating while on a highly dynamic situation and then audio or video-recording and consequent speech-act analysis is more appropriate (c.f. Svensson & Andersson, 2006). These questions might be considered self-evident, but surprisingly often they are also taken for granted, ignored, or even worse both taken for granted and ignored. It is therefore of greatest importance to present and communicate such higher level goals.

Explore questions relevant to the training goals

When the goals have been determined there is a need to further explore or decompose questions that are adjacent to the higher level goal. Evaluative and training questions might not only be directly subordinated the goals, but can also be questions that elaborate and nuance questions around the goal. The goal might be quicker handling of a system, while trust in the system or the training supervisor is a question that is important in order to train the handling. Given the higher goals determined in the former stage evaluators need to explore questions that pertain to these goals.

Choose the evaluation paradigm and techniques to answer the questions

As has been mentioned above, the choice of method and approach to the training and evaluation is subordinated the goals. Experimental studies might be best when determining which of two or more conditions that is the best in some regard, while a field study or detailed case study might be better for more explorative questions. At this point one must decide upon different forms of trade-offs in terms of ethical, practical, and theoretical aspects. The outcome of this phase must specify consistency regarding the paradigm and the methodological approach.

Identify practical issues

Any form of evaluation has practical issues which must be dealt with at the earliest possible opportunity. The most important aspects are users, facilities, budget, and required expertise of the evaluators. These practical details may impose a number of constraints on ambitious approaches. Access to users, both in training and for evaluation of computer systems, is of course of utmost importance. For highly specialised and limited user groups availability is often a problem. Even within a limited group of professionals evaluators might have to decide if they are interested in novices or experts in terms of skills and experience. The choice of facilities is also often limited, but evaluators might have to consider constraints regarding equipment. For example if video-recordings or eye-tracking is to be used specific permits might be needed. Physical integration constraints concerning the number and placement of cameras might be present. Budget is always a constraint which must be kept in mind. The data collection is one thing, but afterwards the analysis of large datasets, not the least if it is qualitative data, can be overwhelming. Expertise issues of users and the trained audience is covered above, but the competence of the evaluators must also be considered. The expertise of the evaluators must cover both domain expertise and expertise in data analysis. Evaluators must secure the competences that are needed and appropriate the time for the analysis.

Decide how to deal with ethical details

The research code always requires evaluators to consider different forms of ethical issues where they must consider issues of privacy and confidentiality. In a close community, where most people know each other, it is especially important that the evaluators will not present individual data in such way that it is possible to know which user/trainee that is scrutinized. During collaborative evaluations where several users/trainees work together or work in such way that each user might be seen it is of utmost importance to facilitate the sessions in such way that one is of courteous nature and that the group is well informed about the circumstances for the evaluation. The sessions should start by giving the user / trainee information about the session(s) and what is expected, measured and how the information is upholding a certain level of privacy and integrity.

Evaluate and interpret the data

The design of the data collection is to a large extent given by the objectives and the approach chosen for the evaluation. The data must after collection be organized, interpreted and presented to a wider public or specific client. Is it possible to use pictures or video sequences of the session? What form of statistical analysis is appropriate given the data? Issues of reliability and validity is always of great importance, i.e., are the measures measuring what's being assumed that they measure (validity) and are the measures consistent in their ability to measure this quality or property (reliability).

The following two sections on frameworks and methods for analysis of training, and tools for evaluation of training, provide specific examples of how ADDIE/DECIDE can be operationalized during the Analysis and Evaluate phases of training.

1.4 Frameworks & Methods for Analysis of Training

In times of continuous introduction of advanced technologies and extended use of automation in complex human-operated systems, the development and strategies for training programs become increasingly important in order to prepare human operators of these systems for current and future operational requirements. To deliberately analyse and design for skill acquisition, retention, transfer, and mitigation of skill decay is critical for successful training of personnel.

Advanced methods for analysis and evaluation of training objectives and performance criteria is essential to continuously develop and adapt personnel, organizations, and operations to dynamic changes of operational requirements. In other words it is essential to define a process for analysing, developing, delivering and evaluating training that inherently provides feedback between and across phases, as well as support the development of relevant measures of the effectiveness of training. This section highlights various methodological approaches primarily applicable during analysis and evaluation of training.

Training Needs Analysis (TNA)

A TNA is primarily conducted to determine where training is needed, what needs to be taught, and who needs to be trained (Goldstein, 1993). The UK Tri-Service Guide to Training Needs Analysis (TNA) for Acquisition Projects (CSE, 2001) is often used as a reference for issues to consider during the design of a TNA.

Task and Work Analysis

An extensive set of task and work analysis methods have been developed by the scientific community in order to describe human work. These methods have different formalisms and scope, suiting different applications and effort/resource levels. For a recent review see for example *The Handbook of Work Analysis* by Wilson, Bennett, Gibson, & Alliger (2012).

Competency Modelling

In competency modelling the competencies that underlie successful performance are in focus more explicitly than in task or work analysis. Competency modelling has received increasing attention from many organizations over the past two decades. Competency modelling strives to distinguish superior from average performers, and describe competencies that are measurable and observable in some way.

Training Criteria

One of the most renowned works on training criteria is Kirkpatrick's training criteria taxonomy (Kirkpatrick, 1959a, 1959b, 1960a, 1960b), which has received widespread attention from researchers and practitioners since its original definition. Kirkpatrick's taxonomy was originally designed as a set of practical recommendations, primarily drawn from Kirkpatrick's personal experience, offering a useful heuristic for what can and should be measured when evaluating the effectiveness of training. Kirkpatrick's taxonomy has been debated by many researchers, and flaws in the taxonomy have been highlighted by for example Alliger, Tannenbaum, Bennett, Traver, & Shotland (1997). They proposed an augmented framework based on a meta-analysis of various training effectiveness studies. Bell & Waag (1998) also presented yet an alternative model for evaluation of training, primarily developed to address evaluations of simulation fidelity and training effectiveness (e.g., Borgvall, 2013). Even though their model was explicitly focused on simulation it was based on Kirkpatrick's training criteria and shows clear similarities with the augmented taxonomy of Alliger et al. (1997).

Castor, Borgvall, & Bennett (2009) describe how the Alliger et al. taxonomy was used as the foundation for evaluations of large scale air force exercises at various levels of training effectiveness evaluation. The different levels of the Alliger et al. augmented taxonomy are briefly described below, with the corresponding levels of Kirkpatrick's taxonomy as well as Bell and Waag's simulator training evaluation model within parenthesis:

- Level 1. Reactions (Kirkpatrick level: Reactions, Bell & Waag level: Utility evaluation).
    - Level 1a. Affective reactions: measures assessing to what extent trainees liked and enjoyed the training.
    - Level 1b. Utility reactions: measures assessing the perceived utility value of the training.
- Level 2. Learning (Kirkpatrick level: Learning, Bell & Waag level: Performance improvement [in-simulator learning]).
    - Level 2a. Immediate post-training knowledge: measures assessing how much trainees know about the training topic directly after the training.
    - Level 2b. Knowledge retention: measures similar or identical to level 2a measures but administered at a later time than directly after training.
    - Level 2c. Behaviour/skill demonstration: measures assessing the behavioural proficiency within the training, rather than the work environment.
- Level 3. Transfer (Kirkpatrick level: Behaviour, Bell & Waag levels: 3.a. transfer to alternative simulation environment, and 3.b. transfer to operational environment). Measures that assess to what extent the knowledge and skills attained during training actually are usable in the real work environment.
- Level 4. Results (Kirkpatrick level: Results, Bell & Waag level: extrapolation to combat environment). Measures that assess the organizational impact of training such as, for example, productivity gains, cost savings, etc. Measurements on the results level are the most distal from the actual training, but by some perceived as the most fundamental when judging training success as they are linked to the underlying reason why the training was performed.

