Assessing the Training Potential of MTDS in Exercise First Wave

Sara Elizabeth Gehr
6030 S. Kent St.
Mesa, AZ 85212
USA

Liz.gehr@mesa.afmc.af.mil

Margaret Schurig
6030 S. Kent St.
Mesa, AZ 8521
USA

Margaret.Schurig@mesa.afmc.af.mil

Lesley Jacobs
TNO Defence, Safety & Security
Waalsdorperweg 63
2509 JG The Hague
The Netherlands

Lesley.Jacobs@tno.nl

Jelke van der Pal
National Aerospace Laboratory NLR,
Anthony Fokker Weg 2
1059 CM Amsterdam
The Netherlands

pal@nlr.nl

Winston Bennett, Jr.
6030 S. Kent St.
Mesa, AZ 85212
USA

Winston.Bennett@mesa.afmc.af.mil

Brian Schreiber
6030 S. Kent St.
Mesa, AZ 85212
USA

Brian.Schreiber@mesa.afmc.af.mil

ABSTRACT

Exercise First WAVE (EFW) was conducted in November 2004 and was the first large-scale NATO MTDS event that focused on investigating and providing COMAO training for warfighters and mission support staff in a distributed synthetic environment. To assess the training potential of EFW, the NATO partners of the seven countries participating in EFW developed a competency-based approach to training needs assessment and evaluation. The approach, which is derived from the Mission Essential Competencies process, is designed to determine training and rehearsal needs and gaps, and to assess the effectiveness of distributed mission training exercises. This paper starts with a description of the development and application of a comprehensive set of training evaluation and performance assessment instruments and methods, which were based on identified requirements and gaps in other recent training research events. These methods were further adapted to the requirements of EFW and include both subjective and objective data collection and tracking of training events. The overall EFW assessment results will be presented and data and experiences from the Netherlands will be discussed with a focus on the lessons learned from an operational point of view. The usefulness of the chosen assessment approach, with respect to the assessment and evaluation of the training potential of Mission Training through Distributed Simulation (MTDS) will be discussed. Implications from EFW for broader, international implementation of a competency-based assessment approach for high fidelity, distributed simulation training will conclude this paper.

Gehr, S.E.; Schurig, M.; Jacobs, L.; van der Pal, J.; Bennett Jr., W.; Schreiber, B. (2005) Assessing the Training Potential of MTDS in Exercise First Wave. In *The Effectiveness of Modelling and Simulation – From Anecdotal to Substantive Evidence* (pp. 11-1 – 11-16). Meeting Proceedings RTO-MP-MSG-035, Paper 11. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.asp.

1.0 INTRODUCTION

Assessing training effectiveness in large scale exercises, for all training audiences involved, is challenging. This also applies to virtual large scale exercises often referred to as Mission Training through Distributed Simulation (MTDS). The challenges faced are threefold:

- As in any large scale training event, establishing training effectiveness and value for all participants is complex and requires careful scenario development and monitoring as well as establishing specific training objectives for the entire training audience.
- In MTDS, unlike e.g. FLAG exercises, participants operate at distributed nodes during all phases of the training (plan, brainstorm, brief, execute and debrief). This requires an assessment framework that supports distributed data-collection. For objective data assessment software is needed that can capture relevant network data. For the subjective assessment measurements this calls for an extensive assessment protocol, including observer training and assessment support tools for experts conducting the assessment such as White Force Liaisons Officers (WFLOs), researchers, engineers, and exercise management (White Force).
- Finally, MTDS is a new training “medium”. Assessing the training potential of this type of training and incorporating this into future MTDS events is therefore equally important as assessing the actual training value and effectiveness.

To be able to assess training effectiveness and potential it is essential to define an assessment framework that is based upon well defined (training) objectives. In Exercise First WAVE (Warfighter Alliance in a Virtual Environment) this was done by using the Mission Essential Competencies (MEC) approach. Furthermore, it is essential to include all relevant roles in the assessment effort, to collect both objective and subjective performance measurements, and for the latter to conduct pre, during and post exercise data collection. EFW investigated the first-ever comprehensive approach to distributed assessment for large scale virtual exercises and there are many lessons learned that can be taken away and applied to future MTDS events. This paper describes this effort and the lessons learned from an international assessment point of view combined with the specific experiences in the Netherlands, who participated in the exercise with substantial operational involvement- an F-16 flight (4-ship), a Forward Air Controller (FAC), a mission commander, WFLOs and mission support personnel - making it an interesting MTDS case study to report the actual operational experiences and assessment results.

2.0 DESCRIPTION OF EXERCIZE FIRST WAVE

Exercise First WAVE (EFW) was designed to investigate and demonstrate the potential of Mission Training via Distributed Simulation (MTDS). Seven countries- Canada, France, Germany, Italy, The Netherlands, The United Kingdom and The United States- participated with their national simulator assets and military, scientific and industrial personnel. The primary training audience consisted of fast fighter jet pilots and mission support personnel, who conducted Composite Air Operation (COMAO) missions based upon Operation Allied Force (OAF) - the allied air campaign against Serbia and Kosovo in 1999.

