2009

Knowledge and Skill-Based Evaluation of Simulated and Live Training – from Evaluation Framework to Field Application

Martin Castor

Jonathan Borgvall

Winston Bennett Jr.

KNOWLEDGE AND SKILL-BASED EVALUATION OF SIMULATED AND LIVE TRAINING – FROM EVALUATION FRAMEWORK TO FIELD APPLICATION

Martin Castor
Jonathan Borgvall
Swedish Air Force Air Combat Simulation Centre
Swedish Defence Research Agency
Stockholm, Sweden

Winston Bennett, Jr.
Warfighter Readiness Research Division
711 Human Performance Wing
Air Force Research Laboratory
Mesa, AZ

In a study, a simulated spin-up exercise and the corresponding large-scale live military flight training exercise was evaluated based on the Alliger et al. augmented taxonomy of Kirkpatrick's training criteria. The data collection was developed and designed to assess the training from reactions to in-simulator knowledge and skill development to operative training effect. The basis for the evaluation was knowledge and skills identified with the Mission Essential Competencies (MEC) process. Using surveys, quantitative and qualitative data from 14 fighter pilots were collected regarding reactions to training, perceived training value and additional training needs. This paper will present the rationale and theoretical framework behind this methodological approach. The main contribution is the description of how the underlying theoretical frameworks have been transformed into measures allowing structured evaluation of training "in the wild".

In July 2008 the Swedish Air Force (SwAF) for the first time participated in the world's largest military live flying exercise – Red Flag (RF). The exercise is managed by the US Air Force at Nellis Air Force Base (Las Vegas, NV) and the flying takes place over the Nevada Test and Training Range (NTTR). Four Red Flag exercises are usually conducted each year and international participation is becoming increasingly common. The exercise is the largest of its kind in the world and by many considered as the most realistic training event available for military aircrew. During this specific RF exercise, RF 08-3, up to 65 aircraft flew two sorties per day for a period of two weeks. The Swedish unit that participated in RF 08-3 was the SwAF rapid reaction unit, which consists of highly motivated and experienced combat ready pilots flying the fourth generation fast-jet JAS39 Gripen. In many ways the SwAF participation in RF 08-3 was considered to be not only an excellent training event but also a war-like criterion of operative performance – a way of confirming developed tactics, techniques, and procedures (TTP) at an individual, team and organizational level.

In order to prepare for the live exercise RF 08-3, the pilots of the SwAF RF 08-3 contingent conducted a series of preparations, in the class-room, live at their squadrons and in a simulator environment. One of the major preparatory events was a simulated spin-up exercise, called RF spin-up, at the Swedish Air Force Air Combat Simulation Centre (FLSC). Several processes of monitoring and documenting the SwAF participation in RF 08-3, ranging from preparations to return to home base, from logistic procedures to fast-jet performance, were undertaken simultaneously. One of them is the work presented here – a training evaluation study monitoring the fast-jet pilots' individual reactions to the spin-up training together with their perceived training value and self-assessed performance readiness before, during and after the exercises.

Training criteria and training evaluation taxonomies

Kirkpatrick's taxonomy (1959a, 1959b, 1960a, 1960b) with four levels of training evaluation criteria has received widespread attention from researchers and practitioners since its original definition. Kirkpatrick's taxonomy originally was a set of practical suggestions, drawn from personal experience, providing a useful heuristic for what can be measured, not necessarily providing strict directions on what should be measured. The taxonomy has been debated by a number of researchers, and flaws in the taxonomy have been highlighted by, for example, Alliger, Tannenbaum, Bennett, Traver and Shotland (1997), who provide an augmented framework.

Bell and Waag (1998) suggests an alternative model for evaluation of simulator training. Even though their model was explicitly focused on simulator training evaluation it was based on Kirkpatrick's training criteria and shows clear similarities with the Alliger et al. augmented taxonomy (1997).

Kraiger, Ford and Salas (1993) describe, as learning is multidimensional, how the result of training should be assessed terms of changes in affective, behaviour (skill-based) and cognitive (knowledge-based) capacities. For the current study, the Alliger et al. augmented taxonomy (1997) was used as the fundament when conceptualizing the surveys. The different levels of the Alliger et al. augmented taxonomy are briefly described below (the corresponding levels from Kirkpatrick's taxonomy and 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: measures that assess to what extent the knowledge and skills attained during training actually are usable in the real work environment. (Kirkpatrick level: Behaviour, Bell & Waag levels: 3.a.) transfer to alternative simulation environment, and 3.b.) transfer to operational environment)
- Level 4. Results: 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 they underlying reason why the training was performed. (Kirkpatrick level: Results, Bell & Waag level: extrapolation to combat environment)

Training Evaluation Method

Participants

Fourteen SwAF JAS39 Gripen pilots participated in the study. They constituted the current SwAF rapid reaction unit at the time, which means they were all experienced combat ready pilots. Individual total military flight hours varied from 950 hr to 2500 hr (M = 1705 hr, SD = 520 hr). Based on their experience, all participants were considered as subject matter experts (SMEs), able to provide highly reliable estimates concerning their own training. This was a pre-condition for the methodological approach of the study.