1.5 Tools for Evaluation of Training

Over the past decade, a broad range of tools have been developed for different domains and applications as the need for rapid and structured human performance measurement often becomes apparent (Bennett, Lance, & Woehr, 2006). A number of different measurement approaches exist and the list below summarize the main categories:

- **Subjective ratings**
    - Ratings of own performance
    - Ratings of team members performance
    - Ratings from observers
- **Measured task performance**

  - Primary task performance, i.e., the main tasks being measured by a) effectiveness related measures, e.g., task completion time, or b) efficiency related measures, e.g., workload during task completion
  - Secondary task performance, i.e., other tasks which are more or less ecologically valid, e.g., ability to do some recordable action when a certain code word is heard over the radio
- **Psychophysiological measures**
  - Heart activity, e.g., heart rate, heart rate variability
  - Brain activity, e.g., EEG, NIRS
  - Skin conductance, e.g., Galvanic Skin Response (GSR)
  - Eye activity, e.g., search patterns and durations, blinks
- **Communication measures**
  - Verbal pattern analysis, e.g., analysis of frequency of transmission
  - Verbal content analysis, e.g., analysis of speech acts uttered on aircraft radio
  - Non-verbal pattern analysis, e.g., when and to whom are data link messages sent
  - Non-verbal content analysis, e.g., content of data link messages that are sent
- **Qualitative descriptions of performance**
  - How well team members seem to understand each other in terms of concepts (shared conceptual model)
  - How well team members know each actor's task responsibility (shared mental model)
  - How well team members share an understanding of the situation (situation awareness)
  - How coordinated the actions are in collaborative team work (skill based coordination)

Below, a selection of tools with high applicability for operations in a complex environment are presented. Two tools belong to the category that base their measurement of recording and analysis of technical parameters from a simulator (or live) environment, i.e., PM-Engine and PETS. However, PM Engine is integrated in the product suite named A-Measure, where also the SPOTLITE tool is integrated. They are both described in the section on A-Measure. SPOTLITE evaluations are based on subjective ratings from observers, as in the TPM tool. STAT measures are also based on subjective ratings, but here from the trainees themselves or team members.

Common for the tools described is that their value as pedagogical tools not primarily is based on the specific software features of the tool itself, but rather on how the tools are used in practice and the efforts to populate the tool with specific training audience related content. Technically tools can be similar, but it is the embedded content and the methodology behind its use that decides whether the tool really supports learning. For example, for the PETS tool considerable effort from subject matter experts and researchers have been invested in definition and parameterization of the air combat metrics which gives PETS the capability to detect violations and departures from recommended behaviour. Crucial for success is also that the measures are related to clear training objectives.

1.5.1 A-measure

A-measure is a product suite developed by Aptima Ltd. The text below describes the components of this comprehensive human performance measurement tool.

The following description is directly adopted from Aptima, Inc.'s web-page at http://www.aptima.com/a-measure/A-MeasureBrochure.pdf

Create, collect, and visualize performance measures in live, virtual, and constructive environments.

In measuring performance, there is no single best source of data. Valuable information can come from the training simulator, observers watching the training session, and from the trainees themselves. Aptima has used its years of training experience to create the A-Measure product family, the industry's first full-performance measurement solution. A-Measure lets you combine all aspects of performance measurement in one solution, providing the complete picture that instructors need to improve learning.

Modular software for complete flexibility

A-Measure's components can be used individually or as an integrated suite for realtime performance measurement.

- ***Create***
  - *Performance Workbench™ – Author measures via graphical interface*
- ***Collect***
  - *PM Engine™ – Collect system-based and physiological measures*
  - *SPOTLITE™ – Collect observer-based measures*
- ***Visualize***
  - *Real-time Performance Dashboard™ – Understand results in real time*
  - *Feedback Toolkit™ – Understand results for debrief*
  - *Trending Tool™ – Understand results over time*

Create

Performance Workbench™

Author measures via graphical interface

Performance Workbench equips instructors with the tools to rapidly specify and implement performance measures. Instructors can draw new measures using a graphical diagramming palette, or define measures using a step-by-step wizard. By automatically converting system based measures into the Human Performance Mark-up Language (HPML), Performance Workbench can capture data from a variety of training simulators. Observer-based measures can also be exported to A-Measure's SPOTLITE™ for assessment by instructor-observers.

Collect

PM Engine™

Collect system-based measures

A-Measure's performance measurement engine – the PM Engine – calculates and assesses measures from data generated during simulation-based exercises, to provide real-time feedback. The PM Engine uses HPML to handle data from a variety of sources, such as a distributed simulation network (e.g., HLA or DIS), log files, and SPOTLITE.

Collect physiological measures

As data from sensors monitoring an operator's body signals, such as brain, heart, muscle-activity, eye movement, respiration, or galvanic skin response, are fed into A-Measure, they are computed into meaningful measures in real time by the PM Engine. When combined with system- and observer based measures, this results in a comprehensive assessment of an operator's state and mission performance.

SPOTLITE™

Collect observer-based measures

SPOTLITE – the Scenario-based Performance Observation Tool for Learning in Team Environments – allows observers to assess in real-time the performance of teams in simulated and live training environments. SPOTLITE enables instructor/observers to collect, analyse, and visualize performance data for debrief immediately following an exercise.

SPOTLITE Mobile App

In addition to applications for use on a tablet PC such as the Panasonic U1, SPOTLITE is also available as a mobile app that runs on the Android-tablet-based versions have been optimized to perform with maximum visual acuity in out-of-doors daylight conditions to enhance SPOTLITE's usability in the field.

Context-relevant measure cueing

A-Measure's measurement cueing capability allows a user to define criteria, such as ratings in the PM Engine, to cue the SPOTLITE user to collect additional data, directing those measure cues to occur in real-time as the training exercise unfolds. For example, if the PM Engine receives a certain measure result, it might cue the SPOTLITE observer-instructor to collect a set of pre-defined measures by automatically adding them to the SPOTLITE measurement tree and notifying the user of their addition.

Visualize

Real-time Performance Dashboard™

Understand results in real time

As performance is tracked and calculated by the PM Engine, these measures can be displayed through a real-time Performance Dashboard. The Performance Dashboard is reconfigurable to show instructors the real-time team performance measures as the exercise unfolds, as well as provide dynamic feedback for distributed trainees.

Feedback Toolkit™

Understand results for debrief

The Feedback Toolkit supports the development of after action review tools that provide feedback for any simulation, from individual to large distributed simulations. The Feedback Toolkit offers features to enhance learning and make feedback meaningful and relevant, including tactical maps and overlays, timelines, audio communications and transcripts, communications pattern analysis, and other performance measurement visualizations.