EFW had multiple overall, training and training management objectives. To meet these objectives a realistic COMAO scenario had to be developed that was operationally challenging for all participating assets and operators. Therefore, the missions were designed to promote realistic interactions between aircrew and placed the aircrews in a dynamic and complex synthetic environment. The OAF scenario allowed for a range of force mix and mission tempo to be considered and a suitable mission profile was developed for each day of the exercise. To be able to manage, monitor and execute this scenario, exercise management via a White Force and (flexible) exercise management facilities, such as monitoring and logging tools, were installed. Given the distributed state of the participants during each phase of the mission, this also included tools enabling distributed planning, briefing and debriefing, such as Video Telephone Conferencing (VTC) and Smartboard technology. Traditionally, a White Force is a team of experts who organize and run large-scale live training exercises. For EFW the White Force included the following roles and responsibilities: Exercise Director, Blue Force Leader, Instructor Pilot of Record (IPOR), Mission Planning Support, Blue Force Liaison, CGF Support, Tech Liaison, VTC moderator, Intelligence, Administrative Support/ATO and a Red Force Liaison Officer. Furthermore, for training assessment, national White Force Liaisons Officers (WFLOs) were located both at the Exercise Control Centre located in Lossiemouth, UK and at each distributed (national) site.

EFW was conducted over a period of one week from November 14 until November 19, 2004. The first day was intended to give the White Force an opportunity to familiarize themselves with the Exercise Management System and for the aircrew and the technical teams at each node to prepare for the training. The next three days consisted of full missions flown in the scenario, with mission commanders from Canada, The Netherlands and the UK. For formation debriefs, each node had the ability to fully reconstruct the mission, enabling replay of any mission event. The overall mass debrief was led by the mission commander and was accessible via VTC to the EFW participants at all sites. A synchronized mission play-back capability was therefore required at each node. On the last day a debrief session, the so called Hot-Wash was conducted under supervision of the various EFW Task Team Leaders, to ensure that lessons identified, from both a technical and a training perspective, were captured. After the exercise, an international three day After Action Review (AAR) was conducted in December 2004.

Given the fact that EFW was the first NATO research project to connect various legacy simulators in an attempt to asses the training potential of MTDS, there was a strong focus in the project to come up with feasible and achievable solutions. Throughout the development and preparation of EFW there was a strong focus upon matching operational needs with what was technically feasible. This resulted in a highly iterative EFW development process. An example of this iterative process is the continuous effort the team undertook to adapt the scenario and missions to all players' needs. The major challenge in EFW came with the withdrawal of the virtual assets of the US and the UK in July 2004. Although this was identified as a significant loss, the operational members of EFW agreed upon the usefulness of going ahead with EFW to assess the training potential of MTDS. The operational challenges were addressed sufficiently by revising the scenario and re-allocating various roles to blue computer generated forces (CGF) entities. Without the option to postpone the exercise to 2005, EFW had to run in the September to November 2004 timeframe. Despite difficulties with the availability of operational members of participating countries, operators with the appropriate level of experience and qualifications (similar to FLAG exercises) could participate in the exercise. Finally, the change in dates and rework on the EFW network resulted in very limited technical testing time: one week instead of the initially planned three month period.

3.0 DEVELOPMENT OF ASSESSMENT INSTRUMENTS AND METHODS

The assessment team, one of the five EFW task teams, was responsible for developing the assessment instruments and methods used during EFW. The team consisted of scientific and operational members of each participating nation. At the various EFW project meetings, the team established what performance measures should be collected during EFW to assess training based on the overall objectives of the exercise. This data fell into two broad categories, subjective and objective, each of which will be discussed in detail in the following sections. For the subjective data, the team also worked together to decide what methods should be used to collect the data. That is, who receives each survey, when should they complete it, what surveys do different types of participants (engineers, observers, pilots, controllers, etc.) need to fill out. This was an essential part of the planning process, as the nature of EFW dictated that data would need to be collected at the eleven different participating sites. The assessment team also worked to ensure that there would be someone at each site who was familiar with the assessment goals, surveys and methods. In many cases, this required a member of the assessment team to travel to the various locations participating in EFW to collect the data during the exercise week. Thus, the protocol needed to be decided upon, agreed to, and explained to all who would be collecting the data in advance of the exercise. Another essential part of the assessment process was to ensure, in advance, that when the schedule for the exercise week was developed, that there was sufficient time in the schedule (pre, daily, and post) for assessment. This time in the schedule also served to partially manage the expectations of the participants so that they knew the assessment would happen, and they did not expect to leave immediately upon completion of the debrief.

4.0 SUBJECTIVE DATA COLLECTION

The main focus of the subjective data collection was on individual and role-specific surveys. Table 1 gives an overview of the developed surveys and participants that completed each survey.

Survey number	**Survey Title**	Air to Air Pilot Air to Ground Pilot	Mission Commander	Red GCI AWACS	Controller FAC	White Force Liaison Officer (WFLO)	Researcher Engineer Observers	SME
0	Initial and Extended Demographics forms	✓			✓			
1	Experiences and training opportunities surveys	✓		✓				
2a	Mission specific expectations (pre-exercise)	✓		✓				
2b	Mission specific experiences (post-exercise)	✓	✓	✓				
3	Mission process assessment							✓
4	Attitudes toward First WAVE and distributed mission training	✓			✓			
5	Engineering and behavioural observer grade sheets					✓	✓	
6	Daily summary “Top 3” and “Bottom 3” events	✓	✓	✓	✓	✓	✓	✓
7	WFLO observational surveys					✓		

Table 1: Overview of the EFW surveys.

Surveys 1-3 are based on Mission Essential Competencies (MECs). MECs are “Higher-order individual, team, and inter-team competencies that a fully prepared pilot, crew or flight requires for successful mission completion under adverse conditions and in a non-permissive environment” [1]. MECs can be broken down into experiences, knowledge, skills, environments and supporting competencies By using MECS in combination with the overall established (training) objectives, the surveys investigated whether EFW did or did not provide a useful experience to obtain a certain mission essential competency. Therefore, also based upon the MEC process conducted previously in the US, UK and in NATO at TLP, experiences, environments and supporting competencies were used in the survey development for the various EFW participants.