Design of the study

The study was conducted as a within-participant design. The dependent measures were the reactions, and the ratings of perceived training value and additional training needs. Independent measures were the training during the RF spin-up exercise and RF 08-3 respectively. All data collection was based on surveys.

Description of the two exercises

Red Flag spin-up. The RF spin-up exercise was conducted over four days at the SwAF simulation facility FLSC about one month prior to the live exercise RF 08-3. All simulation scenarios were flown over a satellite photo generated geographical database of NTTR. During a number of workshops, pilots of the unit in dialogue with instructors and training designers at FLSC analyzed training needs for the live exercise and identified essential experiences that could be provided at FLSC. Each pilots' levels of tactical execution performance was considered to meet or exceed the requirements for entering the live RF 08-3 exercise. Based on that, for the RF spin-up training, tactical execution was not considered a prioritized training objective. On the other hand, and to enable highest level of tactical execution at RF 08-3, the elements that were identified as needs among the pilots were knowledge closely attached to Nellis AFB and NTTR. Examples of such elements are local flight control procedures, ground operations, communication protocols, training rules and regulations, geographical knowledge and familiarity, and time and fuel management. The goal for the RF spin-up training was the provision of the experiences supporting the development of the knowledge and skills associated to these elements in order to allow a strong focus on tactical execution and highest possible performance at RF 08-3.

During RF spin-up every pilot was exposed to five different scenarios: a familiarization flight over NTTR, a Nellis AFB air traffic control procedures, a four vs. four air-to-air engagement over NTTR, and two large force employments (LFE) where airspace and time management was in focus. All through the week, researchers and SMEs (retired pilots) from the US Air Force Research Lab (AFRL) Warfighter Readiness Research Division (Mesa, AZ) and one pilot from the 65th Aggressor squadron at Nellis AFB supported the training, based on their local knowledge and experience. The aggressor pilot provided briefings on Nellis AFB air traffic control procedures, RF training rules and regulations, and NTTR airspace and target areas. The aggressor pilot also provided feedback on the performed sorties and was available for questions.

Red Flag 08-3. For the SwAF pilots the primary training goals for the exercise was to participate in LFE sorties (sorties with many and usually dissimilar aircraft types) in the air-to-ground role to further develop their capacity to participate in coalition operations. Another goal was to enhance and validate their close air support (CAS) and dynamic targeting (DT) ability and tactics. A number of different aircraft types participated in the exercise such as fighter aircraft in air-to-air and air-to-ground roles, tankers, command and control aircraft, and electronic warfare aircraft. Missions were flown twice per day and the enemy side consisted of special aggressor squadrons flying air-to-air missions to counter the LFEs. Simulated surface-to-air missile platforms provided a challenging ground based air defence (GBAD) "threatening" the aircraft.

Contributing research efforts

Mission Essential Competencies. Under a project agreement between FOI and AFRL, the Mission Essential Competencies (MEC) process (Colegrove & Alliger, 2003; Alliger, Beard, Bennett, Colegrove, & Garrity, 2007; Alliger, Beard, Bennett & Colegrove, in press) have been utilized to identify the essential core of experiences, knowledges and skills necessary for a combat mission ready JAS39 Gripen pilot (Bennett et al., 2006). MEC entails competence descriptors at various levels. For this evaluation a set of 36 knowledge and 50 skill requirements was used as a thorough description of "all the pilots need to know". The MEC knowledge definition is "info or fact that can be accessed quickly under stress" and one example of a knowledge statement for SwAF JAS39 Gripen is "radar warning and threat reactions". The MEC skills are defined as "compiled actions that can be carried out successfully under stress" and one example is "interpret rules of engagement".

Similarities and differences to previous MEC-based training evaluations. In the US/UK Red Skies study (Smith et al., 2007) similar research objectives such as the ones of this study were present. The Red Skies study was a DMO/MTDS (Distributed Mission Operations/Mission Training through Distributed Simulation) research trial, in addition to being a training event. This was not the case for the present study, thus the surveys that were developed and used placed a somewhat different emphasis on what was assessed. In the current study measures of development of knowledge and skills were in focus, whereas the assessment the Red Skies study placed an emphasis on MEC experiences. Further, no observer protocols of performance were used in this study. Given the criterion issues in performance measurement for a knowledge and skill set as extensive as that identified during the MEC process, training effect rather than performance was chosen as the primary construct to assess.