Trending Tool™

Understand results over time

The Trending Tool is a web-based application that consolidates the individual performance measures and assessments over time to provide insight into performance changes. Initially implemented for data collected by instructor observers on their SPOTLITE handheld devices, the Trending Tool provides comprehensive, automated trending and effectiveness reports and analyses.

1.5.2 PETS: Performance Evaluation & Tracking System

The Performance Evaluation and Tracking System (PETS) is a performance measurement tool that has been developed by the US Air Force Research Lab (AFRL), Warfighter Readiness Division since 2001, supported by contractors such as Lumir Research Institute (www.lumirrresearch.com). See for example Portrey, Keck, & Schreiber (2006) who described some of the challenges faced during the development of PETS. The original PETS was designed as a proof-of-concept human performance measurement system that could collect some 80-100 'core' air-to-air and air-to-ground combat performance measures in real-time from a distributed network using the DIS protocol. Apart from the resources spent on the development of the PETS software, AFRL has also made a considerable effort concerning parameterization of the measures for different platforms and weapon systems, e.g., what is the maximum acceptable time spent in an enemy's weapon engagement zone. Continued development has led to the Coalition PETS (C-PETS) version used today which is an authoring shell for logging and analysis of air combat related variables. During 2012 PETS has been integrated into the training range at Nellis Air Force Base where the large Red Flag exercises are conducted. PETS is there used in a Live-Virtual-Constructive (LVC) application. PETS has also been integrated with the debriefing tool LNCS that can be used to visualise air combat sorties during debriefings. During debriefings at AFRL approximately ten measures from PETS are presented to participating pilots.

1.5.3 Training Progression Manager in Exonaut

The company 4C Strategies develops a training management tool named Exonaut (http://www.4cstrategies.com/SiteAssets/Exonaut_Software_Suite.pdf) that supports planning, execution and assessment of exercises. One of the modules in the Exonaut software suite is the TPM module (Training Progression Management).

The following description is directly from 4C Strategies web-page http://www.4cstrategies.com/SiteAssets/Exonaut_TPM.pdf

The TPM module is developed in order to provide a capability to follow training progression and enable assessments whether a military unit's performance meets the required standards. TPM keeps track of training progression from the first observation to the final assessment, to ensure early capture of training risks. Progression is monitored against a set conditions and standards (objectives that are defined by subject matter experts) and are continuously updated as the training cycle progresses. TPM will automatically aggregate observations and propose an assessment of unit status. Assessments are displayed in results matrices, giving commanders real-time access to training results. TPM will enhance the exercise manager's control over training, during and between training events. TPM allows for better and more informed decisions, taking training risks into account. The tool can also generate a number of reports that provide for analysis of training trends during a training event, as well as analysis of long-term progression.

TPM includes features such as:

- *Observations with media capture from mobile devices*
- *Assessment of performance*
- *Analysis of training trends*
- *Automatic feedback on results to exercise planners*

1.5.4 STAT: Squadron Training Assessment Tool

The US-based company GoE (The Group for Organizational Effectiveness) has on contracts from AFRL developed STAT (Squadron Training Assessment Tool). STAT enables longitudinal assessment of training development, for example for the pilots of a fighter squadron, through collection of subjective ratings of performance and scenario realism during training. The content of STAT, i.e., the questions that pilots answer after a sortie, are based on a MEC competency model (see Section 2.1). STAT has been fielded and tested at two operational fighter squadrons of the Swedish Air Force during 2013 as a part of the PROFET R&D project in order to gather experiences and feedback from operational use.

2 Training at the Swedish Air Force Simulation Centre (FLSC)

The Swedish Air Force runs a program for simulation-based training, research, and development at the Swedish Air Force Combat Simulation Centre (FLSC). FLSC primarily provides training for fighter pilots and the facility is designed to provide experiences that develop the trainees' knowledge and skills in decision making, planning, communication, tactical behaviour, and situational awareness. The emphasis is on developing skills and knowledge on a team and inter-team level using eight fighter cockpits and four fighter controller stations (ground control intercept, GCI). The research activities are mainly focused on training effectiveness and human performance, while the development program conducts simulation-based acquisition studies as well as tactics development and concept of operations (ConOps).

As any training program should be based on a model of skill acquisition, FLSC has adopted a model influenced by Kolb's Experiential Learning Theory (Kolb, 1983), in the spirit that learning is the process whereby knowledge and skills are developed through the active transformation of experiences. In short, the practical implementation of this model is that the trainees train themselves using the resources provided in the facility under observation by and guidance from peer pilots and instructors. The intent is to provide and maintain an environment which stimulates the motivation of deliberate practice and active learning among the trainees, improving their performance towards expertise levels (Ericsson & Charness, 1994; Ericsson, Krampe, & Tesch-Römer, 1993; Kozlowski, 1998). The trainees are exposed to concrete experiences in the simulator, offering observations and reflections during planning, execution, and debriefings. The trainees own influence over the training aims to stimulate them to assimilate their reflective observations into concepts which they can actively discuss, test, and explore.

In addition to delivering training according to a well-designed skill acquisition model, the content of the training must meet a set of operational requirements to be meaningful and successful across environments, moving the trainees closer to operational expert performance (e.g., Ward, Williams, & Hancock, 2006). The training scenarios at FLSC has been carefully designed and evaluated with the support of senior pilots possessing extensive experience from operational deployments, before implemented in the training program.

2.1 Training Analysis

To further develop the training, FLSC has utilized the Mission Essential Competencies (MEC) to identify the competencies required for SwAF JAS39 Gripen multi-role missions. A recent description of this comprehensive competency modelling is provided in Alliger, Beard, Bennett, & Colegrove (2012). Borgvall & Castor (2006) offer a descriptive review of the methodology (in Swedish). With the support of the US Air Force Research Lab (AFRL), knowledge and skills essential for JAS39 Gripen multi-role and their developmental experiences across environments have been identified and evaluated with regard to current training. The MEC method has been developed by the US Air Combat Command (ACC), US Air Force Research Lab (AFRL), the Group for Organizational Effectiveness (GoE), and Aptima and has been used to map the competency requirement of a wide variety of complex human jobs since its inception (Colegrove & Alliger, 2003). Through the application of the MEC method a comprehensive analysis of required competencies is performed which in the end results in a training gap analysis.

2.2 Design & Delivery

The training concept at FLSC is based on Kolb's Experiential Learning Theory (ELT) and his four-stage cycle of learning: concrete experience, reflective observation, abstract conceptualization, and active experimentation (Kolb, 1983). At FLSC, this basically means that the trainees (i.e., pilots in this case) have a high degree of responsibility and influence on their own training. For example, they plan their training together with the instructors and run the debriefings themselves (even though the FLSC instructors may support with additional comments). By doing so, the pilots, with the support of the instructors, are provided with concrete experiences in the simulator, leading to observations and reflections during execution (pilots flying and pilots not flying) and debriefs. The pilots influence over the training aims to stimulate them to assimilate their reflective observations into abstract concepts which they can actively test and elaborate with. The experience is that this model supports motivation and deliberate practice among the pilots, which in several domains has been found to be a crucial ingredient for developing expertise (Ericsson & Charness, 1994).