Surveys 0 through 2a were to be filled out before the start of EFW. Survey 0 was filled out by the pilots and controllers and collected basic demographic information about each of the pilots and controllers including years of service, experience at various exercises, etc. Each country was able to tailor this survey for their participants. Survey 1 was also filled out by pilots and controllers and there were two parts. Both parts had a list of MEC experiences for the rows that correspond to the primary role of the pilots and controllers, and different training environments for columns. The experiences were based on the combined US, UK & TLP MEC experiences. The environments were specific for each country. The first survey asked participants to tell how often they had had each experience, in each environment, in the past year. The second part of the survey asked to what extent each experience can be experienced in each of the environments. Survey 2a, also filled out by pilot and controller participants, listed the MECs and asked respondents to evaluate how well they expect EFW will be able to train the different experiences. Then, in a free response section, it asked participants to list any specific expectations or objectives they had for the entire exercise.

Four surveys were filled out at the end of each exercise day. Survey 3 was filled out by a Subject Matter Expert (SME) who rated the performance of the local team, the overall blue team, and the overall red team. It was based on the MEC supporting competencies. Survey 5 was filled out by engineers, observers, researchers and WFLOs. It was organized by mission phase (planning, briefing, execution and debriefing) and in each phase listed several possible problem areas. It asked the participant for each area to explain the details of any problems in that area, and what was, or should have been, done to solve the problem, and to rate the severity of the problem. Survey 6 was a short daily survey that was filled out by all participants. It simply asked them to fill in the top three and bottom three events that happened during each exercise day. This was not necessarily what was good or bad about the mission itself, but what was good and bad about the whole day. The final daily survey, survey 7, was filled out by the WFLOs at each site. The survey asked the WFLOs, for each of the four mission phases (planning, brief, execution, and debrief), questions about how each part of the exercise was going from their perspective.

Two of the surveys were given to pilots and controllers at the end of the EFW week. Survey 2b, was similar to survey 2a, but it asked how effective the entire exercise was at training the list of experiences. There was also a version of this survey for the mission commander to fill out on the day that he served as mission commander. This one had no corresponding pre-survey (2a) but asked the mission commander how well EFW was at training a list of mission commander experiences. Survey 4 asked pilots and controllers how much they agreed or disagreed with a list of statements about the usefulness of EFW/MTDS as training medium. Finally, at the end of the week, a researcher performed a structured interview with the WFLOs to get their unique feedback about the events of the week.

In order to make data collection as simple as possible for each participant, three different methods of data collection were available at each site. First, a web-site was developed and each participant (engineer, observers, pilots, controllers, etc) was able to logon with a unique logon. The website would then give that participant access to fill out the surveys designed only for his/her specific role. In addition, the researchers who were members of the assessment team were able to access the website, which was up before EFW, to collaboratively develop and comment on the surveys. Next, a "mini-browser" was developed to collect data. The mini-browser had the same data collection capabilities as the web-site, but could be run on any computer with Microsoft Internet Explorer even without internet access, hence allowing electronic data collection at high security sites (as was the case for most participating military sites). The survey results were saved locally and were e-mailed at a later date for uploading to the results database. A third means to collect the same data were paper surveys. All assessment personnel at the sites had paper copies of the surveys that could be filled out, and then entered in the web-site or mini-browser for incorporation into the results database.

5.0 OBJECTIVE DATA COLLECTION

The subjective data is invaluable for capturing user and observer opinions about the perceived value and utility of an MTDS exercise, but objective data (e.g., kills, bombs on target, shot information, etc.) serves to best quantify performance. Ideally, objective data could be captured automatically from each and every entity involved in an MTDS exercise. These measures could prove useful as training feedback during mission debrief or for using the individual entity and shot data for objective human performance assessment. Furthermore, these data could (theoretically) be aggregated up to a team, package, or a force level (i.e., red/blue) to summarize performance and address important MTDS questions (e.g., learning rates and training transfer). Since the goal was to identify and implement a standardized assessment system, a method was sought that could be used on various future multinational exercises. That is, such a system would have to be capable of assessing exercises using different simulators from different countries. The best approach is to use an assessment system that leverages the exact same technology methods that allow distributed simulation to occur in the first place—namely the interoperability standards of DIS and HLA. In theory, an objective assessment tool relying on the same standards imposed on all simulation entities involved in MTDS would be able to always capture objective measures whenever and wherever an MTDS exercise occurred. That is, the objective assessment tool would be completely simulator independent and allow for maximum flexibility in its use. Therefore, the team set out to capture assessment measures by replaying the EFW DIS/HLA data packets that were shared over the network (i.e., the recorded or "logged" scenarios). This approach would allow for more detailed analysis on EFW and theoretically any other future multinational simulated event.

Recent performance measurement work at AFRL has resulted in a software tool that could capture human performance assessment data from a distributed network, called the Performance Effectiveness Tracking System (PETS) [2]. This system has been demonstrated and used to assess performance on a single DIS network. For a single site, this system has been successful and has been used to objectively quantify the value of distributed simulation training [3]. A new architecture of this system was used to derive outcome metrics (e.g., kill ratios, weapon hit rates, bomb distance errors, etc.) and process/skill oriented metrics (e.g., time spent in enemy weapons zones or measuring at pickle of a weapon the altitude, loft, g-loading, mach, clear avenue of fire) for both friendly and adversary entities in each EFW mission.

6.0 OVERALL EFW ASSESSMENT RESULTS

6.1 Subjective Results

This section will summarise the overall subjective data results and will report data at the *aggregate* level with a focus on the results from the operational participants. Eleven pilots and four controllers completed the demographic survey, answering questions about their background and flying experiences. The pilots ranged in qualification level from student pilot to mission commander. The controllers had an average of 8 years of experience at their current position.

