Measures of Effectiveness and Performance in Tactical Combat Modeling

Dr. P. Dobias
LFORT

Dr. K. Sprague
LFORT

Mr. G. Woodill
LFORT

Mr. Patrick Cleophas
TNO (The Netherlands)

Mr. Wouter Noordkamp
TNO (The Netherlands)

DRDC CORA TM 2008-032
October 2008

Defence R&D Canada
Centre for Operational Research and Analysis

Land Forces Operational Research Team
The Netherlands R&T Organization

National Defence Défense nationale

Canada

Copy No:______

Measures of Effectiveness and Performance in Tactical Combat Modeling

Dr. P. Dobias
LFORT

Dr. K.B. Sprague
LFORT

Mr. G. Woodill
LFORT

Mr. Mr. Patrick Cleophas
TNO

Mr. Wouter Noordkamp
TNO

Distribution of this publication is *Unlimited*

DRDC CORA

Technical Memorandum

DRDC CORA TM 2008-032

2008-10-15

Authors

Original signed by

Dr. Peter Dobias

Approved by

Original signed by

Dr. Dean Haslip

Section Head, Land and Operational Commands

Approved for release by

Original signed by

Dr. Roy Mitchell

Acting Chief Scientist DRDC CORA

The information contained herein has been derived and determined through best practice and adherence to the highest levels of ethical, scientific and engineering investigative principles. The reported results, their interpretation, and any opinions expressed therein, remain those of the authors and do not represent, or otherwise reflect, any official opinion or position of DND or the Government of Canada.

Abstract

Computer simulations are often employed by operational research analysts to evaluate the relative effectiveness of various combinations of military equipment and tactics (i.e., options) for specific tasks within a conflict scenario. If the simulation environment is realistic enough, one can rank the options based on how effective they are when used to complete the assigned objective. The ranking process requires that measures of effectiveness (MOEs) be designed to capture the essence of how well the goal was achieved for any particular option. In this paper, it is shown that the relative ranking of options can be disturbed by omitting options or adding options, dependent on the method used for valuing the MOEs. This has implications for those relying on ranked options as part of a larger decision making process – the omission of one option due to, say, post-analysis logistical, political, budgetary or supply concerns can upset the balance of the remaining rankings and lead to an inappropriate decision if left unchecked. We discuss some circumstances under which rank-order switching can occur. Two methods of valuing MOEs aggregated through weighted sums to produce option rankings are compared and contrasted: 1) a simple *Relative to Best* scheme, and 2) *Valuing with Objective Scales*. The latter is shown to be a better choice when rank-order switching is at issue. Furthermore, it is argued that, in general, only a few MOEs are necessary and that too many can lead to undesirable consequences. Moreover, measures of performance (MOPs) are put forward to capture secondary characteristics of the options that may come into play. Although they do not explicitly enter into option ranking, flagging potential problems early on can help identify options that might be eliminated post-evaluation.

Résumé

Les spécialistes de la recherche opérationnelle ont souvent recours aux simulations par ordinateur pour évaluer l’efficacité relative de différentes combinaisons de tactiques et d’équipements militaires (les options) dans la réalisation de certaines tâches d’un scénario de conflit. Si l’environnement de simulation est suffisamment réaliste, les options peuvent être classées en fonction de leur efficacité dans la réalisation d’un objectif donné. Ce processus de classement suppose l’utilisation de critères d’efficacité conçus pour saisir l’essence de l’intérêt d’une option donnée, c’est-à-dire la mesure dans laquelle elle permet d’atteindre le but recherché. Les auteurs de ce document montrent que le classement relatif des options peut être perturbé si des options sont éliminées ou ajoutées, compte tenu de la méthode appliquée à l’évaluation des critères d’efficacité. Cela a des conséquences pour ceux qui font appel à des classements d’options pour prendre des décisions – l’élimination d’une option, après l’analyse, pour des raisons d’ordre logistique, politique, budgétaire ou matériel risque de perturber l’ordre des options restantes et d’entraîner une mauvaise décision si rien n’est fait. Les auteurs examinent certaines des circonstances dans lesquelles l’ordre de classement peut changer. Deux méthodes d’évaluation de critères d’efficacité réunis sous la forme de sommes pondérées pour donner des classements d’options sont comparées : 1) l’évaluation relative et 2) l’évaluation avec des échelles objectives. Il appert que la seconde constitue un meilleur choix quand le classement peut changer. Les auteurs soutiennent en outre que, en général, quelques critères d’efficacité seulement suffisent et que l’utilisation de trop nombreux critères d’efficacité peut avoir des conséquences indésirables. Par ailleurs, des critères de rendement sont utilisés pour saisir des caractéristiques secondaires des options susceptibles d’entrer en jeu. Même si les critères de rendement n’interviennent pas explicitement dans le classement des options, ils peuvent faire ressortir rapidement des problèmes éventuels et aider à reconnaître des options susceptibles d’être éliminées après l’évaluation.