2.3 Training Feedback & Evaluation

The experience at FLSC is that effective debriefs can substantially enhance the effect of simulator-based team training. Current practice is that instructor pilots from the squadrons perform debriefs supported by instructors from FLSC. A very important tool for these debriefings is the HawkEye tool. Borgvall, Castor, & O'Connell (2011) describes this visualization tool that presents a 3D "God's Eye" view of the simulated world, together with the head-down displays from all eight cockpits in FLSC. The graphical user interface gives the user full control of the view of the world, with full zooming and control of the view angle. The temporal control of the replay provides up to 16 times fast forward or backward playback speed. The functionality to go directly to a specific timestamp also exists. However, the 3D-view not only shows all the aircraft in a selected area, but also in real-time visualizes data and status parameters for each aircraft, for example relating to weapons and sensor systems. Emphasis has been put on the development of visualization of the non-physical parameters (e.g., verbal and data link communication), and information that wouldn't be visible to pilots in reality (e.g., the missile envelope or a missile's predicted point of impact). During replay, the visual presentation of every parameter or function can be turned on and off as different users have different needs and preferences. The tool also provides the possibility to enter time-stamped observer comments that are shown during replay. It is also possible to web-cast the replay which is a useful feature during distributed simulation exercises.

Hawkeye contains functionality and software plug-ins that visualizes the parameters below (although not all plug-ins are used at all times):

- Basic parameters such as speed, altitude, fuel and weapons remaining.
- Flight paths and afterburner status.
- Visualization of radio communication, including statistics (total air time, percent of air time, number of step-ons).
- Visualization of tactical data link messages (Link 16) between aircraft and with ground controllers.
- Radar coverage, radar lock-on and target sorting symbols.
- True location of each entity and/or the location of entities as perceived by aircraft radar.
- Missile envelope, missile lock-on status and the missile's predicted point of impact.
- Shot logs for missiles, including hits/misses and reasons for missile termination/abort.

Separate functionality for logging and analysis of a rather large number of other parameters has been developed and integrated with FLSC core simulator functionality over the years, in response to the needs of specific simulation-based acquisition (SBA) projects. For example a large SBA project concerning future sensors resulted in a major development effort concerning logging and analysis. After porting to a new operating system and refacturing of core simulator functionality a new development effort was initiated during 2011. In Waldemark, Castor, Borgvall, Andersson, Peterson, Anderson, Waller, & Lavén (2011) requirements for future logging were specified on a rather detailed level and many measures were suggested. Grev (2012) added additional suggestions concerning logging and visualization.

3 Recommendations

In this chapter a number of recommendations for future efforts are provided. First a section with over-arching thoughts and ideas, which in the next section are broken down into a number of suggested activity threads. Finally the measures and their development are put in context and a final conclusion regarding continued research and development is provided.

3.1 Innovative Data Analysis, Feedback, & Visualization concepts

Many studies on cognition in practice and everyday life (e.g., Carragher & Schliemann, 2000) have shown that general principles and procedures through which humans learn and then later apply does not seem to exist. Instead, people more strongly rely on resources present in a particular situation. What is to be learned is embedded in activities and can be taught through interacting with teachers, books and other artefacts as well as other learners. By way of the interaction and communication taking place in activities the learner can reflect upon and discover concepts and problems and thereby actively construct and test knowledge. In situated perspectives on learning as described above, there is a strong focus on processes, on activities and use of concepts and other tools (artefacts), and support from experienced practitioners in a given domain. This perspective on learning and training falls within the perspective of socio-cultural perspectives on learning described in Section 1.2. Pedagogical models of learning.

Training in the FLSC simulator environment today strongly builds on measuring and evaluating results in terms of quantitative measures such as kill-ratios, number of launched missiles, etc. There is thus little attention paid to the actual process of flying and fighting during sessions. In the environment of today there is a lack of visualizations that capture the process and features of it. New visualizations would enable focusing aspects of strategy, individual as well as group behaviour that otherwise would become hidden in the more summative evaluation measures. There is today also a lack of tools that aid in bridging the temporal gap between flying in the simulator and carrying out after-action reviews, i.e. there is a lack of tools that in different ways facilitate direct feedback. There is also a lack of tools facilitating observing and following results from training over time, individually and as a group.

Many of the current state-of-the-art simulation environments for learning and training fit very well to crew resource and/or crisis management, both for individualized and collaborative training (Cherry & Ali, 2008; Hicks, Bandiera, & Denny, 2008; Kobayashi, Patterson, Overly, Shapiro, Williams, & Jay, 2008). However, in the FLSC environment it is possible take on research and takes the training one step further. Through the use of the possibilities that are offered by simulation based training by for instance allowing pilots to be subjected to and thereby train also in atypical situations, that is training on something that is perfectly impossible to do in any other environment, this could even further stimulate reflection by the pilots.

Further, based on for instance the theoretical perspective \`Conversational Framework´ (Laurillard, 2009) we wish to develop domain specific support for individual and collaborative reflection regarding what actually takes place during training sessions, pointing at critical instances, alternative possible actions, and un-reflected habitual behaviour/actions. In other words, to a higher degree focus on what occurs during a session in the simulator rather than merely the summative result thereof. After-action reviews could thus be improved and give a more immediate feedback about individual behaviour as well as collaborative team behaviour in the scenarios. This would in turn increase possibilities to evaluate actions taken (individually and collectively).

Design and use of new tools for visualization and tools that bridge the temporal gap between flying in the simulator and having after-action reviews, would also contribute and tap in to experiential learning resulting in better acquisition, retention and transfer of knowledge, increased confidence in decision making, problem-solving and/or trust building in critical cases (Yu, 2001).

3.1.1 Development of new tools & visualisations

Taking a new approach on simulation and training in the FLSC environment would require (at least) development and mapping of scenarios tailoring specific skills and competencies, as well as development of atypical situations that open up for training skills and competencies that would be impossible to train in other environments. Development of new tools for visualization are thus called for, tools that bridge the temporal gap between flying in the simulator and having after-action reviews, as well as between training sessions so as to enable tracking development of skills and competencies. The visualizations we envision would on one hand actively support flying in the simulator and on the other hand also support individual and collaborative debriefing (for instance support reflection regarding actions taken and the appropriateness of these as well as what could have been done differently and why).

Developing and introducing new tools for visualization would also require development of the pedagogical model on which the training and debriefing sessions are based. Instead of iterating ”flying session-debriefing-flying session”, other general models should be investigated. For instance models based on the “The conversational framework´ (Laurillard, 2009) that dynamically facilitate both individual and group debriefing. A plausible general model for this could be 1) synchronous and individual feedback in flight scenarios by use of new visualizations 2) collaborative team after-action reviews in which aspects of strategy can be discussed and elaborated, use of tools that enable visualization of individual and group actions, as well as individual actions as related to team actions, and 3) individual debriefing, use of tools that enable visualization of individual actions in a flying session, and development of skills and competence between flying sessions and over time. An example of a visualization that could be used both in 2 and 3 is to plot individual actions, i.e., individual aircraft and actions taken by a pilot, as compared to an ideal plot (conceptual model of ideal flight pattern and actions taken). Comparisons of the individual flight pattern and actions taken to the ideal plot could then be made and count as anchors for discussions and reflection, individually and collaboratively. Deviations from the ideal plot might be more or less called for or more or less consciously and deliberately made. Important to the pedagogical model assumed, and the training and after-action reviews conducted, is that the goals are clearly defined and measurable, that these are clearly communicated to the pilots, and followed up individually and collaboratively in the after-action reviews.