Before EFW, in survey 2a, pilots and controllers were asked about their expectations about how effective EFW would be at giving them certain experiences based on the MECs. They responded on a 5 point scale, ranging from 0 (not at all effective) to 4 (very effective). They also had the option to enter DNA (does not apply – this experience cannot be replicated in this exercise) if they felt that EFW could not replicate the experience. After EFW, participants filled out a similar survey (2b), asking them how effective the exercise was at providing the same experiences. Again, they had a 5 point scale, and a DNA option. Overall, Air to Ground pilots rated their expectations at 2.2 (std = 1.1) averaged across all 8 pilots and 77 experiences, while A/A pilots had an average expectation of 2.15 (standard deviation = 1.04) averaged across all 8 pilots and 56 experiences. Controllers had an average expectation of 2.13 (standard deviation = 0.98) averaged across all 4 controllers and 67 experiences. After EFW, the experiences were rated lower in effectiveness than the pilots and controllers had expected at the beginning of the week. The 6 air to ground pilots who completed the effectiveness survey at the end of the week rated the effectiveness of EFW an average of 1.77 (standard deviation = 1.39) across all the experiences, 0.55 lower than they had expected at the beginning of the week. The 7 air to air pilots gave an average effectiveness rating of 1.72 (standard deviation = 1.32), 0.49 lower than they expected at the beginning of the week. The average of the 4 controllers effectiveness ratings in EFW was 1.12 (standard deviation = 1.29), 1.06 less than they expected at the beginning of the week. Although experiences at the end of the week were lower than expectations were at the beginning of the week, this was mainly due to simulator or network, rather than mission, isues (this is discussed further in the comments below). In addition, the person who was mission commander of the day filled out a separate, but similar, survey at the end of each day that he served as mission commander, asking how effective EFW was at giving him certain experiences, again, based on MECs. Their average effectiveness rating was 1.97 (standard deviation = 1.35) for the 2 mission commanders who filled out the survey. The mission commander was not given a survey before the exercise to gauge his expectation from training, so no comparison can be made.

For each mission day in EFW a subject matter expert (SME) rated the performance of the team of aircrew (air to air or air to ground) or controllers at their site, as well as the performance of the overall blue force, and the overall red force. They graded them on a scale from 1 (performance indicates a lack of ability or knowledge) to 5 (performance reflects an unusually high degree of ability). They also had the option to choose N/A if they did not have the opportunity to observe a pilot or controller in a particular area. The average performance of the air to air teams, averaged across all sites and days, as rated by the SMEs is 4.15 (standard deviation = 0.44), indicating that even though the teams faced multiple technical obstacles, they were able to perform their role very well. The SMEs rated the overall blue package performance as 4.06 (standard deviation = 0.39), again averaged across all sites and days. This also indicates that the blue force overall performed their tasks well, overcoming the technical obstacles they faced. Although many SMEs chose the N/A option when rating the red force, the overall rating of red team from the SMEs that did respond is 2.89 (standard deviation = 0.87).

At the end of each day of the exercise, each participant - aircrew, controllers, researchers and engineers - completed a “top 3 bottom 3” survey, listing the 3 best and 3 worst events from the day. On days that the connections proved reliable and the missions were more realistic, the positive comments reflected positive potential training value, such as “good SA given by AWACS,” which was said on Wednesday, or “very interesting tactical situation” said on Thursday. Another positive comment was “lessons learned during flight; training habit patterns do not all work in/during wartime. So good simulation of war time operations and good to see that you do different things; switch actions in war compared to simulated training (day to day). Good training.” also said on Tuesday. Many of the negative comments involved poor connectivity, rather than poor performance on the part of players. For example “only intermittent connectivity with Europe during the exercise; no communication with AWACS” said on the Tuesday of the event, reflects on engineering challenges. Another comment “network unstable caused players to disappear,” reported on Monday, also reflects network, not training, issues. Again, the positive comments reflect high potential training value for this exercise, while the negative comments reflect engineering challenges.

At the end of EFW, aircrew and controllers were asked to rate their reactions to EFW and use of MTDS as training tool. The response scale ranged from 1 (Strongly Disagree) to 4 (Strongly Agree). Some questions were worded such that a positive reaction response would require a low score to ensure that the participants were reading the questions, and not simply circling similar numbers for all the questions. Averaged across the 11 pilots who completed the survey, and all 67 questions, reverse coding those that required it, the average was 2.69 (standard deviation = 0.91). Two controllers completed the survey, and, averaged across the 62 questions they were asked, the average response was 2.67 (standard deviation = 0.81). These data indicate that both the pilots and controllers had an overall positive reaction to the exercise, although there is room for improvement.

6.2 Objective Results

As mentioned, PETS is a promising tool for objectively assessing MTDS exercises. Using it in EFW revealed some predictable results and, as the general objective assessment results show, EFW clearly demonstrated that long-term objective assessment is feasible, but more work is definitely needed. The simulation systems in EFW did successfully interoperate, but requirements for a successful automated objective assessment system of any exercise necessitates substantially higher network traffic quality than that demonstrated in EFW (e.g., no missing network data fields, no inaccuracies, no significant time delays, etc.). As an example, a fired missile may be launched, seen fly-out, and destroy another entity successfully (i.e., all seen "correctly" by observers/participants), but some fire network PDU data may be missing. The missing data poses problems for automatically capturing who shot what weapons and subsequently impact kill ratios. Quite contrary to observation and participation for "successful connectivity," collecting objective human performance data from MTDS events *requires a robust MTDS engineering environment*, as the raw data in the exercise needs to be complete, timely, and accurate to allow for proper human performance assessment. Since the primary target for EFW was "successful connectivity," obtaining valid objective human performance data from EFW was never a goal, but merely a desire. The findings and (unclassified) results of the objective assessment work are presented below and although specific examples are reported here of issues identified in EFW, these issues exist in *many* simulation exercises. Therefore, when thinking about the feasibility and extensibility of automated assessment, these precise issues should be carefully examined for each simulation facility.