Executive Summary

Simulated combat environments (i.e., wargames) have proven to be an invaluable tool for operational research (OR) in the past. Wargames typically depend on a wide range of physical parameters and incorporate mathematical models to best represent various key aspects of reality while abstracting or ignoring other, less relevant, details. When military equipment or tactics are in question, wargames can function as a decision aid. Different combinations (options) can be tested through simulation and the results communicated to the decision-maker. Exactly what results are provided and how they are communicated is driven by the military sponsors' key goals and/or desired capabilities, and is constrained by the limits of the simulation environment, the timeframe, the capabilities of the analysis team, and the availability of military personnel to provide expertise. Often the military sponsor prefers that the various options be ranked according to some set of criteria that takes into account the primary factors of interest and also their relative importance. In the literature, arriving at a decision in this manner falls under the general category of Multiple Criteria Decision Analysis (MCDA). For combat simulations, the most relevant factors for evaluating an option normally can be separated into two main classes: Measures of Effectiveness (MOEs) and Measures of Performance (MOPs). MOEs are strongly aligned with the desired outcomes of the mission and often include factors such as mission success, expected number of casualties, time to complete the mission, and/or successful occupation of an area. They enter directly into the option ranking scheme. On the other hand, MOPs are geared to efficiency and performance at the system level and are often characterized by measures such as weapon lethality (the amount of ammunition expended per kill), range of engagement, and the quality of sensor information. Although MOPs may provide valuable data about the simulations, they do not directly contribute to measuring the goal and thus need not (and should not) enter into the ranking scheme. Otherwise, they may obscure the analysis and interpretation of the scenarios and dampen more relevant decision elements.

The aim of this study is to provide a preliminary set of guidelines and best practices for option evaluation and option ranking in combat simulations, so as to present results to military sponsors in a form that is both transparent in meaning and robust under common usage. In particular, establishing how much one option outperforms another was investigated, in addition to how the relative rankings of options might be affected by the removal of one or more options (say, through post-study elimination via budgetary constraints) – the latter can cause an overall shuffle in the ordering of options in a not-so-obvious way. Towards this goal, two methods of valuing MOEs to produce option rankings were compared and contrasted under conditions where the relative ordering was sensitive to the omission (or addition) of an option. The methods investigated were: 1) a simple *Relative to Best* scheme, and 2) *Valuing with Objective Scales*. Furthermore, the separate and distinct roles of MOEs and MOPs towards decision-making in this context were examined and their combined interpretive value was assessed.

For the *Relative to Best* scheme (method 1, above) it was shown that although it is possible under this method to report how much one option is better than another, its meaning is diminished by the fact that the relative order of options is sensitive to changes in rankings if one or more options are eliminated for some reason in the post-analysis stage of a project, especially if there are large differences in one or more MOEs.

Consistent with past experience, it was demonstrated that by using predefined common scales for different MOEs (or *Objective Scales* as per method 2, above), it is possible to develop a scoring system that allows for the determination of how much one option is better than another. This scoring system incorporates both the results of individual MOEs as well as their relative weights. Furthermore, it is obvious that under *Objective Scales* the relative rankings of options are insensitive to the deletion or insertion of one or more options (although absolute ranking can change, of course).