The feedback loop is essential for learning at both individual and group level. The quicker the learners are able to get an overview of the assignment the better. An idea could therefore be to collect and present the individual and the collective feedbacks immediately after a session. This could for instance be done by having each pilot answer questions about the session and their view of the session. This data could then be presented in different forms of visualizations comparing the individual differences, but also enabling to compare different collective assessments given different sessions. This could be done with using tablets where the pilots immediately answer a set of predefined questions, equivalent to what is today answered on paper. During after action reviews it could also be possible to stop the session and rearrange to situation (in terms of flight routes, shot times, sensor utilization, etc.) and discuss the differences in tactics and how one would act in such new fictional context and thereby support reflection on actions that have been taken and exploration of alternative actions.

The use of an eye-tracking system in the cockpit would make visible the visual focus of the pilot during flight sessions. This information could be used to feedback and visualize the visual focus of pilots in both group and individual after-action review, serving complementary goals. As in the case above, visual scan patterns at the instrumentation, alternating between different instruments and the scene outside of the cockpit, may be more or less called for or consciously and deliberately made.

Another path is to data-mine and visualize historical data from 10+ years of previously conducted exercises at FLSC. By processing, filtering and aggregating this data, typical cases of pilot and team behaviour in specific situations can be generated. Visualizing such typical cases, and in particular illustrating the deviation from the expected data, could help instructors and training designers in understanding, analysing and refining the scenarios. In addition, human factors researchers could benefit from such visualizations for analysis of pilot and team performance. Providing such insight into previous experiences and outcomes, could prove beneficial in optimizing the planning and pre-training stages. Conducting such analyses on this vast amount of data could plausibly also reveal patterns that would otherwise be difficult, if not impossible, to observe. Similarly, pilots may benefit from comparative visualizations of their individual performance and such aforementioned typical cases in specific situations. Through direct visual presentation of spatial and temporal deviation from the expected outcome, pilots may be further aided in reflecting on their own performance and decisions throughout the executed sortie.

There are already plans to incorporate psychophysical sensors, such as eye trackers, in the pilot stations. Various measures, including fatigue, workload, and attention, can be extracted from such sensors, and be analysed online or offline in a human factors research context as well as related to the `Conversational framework” to facilitate reflection regarding performance and actions taken. Additional complex composed measures, such as a pilot’s current intent, aggressiveness or awareness, could conceivably be extracted from a combination of the aircraft state (e.g. “determined” or “uncertain” motion pattern), pilot gaze (what the pilot is paying attention to, e.g. surrounding objects or specific data in head-down display devices), and control input (e.g. distinct and clear adjustments vs. constant corrections). Making such often routinely conducted individual actions visible opens up for inspection and reflection on actions and plausibly also improvements that can be made. With such measures reliably and intuitively visualized, scenarios could further be altered in real time to increase difficulty for under-stimulated pilots or ease the workload for stressed ones, as appropriate. Such measures could also aid to improve visualizations, instruments and other human-machine interface in the cockpit.

3.2 Structured & Theory-Driven Measurement Development Process

The above described vision of a possible future for the FLSC would in implementation build on different and complementary theoretical perspectives on training and learning (as described in Section 1.2.) building on the strengths of both individual and social collaborative learning and training. In striving towards the vision one would need to build on an appropriate and legitimate model and process. A suggestion is to build on the ADDIE-model as described in Section 1.3.1 to structure the training and development of the training. Further, we would also need to be more specific regarding what individual and collaborative team measures to make. Further, the choice of tools and the measures collected by the tools should be theoretically justified and structured. Measures made would be on different levels of abstraction, or on different levels of description depending on what type of questions that needs to be answered. See for example the different training criteria described in Section 1.4.

Measuring teamwork

In a team training environment such as the FLSC, effort must be made to measure teamwork, despite of the magnitude of processes involved. One example could be to build on the “big five” framework proposed by Salas, Sims, & Burke (2005). Conceptual mapping between collected measures by any measurement tool and the teamwork models described below are the foundation of a teamwork measurement effort. The components of the “big five” of teamwork model are:

Team leadership: the ability to direct and coordinate the activities of other team members, assess team performance, assign tasks, develop team knowledge, skills, and abilities, motivate team members, plan and organize, and establish a positive atmosphere.

Mutual performance monitoring: the ability to develop common understandings of the team environment and apply appropriate task strategies to accurately monitor teammate performance.

Backup behaviour: the ability to anticipate other team members’ needs through accurate knowledge about their responsibilities. This includes the ability to shift workload among members to achieve balance during high periods of workload or pressure.

Adaptability: the ability to adjust strategies based on information gathered from the environment through the use of backup behaviour and reallocation of intra-team resources. Altering a course of action or team repertoire in response to changing conditions (internal or external).

Team orientation: the propensity to take other’s behaviour into account during group interaction and the belief in the importance of team goals over individual members’ goals.

Cannon-Bowers, Salas, & Converse (1993) state that effective coordination at the team level depends on the emergence of a shared mental model, or common understanding among team members regarding expected collective behaviour patterns. Well-developed mental models help individuals to process and classify information more efficiently and form more accurate expectations and predictions of task and system events. Cannon-Bowers et al. suggest three kinds of knowledge structures in teams: the first is a mental model that contains knowledge about the purpose of the teams, and more specifically the task requirements related to this purpose, called a task model. This model includes task procedures, strategies, and information on how the task changes in response to changes in the environment. The second model represents knowledge about unit characteristics, including their task knowledge, abilities, skills, preferences, and tendencies, called a team model. And finally the third model, and the one that perhaps is most significant when studying collective action, contain information regarding the individual and collective requirements for successful interaction with team members. Cannon-Bowers et al. describe that to be effective, team members must understand their role in the task, which their particular contribution is, how they must interact with the other team-members, who require a particular type of information, and so forth. Related to this, they must also know when to monitor their team members’ behaviour, and when to step in and help a fellow team member who is overloaded, and when to change behaviour in response to the needs of the team. When shared among team members, this model, called the team interaction model, is particularly crucial to effective coordinated action. Thus, both individual and collaborative aspects of training are emphasized and conceptualized in the model.

Measuring on different task analysis levels

Another conceptualization relevant in detailing the vision and more specifically measuring different tasks and levels has been proposed by Myers, Tijerina, & Geddie (1987), which distinguishes between mission, scenario, function, job, duty, task, sub task and task element. Measures mapped on these different levels may have different uses in the pedagogical process. For example, measures of success/effectiveness on the mission level are useful to summarize the training results to higher level of command, but probably too coarse to be useful during a training week at FLSC. Measures on the subtask level on the other hand are probably too detailed to be used for example during an after-action review, but provide valuable data to analyse afterwards by a statistician and can thus contribute with insights regarding the training effects.