Raw Data Deficiencies. Please note that the data deficiencies found and described below might be network latency induced, due to simulator or site specific problems, or created by an issue with the log recorder itself. First, log file data was incomplete (e.g., missing fire, detonate, communication ID, etc.). Some missing data may be technically in accordance with DIS/HLA standards, but to achieve thorough automated objective assessment, all entities from all sites should transmit complete network information. Second, log file data appeared to contain inaccuracies (e.g., aircraft performing 35G turns, neutral force missiles, mismatching fire/detonate IDs). These issues create complications for any software system to automatically count how many unique entities actually participated in the scenario, accurately report shots, determine kills, etc., which again impacts the automatic calculation for important statistics such as force size or kill ratios.

Exercise Protocol. Certain rules should be followed for MTDS exercises to allow for more valid objective metrics, especially outcomes. If, for example, console operators at any given site are allowed to use "shields", then calculating very simple (but important) outcomes such as kill ratios becomes problematic and the validity and generalization of the results becomes highly questionable. Therefore, the rules should reflect a "realism" perspective and should be followed for any scenarios that will be used as part of a formal assessment. The rules should generally revolve around the theme, "If it cannot be done in actual combat, then it should not be allowed in MTDS." During EFW, the agreed exercise protocol allowed shield use. Unrealistic exercise protocols by any site greatly undermine objective assessment. Even if measures for kill ratios can be computed correctly in light of "shield" use, the argument or conclusion strength based on those numbers is severely reduced.

PETS Beta Issues. As a direct result of variability in incomplete fire/detonate PDUs, the PETS system duplicated some shots in the data. Also, as part of the first generation PETS system, substantial logic was programmed to account for missing data and other network anomalies. As one example, if a fire PDU went missing, the original PETS architecture would do proximity calculations on nearby aircraft to the just-appeared missile entity and determine/assign the most appropriate firing entity. In the beta PETS system used for EFW, only a fraction of extra logic had been incorporated. This resulted in a problem with the PETS analysis; the new PETS system treated much of the missing raw logger file data as missing and therefore reported a high percentage of missing values in the output.

Descriptive Summary. During the last two exercise days, PETS was able to collect and calculate some, but far from all, data. Because of the issues mentioned above, the analysis became more of a human effort than an automated one, also requiring the writing of additional software specifically to pull out and analyze certain types of PDUs (e.g., fires, detonates, entity states) to aid in validation and have some degree of confidence in the limited data obtained. First, Blue and Red force size/employment were counted. This was done by examining the total unique participating entities that occurred during the last two days of the exercise. A unique participation is defined as a unique life or entity for the Red or Blue force (e.g., an aircraft or SAM site), and there were 41 and 43 valid red entities and 38 and 41 valid blue entities, respectively. Since data and protocol issues arose, the objective data automatically collected counted a different number of entities. PETS consistently recorded more lives/entities than the true valid number predominantly due to excessive latencies creating time-outs. When these entities re-emerge, PETS recognizes them as a new, unique life/entity, while the true "intent" of the exercise was that they never should have left and were the "same" entity. Continuing the unclassified descriptive summary, the team next counted total munitions (offensive and defensive) employed over the last two exercise days for both Blue force and Red Force. There were a total of 95 Blue force munitions, of which 17 were defensive (i.e., chaff/flare). There were a total of only 11 Red force munitions employed, of which none were offensive. A fairly high proportion of valid measurements were captured for many metrics (e.g., launch ranges). Actual values for those measurements, if desired, will need to be reported in an alternative reporting system. Given the raw data issues previously described, it was a pleasant surprise to find the fairly high percentage of measurements accurately captured for each metric, but it is clear that improvements must be made to incorporate an automated assessment tool like PETS successfully into future MTDS events.

7.0 THE NETHERLANDS EFW PARTICIPATION

The Netherlands participated in EFW with four F-16 flight simulators - three Unit Level Trainers located at Volkel Air Base (Sting 1-3) and the GFORCE simulator, located at NLR Amsterdam (Sting 4) and the FAC simulator (Windmill), located at TNO in The Hague. Furthermore, military personnel from the RNLAF and RNLA were involved, together with scientists and engineers from TNO and NLR. Due to the distribution of the national simulator assets, the Netherlands operated both on a team and on the international collective level in a distributed manner. For the Netherlands the primary reason for participating in EFW was to asses the training potential of MTDS, especially since ULT-Joind, the national project that preceded EFW, already focussed on investigating the technical feasibility of MTDS on a national level.

The work conducted by the overall EFW assessment team was to collaboratively develop a common framework for training delivery and evaluation that could serve as baseline methodology for future MTDS events. The Netherlands assisted the assessment team with operational and scientific support, especially to review and revise the surveys, developed for the subjective data collection. Furthermore, various Dutch experts (WFLOs, researchers and engineers) did performance assessment during EFW at Lossiemouth, Amsterdam, The Hague and Volkel. In addition to the overall assessment work, the Netherlands defined a national assessment approach, based upon national research questions and the Air Force Chief of Staff's research assignment. Therefore, the surveys were complemented with a rather pragmatic approach, given the limited time available, to address the assessment of both EFW international and specific Dutch objectives. The followed approach consisted of briefing the Dutch participants on EFW and its objectives and conducting a so-called Hottest Wash, this was a national evaluation session, on the last exercise day with all Dutch participants. During the Hottest Wash, statements, based upon the objectives, were presented to the participants as a group and they were asked to give individual ratings (e.g. a 5-scale rating from not effective – very effective) to each statement. After gathering the scores, some debate could follow when ratings differed strongly, providing more insight to the various points of view. Finally, some individual interviews (e.g. with the FAC) were conducted to establish the EFW lessons learned.