In addition, analysis showed that there is a high degree of consistency between MARCUS (the Land Force Operational Research Team's (LFORT) traditional MCDA tool) and weighted sums applied to the *Objective Scales* method for the cases considered. While the weighted sums methodology outlined lacks the mathematical rigour of MARCUS, it was demonstrated to be suitable for practical problems that are likely to be encountered during the analysis of combat model outcomes. Both methods are conducive to sensitivity analysis, allowing one to test the robustness of the ranking under variations in the largely subjective weights assigned to decision-making criteria. Note that this scoring methodology is not intended to replace current MCDA practices; rather, it is intended as a supplement to the information available to the decision-maker. In the event that the two methods produce different rank orders and/or the aggregated effectiveness of top options make them nearly indistinguishable (e.g., a tie), MOPs may provide the additional insight needed to tip the scales towards choosing one option over another.

Furthermore, it was noted that a proper definition of measures, as well as proper distinction between MOEs and MOPs, are important aspects of a well-balanced statistical analysis of the system measurement results, supporting informed interpretations and decision-making. A number of *primary* MOEs equal to the (effectiveness) degrees of freedom in the scenario is most desirable. Loosely stated, these comprise the minimal set of factors that contribute to achieving the sought goal. For example, if varying a certain aspect of the simulation does not make a significant difference to the outcome from the viewpoint of satisfying the objective, then it should not enter into the option-ranking scheme. Moreover, MOEs should be defined as independent from one another as is possible in order to facilitate a simple and clear analysis and interpretation. Often, however, it is not possible or overly burdensome to devise strictly independent measures. As a rule of thumb, fewer distinctive measures are better, but not too few. In most instances, three to four MOEs should provide enough of a basis for the option rankings. MOPs, on the other hand, characterize system performance or efficiency, and as such they should be considered separately. Since they normally do not enter into the rankings directly, there is no general need to restrict the number of MOPs made available to sponsors for a given scenario. Together, MOEs and MOPs can provide an estimate of option sustainability, which is an important aspect of modern combat systems.

Based on the findings of this study, the following recommendations for evaluating the results of combat simulations are being made:

- A scoring system can be used to supplement or even replace the traditional MCDA methods employed to determine which option is best and by how much;
- When considering sensitivity to option rank-order switching under a post-analysis elimination of options, the *Objective Scales* method is a better choice for valuing (or rescaling) MOEs than the *Relative to Best* method.

- The number of MOEs used to rank options should be reasonable (3 to 4 MOEs should be sufficient for most studies) and should not exceed the number of degrees of freedom;
- Care should be taken to ensure independence of the MOEs, if possible;
- Proper distinction between MOEs and MOPs should be maintained. MOPs can be used in the final decision-making, but not to rank the effectiveness of options;
- Particular attention needs to be given to the definition of mission success since it is typically the weightiest measure. It needs to be well aligned with the stated mission objectives;
- MOPs can provide insights into sustainability of particular options when logistic considerations are a potential issue;
- Sensitivity analysis should be performed whenever possible. At a minimum it should consist of varying the weights assigned to the individual MOEs, but it can include varying the MCDA method as well; and
- A separation 'distance' between options that measures how much one option is better than another at an intuitively appropriate scale should be provided, or, if fitting, the options should be presented as equivalent in rank.

Dobias, P., Sprague, K., Woodill, G., Cleophas, P., Noordkamp, W., 2008. Measures of Effectiveness and Performance in Tactical Combat Modeling, DRDC CORA TM 2008-032