A major challenge when evaluating data is the fact that there are many skills where the measurable actions from the execution of the skill are hidden and embedded in other behaviour. Also the desired outcome may be achieved by very simple actions or by not performing an action. An operator's handling of a system may have a wide range of effects from immediate to gradual or remote and from trivial to critical. The above described frameworks would aid in detailing the ADDIE model that would count as the structuring element in training development and evaluation.

A pedagogical model for presentation of measures

An issue for research at FLSC, when performance measurement tools are further developed and integrated, is to investigate which measures, along with how they are presented, that are useful in the pedagogical process and provide the "best" results. The capability to quickly present a number of logged and/or automatically analysed and aggregated measures is very desirable. This capability reduces the temporal distance between the actual situation where a tactical manoeuvre is executed and the time when learning/reflection occurs. However, this technical capability needs to be integrated in a carefully designed pedagogical process in order for the learning to really occur. Further scientific elaboration and experimentation regarding how a "pedagogical point/insight" in a team context is designed and managed is needed. The understanding is probably based on exposure to contextualised problems and situations triggering decisions and active reflection, i.e., going through all steps in Kolb's four-stage cycle of learning. A number of research questions concerning how logged and visualized measures support this process with understanding and acceptance of each "learning point" can be identified. A model/method/process for enhanced de-briefings would be of major scientific and operational relevance. It is recommended that this research activity is initiated by R&D at FOI/FLSC during 2014.

Some related issues to consider are:

- How many different measures and which measures should be presented during a "normal" FLSC de-briefing session? How should these measures be presented?
- Measures describing process or outcome can be identified. The question is which type of measures that provides the most useful basis for identification of how the performance can be improved or changed?
- How should the performance data of the training audience of a specific week be related to data collected earlier?
    - Which comparisons between data sets should be common practice or be presented automatically? Some alternatives are:
        - Comparison with earlier performance from the same group, e.g. performance earlier during the training week or from a previous week. Comparisons with all previous data, across the whole population
        - Comparisons between organizational units, e.g., fighter squadrons.
        - Comparisons with behaviour generated by Computer Generated Forces (CGFs). The recently developed CGFs at FLSC are very competent in some regards and quite incompetent in others, but they exhibit stable and repeatable performance which can be useful as a reference.

- Should deviations from the population mean for any measure, although that specific measure is not normally presented, “pop-up” during the de-briefing?
- Should measures that show a positive or negative trend be displayed and how clear should the trend be?

Harmonization of training goals between organizational levels

The success of training is inherently linked to the identification and definition of training goals. For the tools discussed in the current report it is important that they relate to the training goals of different organizational levels in an organization. With the Swedish Air Force as an example the training objectives of the Air Component Command (FTS), the squadron commanders, the flight duty planners/responsible officers at the squadrons, and the individual pilots often are aggregations or subsets of each other. The performance measurement must be to support evaluation of training goal fulfilment from these different levels. The situation is quite often that the same training week should provide information to a) FTS who wants to get a recent estimate of training needs and status of the squadron, b) the squadron commander whose training goal is for the pilots to be able to perform a straightforward, effective and repeatable tactic and fulfil official training requirements, c) the training planner who during the week in real-time wants to pace training to the squadrons progress and d) the individual pilot who wants to improve his own and his team’s performance. One important research question is how the translation between these training goals is made and how this is explained to the training audience.

On a practical level the training planners from the squadrons must be made aware which logging features that are available from FLSC and how they can be tailored to support the current training evaluations. In some aspects Concept of Operations (ConOps) and Standard Operating Procedures (SOPs) develop rather rapidly as new experiences are made.

Desired properties of tools and measures

For further development of performance measurement capabilities a number of desired properties of tools and measures have been identified:

- Measures must be conceptually straightforward and “easy to understand”. The training audience must understand why the measure is presented and how it is computed. Whenever possible measures should be visualized.
- It must be easy to relate measures to a timeline, as time is an important organizing principle for humans to understand a situation. However, other ways to rank and order data can be imagined, e.g., present events in order of importance (rated or derived according to some algorithm).
- The measures should represent both the process and the outcome. Especially for process related measures it is desirable to be able to present the measures in real-time as a mission progress. Under some conditions it might even be beneficial to present them in the cockpit, even though this information wouldn’t be available during real flight.
- Instructors at FLSC should be able to present measures directly after a mission has ended. Measures that are too complex, too computationally demanding or requiring hands-on attention/management from an engineer or analyst are less probable to support training.
- It must be easy to relate different measures to each other and to temporal, scenario or platform specific aspects as it often is a combination or triangulation of information that is needed. For example, a composed measure that captures the offensive stance of a behaviour or tactic in relation to the enemy's offensive stance over time can be an interesting measure. This measure would need to be a

combination of speed towards the enemy, number of missiles per time-unit from certain distances and aspect angles and so on in relation to the same variables from the enemy. It might also be of interest to relate psychophysiological measures to other variables, i.e., the heart rate is increased during certain phases or situations. Any logging should thus be associated with extensive logging of meta-data.

- Collected measures should automatically be compared to data sets that have been collected earlier.
- The performance measurement must be integrated/harmonized with the vision and doctrine of the organization.
- The measures must be based on operational definitions and should harmonize with the mental models of the target audience.
- The measures must reflect factors that the training audience can influence.
- The amount of information presented must be limited/manageable during an after action review, but still provide a holistic view of performance.
- All measures must be reported in relation to other measures.
- The statistical analysis behind the measures must be able to cope with data of different types and be able to integrate them statistically, e.g., through multivariate statistics such as structural equation modelling, c.f., Castor (2009).

Several different user groups for the tools described in the report exists, and their requirements differ somewhat, as described in Borgvall, Castor & O'Connell (2011):

- Training audience - desire "immediate" feedback in a manageable format that describe their performance and process.
- Instructors - need to be able to design exercises with appropriate challenge levels during preparation of training weeks. During the training week they need to pace the training based on the training audience improvement.
- Training researchers - desired access to many variables with different characteristics in order to track learning effects.
- Operational analysts - desire access to many variables similar to the training researchers, but have a propensity to be interested in measures related to the technical systems.
- Technical development team - wants technical feedback concerning bugs and the simulator systems optimization/development, e.g., be able to run standardized batch trials overnight when new versions of software have been integrated.

The needs from all these user group should be considered when future tools are developed.