7.1 Dutch Assessment of International Training Objectives

During the hottest wash with all Dutch participants (operators, engineers and scientists) conducted at Volkel and Lossiemouth (via VTC) all national and international (training) objectives were evaluated and assessed. The international training objectives (ITOs) were:

ITO 1: To practice daytime COMAO procedures, employing fighter escort/sweep, AAR, SEAD, RECCE and AEW in a hostile EW environment.
ITO 2: To exercise procedures for defensive operations with Fighter Areas of Responsibility (FAORs) and point defence tasking.
ITO 3: To employ EW (Electronic Warfare) resources in support of offensive and defensive air operations.
ITO 4: To plan and integrate a multi-national COMAO in a defined threat environment.
ITO 5: To brief a COMAO package generated from dispersed locations.
ITO 6: To conduct mission debriefs from dispersed locations.
ITO 7: To engender efficient team-working skills between nations and differing elements of the COMAO package
ITO 8: To develop a technical appreciation of real world threats
ITO 9: To expose aircrew to situations they would not normally encounter in a peacetime environment.

Each ITO, apart from ITO 2, which was not applicable to the Dutch role in EFW, was assessed by all participants with appropriate roles and expertise to assess a specific objective. For example, objective 8 was only answered by engineers and operators. Figure 1 summarises the assessed level of effectiveness per international training objective (ITO) according to the Dutch participants and shows the percentage of participants that rated per objective the various levels of effectiveness.

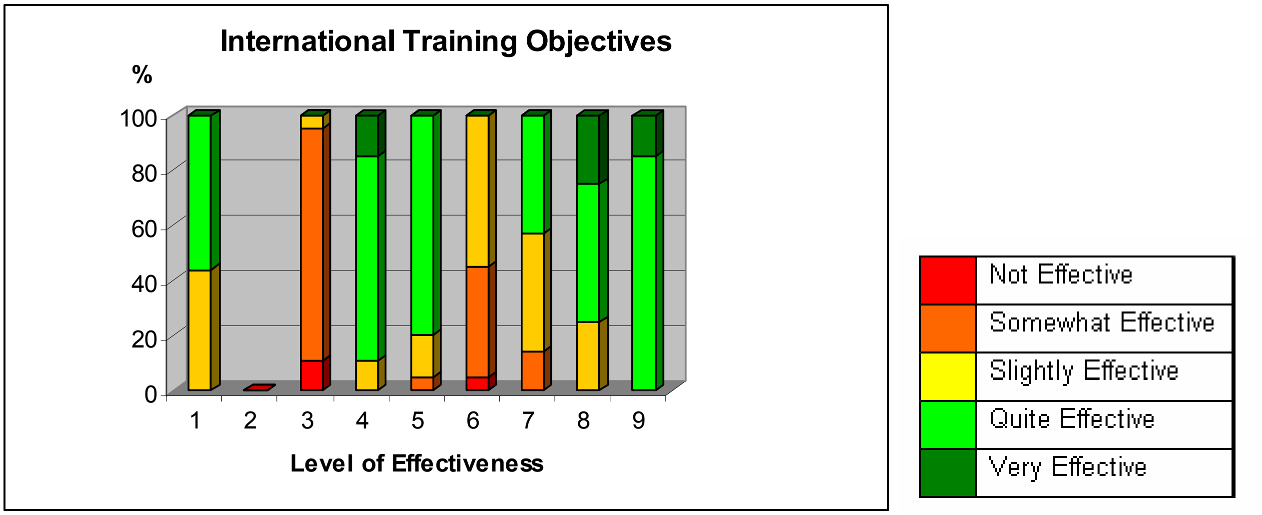

Figure 1: Assessed level of effectiveness per ITO

Figure 1 shows training objectives 1, 4, 5, 8 and 9 assessed as, at least, quite effective by the majority of the participants. This clearly indicates both positive experiences during EFW and also demonstrates the training potential of MTDS. Objective 7 shows equal numbers of participants rating it as quite effective and slightly effective. The Hottest Wash revealed that senior operators were more positive about the demonstrated training potential at a national team level while junior operators were more positive about the training effectiveness perceived at the collective level. It was also concluded, for this objective, that the effectiveness also depends on the international participants in a future MTDS event and adherence to international training objectives by each nation. Objective 3 clearly shows the insufficient demonstration of EFW in this area. Objective 6, 'Dispersed Debriefing', is rated as slightly effective at most, although EFW survey 2b reveals the pilots were less negative about debriefing a large force after the exercise than they expected up front (rating increased from 2.2 to 2.6).

7.2 Dutch EFW Survey Results

From the 77 items in the 'Air to Ground specific experiences' survey 1, 46 items could be used for NL analysis[1]. After the exercise, the average rating on these experiences (assuming all experiences equally important) was 2.3, in between 'somewhat effective' and 'quite effective'. This is a lower score than the pilots pre-exercise expectations, which were 2.6 ('quite effective'). Analysis revels that this is purely a result of limitations of the used simulators and network as shown in Table 2. For items related to the mission organisation, the exercise was – overall - as effective as expected: "quite effective", which is a good result given the first time nature of this MTDS event.