Sommaire

Les simulations d’environnements tactiques (les jeux de guerre) ont montré qu’elles étaient un précieux outil de recherche opérationnelle. Les jeux de guerre font généralement appel à un large éventail de paramètres matériels et à des modèles mathématiques pour représenter au mieux des aspects clés de la réalité en laissant de côté des détails moins pertinents. Les jeux de guerre peuvent aider à prendre des décisions au sujet des tactiques ou des équipements militaires. Diverses combinaisons (options) peuvent être mises à l’essai par simulation, les résultats étant communiqués aux décideurs. La nature exacte des résultats présentés et la façon de les communiquer dépendent des principaux buts des clients militaires ou des capacités qu’ils recherchent, et elles sont tributaires des limites de l’environnement de simulation, des délais, des capacités de l’équipe d’analyse et de la disponibilité de militaires pouvant mettre leur expertise au service des analystes. Il arrive souvent que des clients militaires préfèrent que les options soient classées en fonction d’un ensemble de critères donné qui permet de tenir compte des principaux facteurs d’intérêt et de leur importance relative. Dans les milieux spécialisés, on range cette façon de parvenir à une décision dans la catégorie générale de l’analyse décisionnelle à critères multiples. Dans les simulations tactiques, les facteurs qui présentent le plus d’intérêt dans l’évaluation d’une option peuvent habituellement être rangés en deux grandes catégories : les critères d’efficacité et les critères de rendement. Les critères d’efficacité sont étroitement liés aux résultats escomptés de la mission et ils comprennent souvent des facteurs comme le succès de la mission, le nombre prévu de pertes, la durée de la mission ou l’occupation d’un secteur. Ils interviennent directement dans le classement des options. Les critères de rendement, par contre, concernent l’efficience et le rendement au niveau du système et ils sont souvent caractérisés par des facteurs comme la létalité des armes (la quantité de munitions utilisées pour obtenir un tir au but), la portée d’engagement et la qualité des renseignements fournis par les capteurs. Même si les critères de rendement peuvent fournir de précieuses données au sujet des simulations, ils n’interviennent pas directement dans l’évaluation de l’objectif et ils ne devraient par conséquent pas entrer dans le classement. Autrement, ils risquent de brouiller l’analyse et l’interprétation des scénarios et de masquer des éléments de décision plus pertinents.

Cette étude vise à proposer un ensemble préliminaire de règles et de pratiques d’évaluation et de classement d’options dans des simulations tactiques, de façon que les résultats présentés à des clients militaires soient à la fois clairs et fiables. Les auteurs se sont intéressés en particulier à la façon de déterminer dans quelle mesure une option en surclasse une autre; ils ont également cherché à voir en quoi le classement relatif des options peut se ressentir de l’élimination, après l’analyse, d’une ou de plusieurs options (pour des raisons budgétaires, par exemple), ce phénomène pouvant entraîner une réorganisation inattendue du classement général des options. À cette fin, ils ont comparé deux méthodes qui sont appliquées à l’évaluation de critères d’efficacité dans l’établissement de classements d’options, dans des conditions où le classement relatif était sensible à l’élimination ou à l’ajout d’une option. Les méthodes étudiées étaient 1) l’*évaluation relative* et 2) l’*évaluation avec des échelles objectives*. Ils ont en outre étudié le rôle distinct des critères d’efficacité et des critères de rendement du point de vue de la prise de décisions dans ce contexte et évalué leur valeur interprétative combinée.

Pour ce qui est de l’*évaluation relative* (méthode 1 ci-dessus), l’étude a montré que, même si cette méthode permet de voir dans quelle mesure une option est préférable à une autre, son

intérêt souffre de ce que le classement relatif des options est susceptible de changer si une ou plusieurs options sont éliminées, pour une raison ou une autre, après l’analyse, particulièrement si un ou plusieurs critères d’efficacité présentent des différences marquées.

Comme d’autres avant eux, les auteurs ont montré que l’application d’échelles communes prédéfinies (ou *échelles objectives*) à divers critères d’efficacité (méthode 2 ci-dessus) peut conduire à un système de classement servant à déterminer dans quelle mesure une option est préférable à une autre. Ce système permet de tenir compte à la fois des critères d’efficacité pris un à un et de leur poids relatif. Il est en outre évident que, dans l’évaluation avec des *échelles objectives*, le classement relatif des options n’est pas sensible à l’élimination ou à l’ajout d’une ou de plusieurs options (même si le classement absolu peut évidemment changer).