Proposed activities

- Research and develop tools that support communication analysis.
- Integrate and routinely use several different types of measures and tools such as those described in the report.
- The feedback loop is essential for learning at both individual and group level. The faster the training audience are able to get an overview of the mission the better. It would therefore be a good idea to collect and present the individual and the collective feedbacks immediately after the session. This could be done by having each pilot answer questions about the session and their view of the session. The data could then be presented in different forms of visualizations comparing the individual differences but also being able to compare different collective assessment given different sessions. This could be done with using tablets where the pilots immediately answer a set of questions, equivalent to what is answered on paper today.
- Develop new visualizations that support collaborative and individual debriefing, possibly as a new plug-in to Hawkeye.
- Develop atypical situations that can be used during training weeks.
- Research and evaluate new pedagogical models for individual and collaborative training, e.g., the conversational framework (Laurillard, 2009).
- During debriefing sessions it could also be possible to stop the session and re-arrange tactical situations (in terms of flight routes, shot times, sensor utilization, etc) and discuss the differences in tactics and how pilots would act in such new fictional context. This would imply changes to the pedagogical framework but possibly enhance FLSC as a collaborative setting for “teachers” and “learners”.
- During flight sessions it might in some situations be desirable to provide learning management temporal control, e.g. functionality to Start, End, Halt-Resume, Backup, Replay, Speedup, and Skip during the simulation.
- Recurrent reviews of performance indicators/measures is essential to keep them up to date with current doctrine, concepts of operations (ConOps), and standard operation procedures (SOPs).
- Efforts concerning standardization of human performance measurement, e.g., the Human Performance Measurement Language (Stacy, Ayers, Freeman, & Haimson, 2006) or the Metric Ontology Tool (Pester-DeWan & Oonk, 2006) should be surveyed and monitored over time.
- Conduct statistical analysis of collected and aggregated data (data mining) in order to find interesting patterns, problems and possibilities in the pilots’ behaviour.
- Equip the pilot stations with psycho-physiological measurement systems such as eye-trackers and other sensors. Summaries of available psychophysiological measurement methods are for example provided in the GARTEUR Handbook of Mental Workload Measurement (Castor, Hanson, Svensson, Nählinder, LeBlaye, MacLeod et al., 2003) or the NATO HFM report on Operator Functional State Assessment (Wilson, Balkin, Beamont, Burov, Edgar, Fraser et al., 2004).
- To a larger extent include personnel and systems that represent other service branches apart from the air force, i.e., personnel from the navy and the army.
- Develop new visualization concepts and tools, e.g., visualizations of risk that can be presented in real-time to pilots, c.f., Grev (2012).

- Develop a concept for an experimental training week, with enhanced forms of debriefing and visualization tools. For instance including use of a-typical combat scenarios, new decision support systems and development of new visualizations.

3.3 Strategy for Training and Integration of Supporting Tools

With the purpose of minimized costs and maximized training effectiveness, and ultimately a learning organization and a mission ready force, the authors of the report suggest that the Swedish Armed Forces initiates a multithreaded research and development project concerning development and integration of pedagogical tools such as those discussed in the report. The project should secure access, at least to extended trial integration rights to the tools described in the report and conduct detailed analysis of specific measures that can be accessed through the tools. When technical integration has been performed, trials of the pedagogical utilization should follow along with experiments concerning new pedagogical models. It must also be noted that performance assessment tools are part of a larger complex of training enhancing systems or requirements:

- Competency models, including training needs analysis.
- Well defined training objectives, on several harmonized levels.
- Well-designed injects, which maps to the training objectives.
- LMS, learning management systems.
- ITS, intelligent tutoring systems.
- A defined baseline of performance, e.g., achieved through utilization of benchmark scenarios during training weeks (possibly against FLSC's recently developed computer generated forces).
- High feedback quality and a shortened temporal distance between execution and learning, achieved for example through improved visualization tools.
- An overarching training strategy from the Armed Forces Headquarters and extrapolation from the air force's current effort on explicit linkage between course curriculum elements between educational steps in a pilot's education or "pilot life-cycle of learning (c.f. "röda tråden arbetet").

In close collaboration with the operational squadrons and FTS, this project should define measures, parameterization and hierarchies of training goals, based on information derived from a multitude of sources, such as a recently performed MEC analysis, squadron level training objectives lists (e.g., "målkatalogen"), operational unit requirements specifications (e.g., "krigsförbandspecifikationer, KFS"), typical training setups and best practices guides from the squadrons, and experiences from exercises and operations.

Inspiration and comparisons concerning future tools is available through continued collaboration with AFRL under the IMTR II Project Agreement. AFRL´s vision concerning an "Office Suite of Air Combat Training Assessment" is a benchmark and should be adopted for Swedish future development efforts.

4 References

Alliger, G.M., Beard, R., Bennett, W. Jr., & Colegrove, C. (2012). Understanding Mission Essential Competencies as a Job Analysis Method. In M.A. Wilson, W.Bennett Jr, S. Gwaltney Gibson, & G.M. Alliger (Eds). *The Handbook of Work Analysis: Methods, Systems, Applications and Science of Work Measurement in Organization*s. Taylor-Francis.

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 50,* 341-358.

Aquinis, H., & Kraiger, K. (2009). Benefits of Training and Development for Individuals, Teams, Organizations and Society. *The Annual Review of Psycholog*y, 60.

Bell, H. H., & Waag, W. I. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *International Journal of Aviation Psychology, 8*(3), 223-242.

Bennett, W., Lance, C.E., & Woehr, D.J. (Eds.) (2006). *Performance Measurement – Current perspectives and future challenges*. Mahwah, NJ: Lawrence Erlbaum.

Borgvall, J. (2013). Generating recommendations for simulator design through pilot assessment of prototype utility. *Military Psychology*, *25*(3), 244-251.

Borgvall, J., & Castor, M. (2006). Mission Essential Competencies: Kompetens - Operativa Krav. FOI-R--2106--SE. FOI Command and Control Systems, Linköping. (in Swedish).

Borgvall, J., Castor, M., & O'Connell (2011). Enhanced scenario visualization for simulation-based training. *Proceedings of International Symposium on Aviation Psychology 2011*. Wright State University: Dayton, Ohio.

Branson, R. K., Rayner, G. T., Cox, J. L., Furman, J. P., King, F. J., & Hannum, W. H. (1975). Interservice procedures for instructional systems development. (5 vols.) (TRADOC Pam 350-30 NAVEDTRA 106A). Ft. Monroe, VA: U.S. Army Training and Doctrine Command.

Cannon-Bowers, J.A., Salas, E., & Converse, S.A. (1993). Shared mental models in expert team decision making. In N.J. Castellan, Jr (Ed.), *Current Issues in individual and group decision making.* Hillsdale, NJ: Erlbaum.

Cannon-Bowers, J. A., & Bell, H. H. (1997). Training decision makers for complex environments: Implications of the naturalistic decision making perspective. In C. E. Zsambok & G. Klein (Eds.), *Naturalistic decision making* (pp. 99-110). New York, NY: Lawrence Erlbaum Associates.

Carraher, D. W., & Schliemann, A. D. (2000). Lessons From Everyday Reasoning in Mathematics Education: Realism Versus Meaningfulness. In D. Jonassen and S. Land (Eds.), *Theoretical foundations of learning environments* (pp.173-195). Mahwah, N.J.: L. Erlbaum Associates.

Castor, M. (2009). *The use of structural equation modeling to describe the effect of operator functional state on air-to-air engagements outcomes*. Linköping Studies in Science and Technology Dissertation No. 1251. Linköping, Sweden: Linköping University.

Castor, M., Hanson, E., Svensson, E., Nählinder, S., LeBlaye, P., MacLeod, I., Wright, N., et al. (2003). *GARTEUR Handbook of Mental Workload Measurement*. Group of Aeronautical Research and Technology in Europe.

Castor, M., Borgvall, J., & Bennett, W., Jr. (2009). Knowledge and skill-based evaluation of simulated and live training – From evaluation framework to field application. *Proceedings of the International Symposium for Aviation Psychology 2009*. Dayton, OH: Wright State University.