[1] Items were excluded from the analysis when two or more out of the five pilots indicated 'Does Not Apply'.

Table 2: Survey results for A/G experiences (Dutch pilots only)

	PRE (N= 5)	POST (N= 5)
Simulator/network dependent items (27)	2.6	2.0
Mission-organisation/exercize design dependent items (19)	2.7	2.7
All items (46)	2.6 (SD = 0.9)	2.3 (SD = 0.4)

7.3 Dutch assessment of national objectives

Three national objectives, looking more into the achieved training value during EFW that were also assessed and discussed in the Hottest Wash, will be summarized here.

1. Was training achieved at the team level? The Dutch EFW participants were very positive about the training achieved at a (national) team level. Positive training points mentioned were, for example: for the 2 ship-FL to brief and lead as a 4 ship-FL, to coordinate as flight in a COMAO and to fly/operate the ac/sim without all kinds of restrictions that apply to live training events. The mission support elements were also positive about the fact that they could follow the entire mission with the exercise monitoring tools (this provided more insight). Perceived weak points were that the team as tasked was not co-located during the planning phase while there were no tools for distributed planning available (see also national objective 2).

2. Does working from dispersed locations affect the team training? Briefing and execution from dispersed locations worked excellently with the Smartboard, VTC and CTM software. However, for mission planning and mission debrief at formation level, it was felt a disadvantage to be on multiple sites, due to the fact that fully active involvement of Sting 4 (who was located in Amsterdam) in the planning and debrief was not possible. Apart from the fact that this can be resolved technically by having all tools and software available at every site, there was a strong consensus that a team should preferably always be co-located to engender effective team working skills. This is even more required when the flight provides a mission commander.

3. Was training achieved at the collective level? The Dutch participants were also positive in their assessment of this point. Especially during mission execution, EFW could, despite its technical limitations, be compared with real life exercises such as Red Flag or Maple Flag. Using OAF for the EFW scenario contributed strongly to its realistic value. Finally, it was found useful that EFW provided the opportunity to see how different nations train and operate (also good for benchmarking the national training conops). Disadvantages experienced for achieving training at the collective level were related to the insufficient availability of CTM and VTC for mission planning at several days. The FAC indicated that it would be useful for future MTDS events, to train e.g. CAS and TST type of missions also with other international players in the scenario.

7.4 Discussion

There is an apparent discrepancy between the overall modest results of the international EFW surveys on the one hand (showing that even modest expectations were not met during the exercise) and on the other hand the overall positive feedback received during EFW, the international Hot Wash on the last EFW day, the positive operational evaluation conducted at the AAR in Orlando, December 2004, and the positive Dutch national assessment results (both surveys and national Hottest Wash). Potential causes for this discrepancy between the surveys and the group sessions could be:

- *Different focus of the instruments.* The surveys have been designed to measure *experienced training value* while the group sessions were aiming to assess the *training potential of* MTDS given the experiences. The surveys therefore stimulate reflection on the events as experienced given the limitations of the current network or simulators, while the group sessions stimulate reflection on a more general level of events and to 'this type' of exercises, network, and simulators.
- *Reliability of the data.* The limited number of participants who filled in all surveys might not be representative to all EFW participants. Also, the complexity (as perceived as such by many of the Dutch respondents) and extensiveness of the surveys may have caused misinterpretation and lack of time to fill in the surveys. On the other hand, group sessions may have caused a group bias.
- *Varying national views.* The positive assessment of the Dutch participants may not be shared by other EFW participants. However, this is quite opposite from the international operational and technical experiences exchanged at the AAR in Orlando (both formal and informal). Another possible explanation for the positive Dutch results might be that due to the substantial operational participation in EFW, the primary focus on training for *all participants* and lack of technical problems at a national level. This country was able to operate in EFW in a way that perhaps resembles more closely how future MTDS events should become than could be achieved by smaller operational teams of other participating nations.
- *Operational vs Technology effect.* (Dutch) operators see EFW on-the-whole fail to meet their expectations on simulator and network dependent items (e.g. reacting to realistic type threat) that require highly realistic simulation, while they experienced effective training - as expected - on simulator independent (or less dependent) items (e.g., restricted ROEs and unlocated threats) which require highly realistic mission (exercise) design from ATO to debrief. Given this effect it is not surprising to find the participants more positive in group sessions asking about the potential of MTDS than on the more specific EFW survey items asking about experienced training value. Participants consider serious simulation problems like 'not working' RWR, 'unrealistic threat behaviour' (from CGFs) or 'poor quality of the visual database' as technical issues that can and need be solved in a next MTDS event.

To conclude the Dutch national assessment, the results indicate that negative points were mostly simulator and network related - which may not be too surprising given the fact that the Netherlands participated with legacy flight simulators designed for procedural training and the limited time available in EFW for technical testing. Despite these issues, actual training value was achieved for the Netherlands during EFW, on both team and collective level. This clearly demonstrates the training potential of MTDS as well as the need for improving the Dutch flight simulators for this type of training.

RNLAF's Conclusion: MTDS needs to be incorporated into future training!

The RNLAF wishes to add that participating in EFW not only clearly demonstrated the training potential but also the maturity and technical feasibility of the MTDS concept. Given the first time nature of the event, it was not surprising to find certain preconditions insufficiently met and there certainly is room for improvement. However, also shown by the positive assessment results, the proof of the pudding is in the eating- actual doing is best. Despite all the technical deficiencies, the Dutch pilots, participating in EFW without prior MTDS experience, did experience (already) its training value and potential. To the RNLAF it is clear that it needs to go forward with MTDS to answer the increasing challenges faced to give its pilots sufficient high quality training necessary for successful operational deployments.