L’analyse a également montré qu’il y a une grande cohérence entre les résultats obtenus avec le MARCUS (l’outil d’analyse décisionnelle à critères multiples dont se sert habituellement l’équipe de recherche opérationnelle de la Force terrestre) et les sommes pondérées qui ont été appliquées à l’évaluation des cas considérés avec des *échelles objectives*. Même si la méthode des sommes pondérées n’a pas la rigueur mathématique du MARCUS, elle convient bien aux problèmes pratiques susceptibles de se présenter dans l’analyse des résultats obtenus avec des modèles de simulation de combats. Les deux méthodes se prêtent à des analyses de sensibilité, ce qui permet de vérifier la fiabilité du classement compte tenu des variations des poids largement subjectifs qui sont attribués aux critères de décision. Il convient de souligner que cette méthode ne vise pas à se substituer à l’analyse décisionnelle à critères multiples, mais bien à mettre d’autres renseignements à la disposition des décideurs. Si les deux méthodes produisent des classements différents ou que l’efficacité globale des meilleures options fait qu’il est à peu près impossible de les départager, les critères de rendement pourraient fournir des renseignements complémentaires susceptibles de faire pencher la balance en faveur d’une option.

Les auteurs ont noté en outre qu’il est important de bien définir les critères et d’établir une nette distinction entre les critères d’efficacité et les critères de rendement si l’on veut arriver à une analyse statistique équilibrée des résultats qui débouche sur des interprétations et des décisions éclairées. Le mieux est d’utiliser un nombre de critères d’efficacité *primaires* égal à celui des degrés de liberté du scénario. En gros, ces critères devraient correspondre aux facteurs indispensables à la réalisation de l’objectif souhaité. Par exemple, si le fait de modifier un aspect donné de la simulation ne change pas sensiblement le résultat du point de vue de la réalisation de l’objectif, l’aspect en question ne devrait pas entrer dans le classement des options. De plus, les critères d’efficacité devraient autant que possible être définis indépendamment les uns des autres, de façon à faciliter et à éclairer l’analyse et l’interprétation des résultats. Il arrive souvent, toutefois, que la définition de critères parfaitement indépendants se révèle impossible ou trop complexe. En règle générale, il vaut mieux utiliser un petit nombre de critères distincts, mais pas trop peu. Dans la plupart des cas, trois ou quatre critères d’efficacité devraient fournir une bonne base de classement des options. Les critères de rendement, de leur côté, servent à caractériser le rendement ou l’efficience d’un système et ils devraient donc être considérés séparément. Comme ils n’interviennent normalement pas directement dans le classement, il n’est pas nécessaire d’en restreindre le nombre dans un scénario donné. Ensemble, les critères d’efficacité et les critères de rendement peuvent donner une idée de la soutenabilité d’une option, ce qui est important dans les systèmes de combat modernes.

S’appuyant sur les résultats de l’étude, les auteurs font les recommandations suivantes au sujet de l’évaluation des résultats de simulations tactiques :

- Un système de classement peut compléter et même remplacer les méthodes classiques d’analyse décisionnelle à critères multiples qui servent à déterminer quelle option est la meilleure et dans quelle mesure.
- Au moment d’étudier les effets de l’élimination, après l’analyse, de certaines options sur leur classement, la méthode des *échelles objectives* constitue un meilleur choix pour évaluer (ou reclasser) les critères d’efficacité que l’*évaluation relative*.
- Le nombre de critères d’efficacité utilisés dans le classement des options devrait être raisonnable (trois ou quatre devraient suffire dans la majorité des cas); il ne devrait pas excéder le nombre de degrés de liberté.
- L’indépendance des critères d’efficacité devrait être préservée autant que possible.
- Les critères d’efficacité devraient être distincts des critères de rendement. Les critères de rendement peuvent servir dans la prise de décision finale, mais pas dans le classement des options en fonction de leur efficacité.
- Une attention particulière doit être accordée à la définition du succès de la mission, puisque c’est habituellement le facteur auquel on attribue le poids le plus important. Il importe de bien l’aligner sur les objectifs de la mission.
- Les critères de rendement peuvent fournir des renseignements sur la soutenabilité de telle ou telle option quand des considérations d’ordre logistique entrent en jeu.
- Une analyse de sensibilité devrait être faite chaque fois que cela est possible. On devrait au moins s’assurer de varier les poids attribués à chacun des critères d’efficacité, mais on pourra aussi recourir à une autre méthode d’analyse décisionnelle à critères multiples.
- Une « distance » entre les options qui permette de voir dans quelle mesure une option est préférable à une autre, selon une échelle intuitive, devrait être prévue; sinon, les options comparables devraient se voir attribuer le même rang, si cela convient.