Surveys and data collection

To capture the relevant data from the participants before, during and after RF spin-up and RF 08-3 a battery of five surveys was developed. Table 1 provides an overview of where and when each survey was collected and what it captured.

Demographics survey. A demographics survey, not presented in Table 1, designed to capture background information about the pilots to be able to sort them by previous experience was collected before the RF spin-up training started.

Knowledge and skills addition training needs survey. The rationale for the knowledge and skills additional training needs survey (ATN) was to establish the participating pilots initial performance readiness prior to RF spin-up, and then repeatedly monitor their performance readiness development in terms of additional training needs desired for each of the 36 knowledge and 50 skill requirements. The pilots completed the ATN survey three times:

first and last day of the RF spin-up and last day of RF 08-3 (plan was to collect this data also the first day of RF 08-3 but that data could not be collected due to operational constraints). The participants also provided calibrated values of their previously rated performance readiness at the two last rating occasions (i.e., ATN3, ATN6 and ATN7 in Table 1). Their actual rating from the previous occasion was not presented to them when making this calibration.

Knowledge and skills perceived training value survey. The knowledge and skills perceived training value survey (PTV) assessed the pilots' ratings of perceived training value of each sortie flown at RF spin-up and RF 08-3 right after it was finished (i.e., to what extent did the previous sortie provide training value for the knowledge and skills). In a workshop, pilots, simulator instructors and training designers identified a subset of 28 knowledge and 40 skill requirements (out of the total of 86 from the ATN survey) based on how likely they were anticipated to be developed at RF 08-3. This was to decrease the intrusiveness of the survey since it was used within the training context.

Reactions survey. The reactions survey contained questions concerning both the affective and utility reactions to RF spin-up. It was collected both after the spin-up and after RF 08-3 (in both cases concerning reactions to RF spin-up).

Top 3-Bottom 3 survey. The top 3-bottom 3 survey (T3-B3), not presented in Table 1, was completed once at the end of each day, both during RF spin-up and during RF 08-3. The pilots straightforwardly listed the three best and the three worst events of the day and thus provided a wealth of qualitative data concerning the exercises.

Table 1. *Surveys used for data collection before, during and after RF spin-up and RF 08-3.*

Exercise	Time	Knowledge & skills additional training needs survey (ATN)	Knowledge & skills perceived training value survey (PTV)	Reactions survey (R)
RF spin-up (simulator exercise)	Before	ATN1. Current performance readiness		
	During		PTV1. Perceived training value	
	After	ATN2. Current performance readiness ATN3. Calibrated ATN1		R1. Affective & Utility reactions
RF 08-3 (live exercise)	Before	ATN4. Current performance readiness[a]		
	During		PTV2. Perceived training value	
	After	ATN5. Current performance readiness ATN6. Calibrated ATN1 ATN7. Calibrated ATN2		R2. Affective & Utility reactions

[a] ATN4 data was not collected in this study due to operational constraints at RF 08-3.

It is sometimes claimed that the reliability and validity of subjective ratings of many psychological constructs are insufficient and that it can be difficult to fully determine what has been measured. Doubts about their validity, although sometimes justified, should not be exaggerated. Even if the precision of any single rating is modest, data may still be sufficiently rich of information to be useful. The authors want to highlight the experienced pilot, as an intelligent filter against the complexity of the world, who has the capability to integrate his or her experience into a balanced measure. Note that this capability is dependent upon the nature of the construct that is being assessed, as not all mental processes are introspectively available. However, for assessment of operative field training effects, subjective ratings are useful and in most cases the only practicable way forward. Validity is further increased when ratings are collected before, during and after training, and when calibration ratings of previous status are included as this allows for control of what Golembiewski, Billingsley and Yeager (1967) calls alpha, beta and gamma types of change.

As a note to Bell and Waag's level "extrapolation of transfer to combat environment" the training during an exercise as Red Flag can in many respects provide better training of mission execution than war operations, as these highly realistic exercises allow more comprehensive and concentrated exposure to the full envelope of situations during a mission. In that respect, it constitutes not only an excellent training opportunity but also a representative training criterion.

Experiences from application

Survey data mapping to Alliger et al. augmented taxonomy

The Alliger et al. augmented taxonomy (1997) was used to conceptualize the surveys in order to capture data at all levels and meet expected analysis goals. Table 2 shows how the collected data set was linked to the specific levels.

Table 2. Summary of how data from each survey evaluates training on the different levels of the Alliger et al. (1997) augmented taxonomy.