8.0 EVALUATION OF THE ASSESSMENT APPROACH

Overall, the results of the assessment show that EFW clearly demonstrated the training potential of MTDS. The operators appreciated the training on the collective level, especially mission experiences that are less sensitive to simulator and network deficiencies. The mission experiences that are more dependent on simulator and network quality were trained less than the pilots expected beforehand. Therefore, simulator and network improvements are necessary for next MTDS event. Despite the technical limitations, actual training value was reported by operational participants, and all indicated that MTDS has high training potential. This section focuses on the validity ("Does the approach truly assess training value?") and usability (E.g. "Are the surveys understandable and efficient?") of the assessment approach and its particular instruments underlying the conclusion.

Validity & Reliability. Both validity and usability issues were not a research topic in itself during EFW. However, the use of additional assessment approaches (Netherlands national approach, EFW Hot Wash, After Action Review, and the WFLO interviews) with partly different objectives and approach, do lead to similar results. As such there is support that the overall conclusions are valid. The reliability of the data may be limited by the small number of participants in EFW, leading to a small set of data available to be used in the analysis. Therefore, not all results on specific issues may be reliable, but on a more abstract level (taking all items together, leading to an overall effect and conclusions), the results will be more stable.

Method / Approach. Assessing the training effectiveness in MTDS events is a very challenging and complex task, with many variables involved. It was adequately carried out by assessment proctors at most sites during EFW. While the MEC-basis for the surveys was an excellent start for the development of an international assessment framework, it proved still difficult and time-consuming because countries had diverging objectives for EFW. A strong point of the assessment framework was the role specific subjective data collection played, based upon MECs, prior to, during, and after the exercise. Subjective performance assessment by human experts (e.g. WFLOs, SMEs and Exercise Management Directors) remains, however, absolutely necessary in order to establish a core understanding of the training effectiveness of MTDS.

Objective assessment (PETS). The potential of using automated performance measurement tools, as demonstrated by PETS, is very high. This is shown by the fact that, despite a number of network and exercise issues, a fair amount of data was successfully captured. However, for the participating entities (both simulators and CGF's) strict adherence to data-protocols (DIS/HLA) is required and substantial work in this area remains to be done.

Subjective individual assessment (surveys). For survey administration, availability of both digital ('mini-browser') and paper versions of the surveys were convenient and allowed daily data entry. Although substantial effort was made in advance to review the surveys to eliminate potential language issues for non-native English speakers, a few occurrences of confusing language were found during the exercise week. For this reason, and to make data collection go more smoothly, it was found that observers speaking the 'local language' was a huge benefit. The daily workload for sites with numerous respondents and few assessors was heavy.

Subjective group assessment. Group sessions and a small number of interviews are obviously sensitive to biases, and should not be employed as the single instrument for assessment. However, when complementing objective and survey data, the subjective data from interviews and group sessions can be very valuable and provide quicker and more in-depth insight to the data gathered. For the Dutch national assessment this approach proved useful for EFW and was reviewed positively by the participants.

Overall, the assessment approach produced usable data from which stable, high-level conclusions can be derived. When considering detailed issues, the method may require a larger group of participants. The costs (effort & time) of using the approach are considerable, both for gathering as well as analysing the data. However the results of the evaluation of the assessment methodology show that overall the effort led to valid data, collected for the first time in this way in a first time exercise of this kind.

9.0 RECOMMENDATIONS

EFW investigated the first-ever comprehensive approach to distributed assessment that resulted in an extensive set of assessment instruments that can be further adapted and reused in future MTDS events. It is important to use the same instruments over a series of MTDS events to make comparisons between these exercises possible. Multiple implementations of the same set of instruments will allow for instrument improvement and stronger and more detailed conclusions to be drawn. The assessment approach may be further enhanced to accommodate all multinational objectives. Further, it is absolutely essential to understand the effects and challenges of working from dispersed sites during the entire exercise. Therefore, it is important that future MTDS exercises build on the lessons learned in EFW.

Recommendations for further assessment development include that part of the subjective data collection process should be adapted to deal with the specific needs of future MTDS events. An effort should be made to refrain from using too extensive data collection methods. The approach to assess training value should focus more strongly on the delivery of useful outcomes and, especially for a new training tool such as MTDS, to establish lessons learned and improve each training event. A further recommendation is to increase the support to the human experts, working together in a distributed setting, with e.g. appropriate observer/assessment training prior to exercises and (software) tools to monitor and assess the training effectiveness. The EFW concept of using national WFLOs (senior military experts) at the exercise control centre and at the distributed sites proved essential for training assessment and is highly recommended for future events. Finally, given the potential and advantage of using software like PETS to collect objective performance measurements it is recommended to investigate the use of such a tool further in future MTDS exercises.

It is believed that applying these recommendations will lead to an effective and efficient assessment framework, which not only provides the military operators with useful and relevant training feedback, but is a means as well for all participants to ensure that the very promising training potential of MTDS can indeed be achieved.

10.0 REFERENCE SECTION

[1] Colegrove, C. M., & Alliger, G. M. (2002). Mission essential competencies: Defining combat readiness in a novel way. Paper presented at: *NATO Research & Technology Organization, Studies, Analysis, and Simulation Panel, Conference on Mission Training via Distributed Simulation (SAS-38)*, Brussels, Belgium.

[2] Schreiber, B.T., Watz, E.A, Bennett, W., Jr., & Portrey, A. (2003). Development of a Distributed Mission Training Automated Performance Tracking System. In Proceedings of the 12th Conference on Behavior Representation in Modeling and Simulation. Scottsdale, AZ.

[3] Schreiber, Stock, & Bennett, in preparation (Vol II)