Dobias, P., Sprague, K., Woodill, G., Cleophas, P., Noordkamp, W., 2008. Measures of Effectiveness and Performance in Tactical Combat Modeling, DRDC CORA TM 2008-032

Table of contents

List of figures

List of tables

1. INTRODUCTION

1.1. Background

Simulated combat environments (i.e., wargames) are routinely employed by operational research analysts to assess the operational effectiveness of equipment, tactics and force composition on behalf of military sponsors. Wargames typically depend on a wide range of physical parameters and incorporate mathematical models to represent various aspects of reality. The conditions (physical environment), circumstances (objectives, obstacles) and automata capabilities (equipment, knowledge of tactics, decision-making) within a wargame are prescribed by the user. The degree of automation varies. With some, such as CAEn [1,2], interactors (human players) define the movements and actions of each automaton (or *entity*) on the battlefield and make all tactical decisions. In others, such as MANA [3], EINSTein [4] and HiLOCA [5], rules for movement, interactions and use of capabilities are prescribed before the simulation begins. The simulation then runs its course without human intervention. It is even possible to automatically generate an optimal set of rules to best suit the conditions of a scenario [6,7]. In any event, entities within a combat simulation move and interact with one another and their environment according to their instructions. Generally, the step-by-step movements and actions are recorded over time in digital files for later analyses.

Once the simulation environment for a given scenario has been set up, variations in equipment, tactics, and/or entities that enter into the simulation comprise the various *options* targeted for consideration. At the onset of a study, generally it is not known which of the possible options is the best overall. The goal of the simulations is to provide data that allow one to compare and rank the available options within an otherwise consistent environment, often to support important decisions that involves the commitment of limited resources and potentially puts lives at risk. The simulations themselves are typically evaluated by employing a set of predefined measures of effectiveness (MOEs) and/or performance (MOPs). The values of the individual measures for a given option are computed and then combined in a consistent, predetermined manner to provide an overall assessment of the option for comparison with other options. This kind of process is known as multi-criteria decision analysis (MCDA).

Currently, for land force combat simulations anywhere up to ten or more MOEs and MOPs are used to analyze the outcomes. These measures are assigned weights based on their importance and relevance to the sponsor, and are used to obtain an overall ranking for each option. Normally only a few measures account for the majority of the decision weight (past experience suggests between 70-90%), while the remainder contribute very little to the final outcome. This has been described as *measure overload*, raising concerns that too many measures can obscure simple relationships derivable from fewer, more important measures. On the other hand, by ignoring important aspects of the system or grouping them inappropriately, having too few measures certainly has the potential to mislead decision-makers. Thus a balance must be struck that is dependent upon the inherent complexity of the problem at hand and the degrees of freedom that are available for exploitation. Having said that, what really matters, in the end, is the *quality* of the final aggregate measure used to rank the options, rather than the number of elements that went into computing it, their independence, or whether they should have been classified as MOEs or MOPs. There are no set rules. Nevertheless, one can suggest guidelines or ‘rules of thumb’ that, if followed, help to ensure that the quality of the measures is as high as possible under the circumstances, and that the options are ordered and compared in a meaningful way.

Standard procedure within the Land Forces Operational Research Team (LFORT)(Canada) is to rank options based on observed statistically significant differences for all of the MOEs. These ranks are then used in conjunction with the assigned weights to obtain overall rankings for the analyzed options.