Colegrove, C. M., & Alliger, G. M. (2003). Mission essential competencies: Defining combat readiness in a novel way. In NATO Research and Technology Organization (RTO) System Analysis and Studies Panel's (SAS) Symposium, *Air Mission Training Through Distributed Simulation (MTDS) – Achieving and Maintaining Readiness (SAS-038).* Brussels, Belgium: NATO RTO.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Cherry, R.A., & Ali, J., (2008). Current concepts in simulation-based trauma education. Journal of Trauma-Injury Infection & Critical Care. 2008 Nov; 65(5):1186-93.

CSE (2001). *Joint Service Publication (JSP) 502, The Tri-Service Guide to Training Needs Analysis (TNA) for Acquisition Projects*. UK Central Service Establishment (CSE).

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363-406.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*(8), 725-747.

Freeman, J., Salter, W. J., & Hoch, S. (2004). The users and functions of debriefing in distributed, simulation-based team training. *Proceedings of the Interservice/Industry Training, Simulation & Education Conference (I/ITSEC) 2004*. Orlando, FL: National Security Industrial Association.

Goldstein, I.L. (1993). *Training in Organizations: Needs Assessment, Development and Evaluation*. Monterey, CA: Brooks/Cole. 3rd ed.

Grev, M. (2012). Utvärdering av flygning [in Swedish]. Working Document. Division for Information Systems and Aeronautics, Swedish Defence Research Agency (FOI).

Henley, I. M. A (2003). *Aviation Education and Training: Adult Learning Principles and Teaching Strategie*s. Avebury Aviation, Aldershot, UK.

Hicks, C.M., Bandiera, G.W., & Denny, C.J. (2008). Building a simulation-based crisis resource management course for emergency medicine, phase 1: Results from an interdisciplinary needs assessment survey. *Academic Emergency Medicine*. 2008 Nov; 15(11):1136-43. Epub 2008 Jul 14.

Holding, D. H. (1987). Concepts of training. In G. Salvendy (Ed.), *Handbook of Human Factors* (pp. 939-962). New York: John Wiley & Sons, Inc.

Jander, H., Borgvall, J., & Castor, M. (2011). Brain budget – Evaluation of Human Machine Interaction in System Development for High Risk and Task Critical Environments. *FOI- R--3272--SE*. FOI Information and Aerosystems.

Kirkpatrick, D. L. (1959a). Techniques for evaluating training programs. *Journal of ASTD*, 13, 3-9.

Kirkpatrick, D. L. (1959b). Techniques for evaluating training programs: Part 2-Learning. *Journal of ASTD*, 13, 21-26.

Kirkpatrick, D. L. (1960a). Techniques for evaluating training programs: Part 3-Behaviour. *Journal of ASTD*, 14, 13-18.

Kirkpatrick, D. L. (1960b). Techniques for evaluating training programs: Part 4-Results. *Journal of ASTD*, 14, 28-32.

Kobayashi, L., Patterson, M.D., Overly, FL., Shapiro, M.J., Williams, K.A. & Jay, G.D. (2008). Educational and research implications of portable human patient simulation in acute care medicine. *Academic emergency medicine.* 2008 Nov 1;15(11):1166-74.

Kolb, D. A. (1983). *Experiental learning: Experience as the source of learning and development*. Upper Saddle River, NJ: Prentice Hall Inc.

Kozlowski, S. W. J. (1998). Training and developing adaptive teams: Theory, principles, and research. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 115-153). Washington, DC: American Psychological Association.

Laurillard, D. (2009). The pedagogical challenges to collaborative technologies. *International Journal of Computer-Supported Collaborative Learning*, 4(1).

Lave J., & Wenger, E. (1991). Situated Learning: Legitimate peripheral participation, Cambridge, UK, University of Cambridge Press.

LeFrançois, G. R. (1997). Psychology for Teaching. Wadsworth Publishing Company.

Myers, L.B., Tijerina, L., & Geddie, J.C. (1987). *Proposed military standard for task analysis.* U.S. Army Human Engineering Laboratory.

Pester-DeWan, J., & Oonk, H. (2006). Human Performance Benefits of Standard Measures and Metrics for Network-Centric Warfare. *Proceedings of Human Factors Issues in Network-Centric Warfare Symposium*. TTCP Human Resources and Performance Group TP 7 Human Systems Integration.

Portrey, A. M., Keck, L. B., & Schreiber, B. T. (2006). Challenges in developing a performance measurement system for the global virtual environment. *AFRL-HE-AZ-TR-2006-0022*. Mesa, AZ: Air Force Human Research Laboratory, Human Effectiveness Directorate, Warfighter Readiness Research Division.

Preece, J., Rogers, Y., & Sharp, H. (2002). Interaction Design: Beyond Human-Computer Interaction. New York: Wiley.

Ramberg, R., Nählinder, S., Borgvall, J., & Castor, M. (2012). Lägesrapport - Intervjuserie vid FM träningscenter. *FOI Memo 4121*. FOI Information and Aerosystems. (in Swedish).

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a “big five” in teamwork? *Small Group Research*, 36, 555-599.

Salas, E., Bowers, C.A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *International Journal of Aviation Psychology*, 8 (3), 197–208.

Stacy, W., Ayers, J., Freeman, J., & Haimson, C. (2006). Representing human performance with Human Performance Measurement Language (HPML). *Proceedings of the Interservice/Industry Training, Simulation and Education Conference*. Arlington, VA: NDIA.

Stanton, N., Hedge, A., Brookhuis, K., Salas, E., & Hendrick, H. (2005). *Handbook of Human Factors and Ergonomics Methods*. Florida, CRC Press.

Svensson, J., & Andersson, J. (2006). Speech acts, communication problems, and fighter pilot team performance. *Ergonomics*. 2006 Oct 10-22;49 (12-13):1226-37.

Vygotsky, L. S. (1976). *Mind in society*. Cambridge, MA: Harvard University Press.

Ward, P., Williams, A. M., & Hancock, P. A. (2006). Simulation for performance and training. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 243-262). New York, NY: Cambridge University Press.

Waldemark, K., Castor, M., Borgvall, J., Andersson, C., Peterson, S., Anderson, H., Waller, B., & Lavén, P. (2011). Kravspecifikation för ny struktur på dataloggning i FLSC 2011[in Swedish]. Working Document. Division for Information Systems and Aeronautics, Swedish Defence Research Agency (FOI).

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall Inc.

Wilson, M.A.,Bennett Jr, W., Gibson, S., & Alliger, G.M. (Eds). (2012). *The Handbook of Work Analysis: Methods, Systems, Applications and Science of Work Measurement in Organizations*. Taylor-Francis.

Wilson, G.F., Balkin, T., Beamont, M., Burov, A., Edgar, G., Fraser, W., et al. (2004). Operator Functional State Assessment. *RTO-TR-HFM-104*. Paris: NATO Research and Technology Organisation.

Yu, F. (2001). Competition within computer-assisted cooperative learning environments: Cognitive, affective and social outcomes. *Journal of Computing Research*, 24(2) 99 - 117.

FOI
Defence Research Agency
SE-164 90 Stockholm

Phone: +46 8 555 030 00
Fax: +46 8 555 031 00

www.foi.se