1. Reactions		2. Learning			3. Transfer	4. Results
1.a. Affective	1.b. Utility	2.a. Immediate Knowledge	2.b. Knowledge retention	2.c. Behavioural/Skill demonstration		
R1 R2	R1 R2	PTV1[a] PTV2[a]	ATN2[b]	ATN5[c]	ΔATN6-ATN7[d]	Post analysis[e]

[a] PTV classified as immediate knowledge based on the assumption that a perceived training value indicates that learning has occurred. [b] Performance readiness status after one week of exposure to RF spin-up training. [c] Performance readiness status after exposure to RF 08-3 live training and performance criterion. [d] The delta between ATN6 and ATN7 is a measure of how much of the total training effect from before RF spin-up to after RF 08-3 that each pilot attributes to RF spin-up. This is based on the fact that ATN5, ATN6 and ATN7 for each knowledge and skill are rated at the same time. [e] Linking ATN training effects to the official SwAF training objectives for RF 08-3, for example to support cost-benefit analyses comparing magnitude of training effects and associated cost levels.

Examples of data

The data from the ATN surveys distinguished which knowledge and skill statements that received the most development during RF spin-up and during RF.

The mean of means from the ATN survey, together with SME mappings of each knowledge and skill to the SwAF training goals enabled a post-hoc analysis concerning the fulfilment of the same. The results of this type of evaluation yield interest at all level of the military hierarchy, decision maker not the least, and provide the often difficult mapping to the Results level.

In the MEC training gap analysis previously conducted for the JAS39 Gripen, a number of experiences for which the pilots desired more exposure were identified. Through SME mappings between these training gaps and the exposure to experiences that lead to reduction of training gaps during RF08-3, indications to what extent RF08-3 addressed the training gaps was extracted from the data.

The PTV survey provided detailed data from each sortie that can be used when analysing the training contribution from each specific sortie. The data also entail information useful when discussing the design of the sorties and when the SwAF in future exercises express expectations and training needs to the RF staff.

Through the T3-B3 survey a large amount of qualitative data concerning the exercises was captured, which corroborates the quantitative data.

Future analyses & methodological enhancements

The Alliger et al. meta-analysis (1997) reported that few studies of training present data and correlations between the different levels of Kirkpatrick's original taxonomy. One of the goals of this study is to analyze the effects of each measure but also to correlate between the levels of the Alliger et al. augmented framework as presented in Table 2.

In order to assess the training effect for the full range of knowledge and skills, which can be described as the lion's share of the pilots' full professional competence, comprehensive and structured assessment approaches such as the one described here are needed when studying training outside of the laboratory. Ultimately a mix of structured self-ratings, instructor ratings and logged performance data from simulator and/or aircraft would be collected. However, for studies of training with high ecological validity there are often issues with ceiling effects in the performance measures, such as number of air-to-air missile hits or bomb accuracy, when studying high performing pilots. This leads on to the almost the philosophical question whether training value and training effect are more representative constructs to evaluate than performance, although performance increases are the end goal of training.

The formulation of organizational goals and training objectives has always been a challenge for both researchers and the operational community. An observation from this study, describing a point observed many times before, is that if training goals were expressed in a more clear and precise form, training evaluations, and in the long run the training itself, could be developed much further.

References

Alliger, G. M., Beard, R., Bennett, W., Jr., & Colegrove, C. M. (in press). Mission essential competencies: An integrative approach to job and work analysis. In M. A. Wilson, G. M. Alliger, W. Bennett, Jr., R. J. Harvey., F. P. Morgeson, K. J. Nilan & E. Salas (Eds.), *The handbook on job and work analysis: The methods, systems, applications, & science of work measurement in organizations.* Mahwah, NJ: Taylor Francis.

Alliger, G. M, Beard, R., Bennett, W., Jr., Colegrove, C. M., & Garrity, M. (2007). Understanding Mission Essential Competencies as a Work Analysis Method (AFRL-HE-AZ-TR-2007-0034). Mesa, AZ: Air Force Research Laboratory.

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.

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., Jr., Borgvall, J., Lavén, P., Castor, M., Gehr, S. E., Schreiber, B., et al. (2006). International mission training research (IMTR): Competency-based methods for interoperable training, rehearsal and evaluation. In the proceedings of the Simulation and Interoperability Standards Organization (SISO), *European Interoperability Workshop (EuroSIW).* Stockholm, Sweden: SISO.

Colegrove, C. M., & Alliger, G. M. (2003). Mission essential competencies: Defining combat readiness in a novel way. In proceedings of 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.

Golembiewski, R.T., Billingsley, K. R., & Yeager, S. (1976). Measuring change and persistence in human affairs: types of change generated by OD designs. *Journal of Applied Behavioural Science*, 12, 133-157.

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.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, 78, 311-328.

Smith, E., McIntyre, H., Gehr, S.E., Symons, S., Schreiber, B., & Bennett, W., Jr. (2007). Evaluating the impacts of mission training via distributed simulation on live exercise performance: Results from the US/UK "red skies" study (AFRL-HE-AZ-TR-2206-0004). Mesa, AZ: Air Force Research Laboratory.