For example, in a recent study upward of ten measures were used to rank individual options. While one of the options far outperformed all the rest on the mission success, this large difference was eliminated by reducing the consideration to the ranks of options for individual measures. So while the leading option was much better than the rest on key criteria, the final rankings did not show this. For some scenarios it actually ranked lower than another option due to a weaker performance on one of the less important measures. This clearly revealed the need to introduce another MCDA method which would capture the actual performance on individual measures.

Furthermore, in another study most of the measures including mission success were heavily dependent on the attrition which led to a bias in favour of the options that were strong on eliminating the opposing force although it was not the objective of the mission. This highlighted the need for proper selection of measures more in line with the actual mission objectives.

To summarize, the currently used MCDA method suffers from several setbacks:

- The method does not provide an answer to the question “how much is one option better than another?” This is often a very relevant question, since there are normally factors that are not considered in the simulation, but which must be considered by the sponsor in making the final decision (such as the cost of individual options);
- In some instances the difference between options can be statistically significant, yet the difference may not be large enough to be practically relevant;
- Often, there are too many measures used in option ranking. Consequently, the low-weight (and thus relatively low importance to the sponsor) measures are analyzed to the same degree as the most critical measures. The large number of measures clutters the analysis.
- The current method has the potential to inappropriately rank options, for example in circumstances of one (A) being dramatically and significantly better than another (B) on the most critical measure (e.g., mission success), but marginally significantly worse on other measures. Since only the relative ranks are considered, it might happen that B would be ranked better than A;
- There is no delineation made between MOEs and MOPs. The measures are lumped together, ignoring the fundamental difference between characteristics of a system (MOP) and the operational effectiveness (MOE). MOPs often provide insight into why MOEs differ and can also be of interest to the sponsor for a variety of reasons, but they should not be used in ranking operational effectiveness.

Therefore, a study was initiated by LFORT, with support from Director of Land Requirements (DLR) to explore an alternative to the current methodology of assessing the outcome of combat simulations to improve the quality of answers provided to the Chief of Land Staff (CLS).

Collaboration with the Netherlands Research and Technology Organization (TNO) was initiated at a bilateral meeting held in the fall of 2007 in The Hague, NL. This provided an opportunity to compare experiences and approaches of LFORT and TNO, and provided valuable inputs into this study.

1.2. Aim

The aim of the study was to propose a revised methodology to evaluate the results of combat simulations, which would provide a means of comparing the effectiveness of two or more options that was provably more robust and meaningful than could be obtained using the current MCDA approach.

1.3. Objectives

The objectives of the study were as follows:

- Develop an MCDA methodology that will provide some insight into the degree of difference between two or more studied options, that has following characteristics:
    i. The method must not only determine which option is better, but also by 'how much';
    ii. If an option is removed post-study due to some unforeseen circumstance, the method should preserve the relative option rankings of those that remain;
- Provide guidance for recognizing and dealing with situations where the differences between two options are statistically significant but not practically relevant;
- Outline a prescription for choosing an appropriate number of measures under typically encountered circumstances capable of characterizing the most important aspects of a combat scenario; and
- Summarize the overall best practices for combat simulation option evaluation, including guidelines that reduce measure overload whilst maximizing MOE-alignment with the sponsors' wishes and making full use of the distinct information provided by MOPs.

1.4. Methodology

A two-pronged approach to the problem was used

- Best practices and options were proposed through the creation of a new combat simulation assessment framework (CSAF), based on 1) a brief literature review of the current state of MCDA , 2) the authors' experiences with combat evaluation, and 3) anticipated needs of the sponsor and affected decision-makers; and
- The methodology was tested against the results of previous studies by comparing the rankings obtained using the scoring methods to the rankings obtained using MARCUS.

2. SELECTION OF MEASURES

As was already mentioned in Section 1.1, it is very important to select proper measures (and a proper number of measures as well) to enable efficient and reliable MCDA. Too few MOEs will not provide sufficient granularity, while too many of them will clutter the analysis. Furthermore, a lack of independence between measures can lead to a bias in the analysis, and thus invalidate the results. It may also become difficult to assign weights to criteria that share common factors. For example, when evaluating a combat scenario a common MOE is the number of casualties suffered. A second MOE might be the number of shots fired by the opponent and there may be other MOEs defined for subsequent decision criteria, each with a corresponding weight (importance). Under a slowly varying kill-rate though, it is easy to see that attrition actually weighs in at least twice – once fully and then partially. So the question is: ‘What overall weight are you actually assigning to attrition if several measures depend on it?’

Inappropriate measures that do not properly reflect the mission objectives for an analyzed scenario can also invalidate the analysis, or lead to absurd results. Lastly, the measures need to capture adequately the differences in the performance of analyzed options.

It is important to distinguish between two types of measures (MOEs and MOPs) in evaluating the results of combat simulations. They can be defined as follows:

Measure of Effectiveness: A measure that characterizes the operational effectiveness of a unit or force in achieving its objectives during a mission. The measure must relate directly to the mission objectives and it must provide insight into the degree to which these objectives were satisfied.

Measure of Performance: This measures the performance of a particular system, and as such it is indirectly related to the mission objectives. It is usually related to technical properties of the analyzed systems, and should be consistent for corresponding systems across options.

It is important that both kinds of measures be considered when evaluating options in a simulation. Ineffective but highly performing systems will not achieve their objectives and eventually be discarded by BLUE or overwhelmed by RED. They are not sustainable. Effective but poorly performing systems will be useful but could be expensive to maintain, again not sustainable. The best options are those that are both effective and highly performing.

2.1. Measures of Effectiveness

As defined above, MOEs measure how effective a particular option is in achieving the mission objectives. Therefore it is important to ensure that the selection of measures reflects these objectives. For instance, if the mission objective is to destroy a particular target, it should be reflected in the MOEs. Likewise, if the objectives include minimal BLUE casualties, this should be reflected as well.

At the same time it is important to ensure a reasonable number of MOEs. Too many can lead to problems in the assessment such as dependence between MOEs, and it can also dilute the importance of individual measures. The number of MOEs directly affects the range of weights that can be assigned to them.

For practical purposes, under ‘normal’ circumstances consistent with the authors’ experience, a few (three to four) MOEs are sufficient, especially when supplemented by MOPs to describe the system characteristics that are of interest to the sponsor. The limited number of measures allows for better insight into the system's main-driver dynamics, because it enables the analyst to look into the relative effectiveness of the options. Also, if the number of measures is greater than the number of free parameters such as weapon mixes considered, the independence of the measures will be difficult or impossible to ensure. This can complicate an intrinsically simple dependency or interpretation of the results.

Good candidate MOEs for evaluating options within many combat scenarios include:

- Mission Success – evaluating the overall effectiveness in achieving the mission objectives. Since this is often the most important measure, it is discussed in detail in the next section;
- RED and BLUE casualties – this can take on a variety of forms, including straight casualty numbers, LER (ratio between RED and BLUE killed), or RCS (ratio of the final unit strength to the initial strength). Note that while RCS is casualty based it serves the purpose of estimating whether or not a unit continues to be battle ready while loss exchange ratios are an estimate of the unit's effectiveness in the just concluded action. Both of these are often of high importance to the sponsor;
- Time to accomplish mission – this MOE is valuable in situations wherein the timely accomplishment of the mission is critical; or
- Area occupation – this MOE enters into scenarios if the purpose of the game is to occupy or clear a certain area. It can also be often found as a component of mission success.
- Detections – this MOE is especially useful in scenarios where surveillance plays a significant role. It can include a variety of derivatives such as the ratio of the number of detected enemy units to their total number present in a particular area or the distance at which opposing units were detected;

The above measures are just examples of possible MOEs. Of course, the actual selection of MOEs must reflect the particular scenario or mission type that is to be modeled and what the sponsor hopes to gain by studying it.

2.2. Mission Success

Mission success is usually the primary MOE used to assess simulation results. In many simulations it is assigned 30% or more of the total weight toward the ranking of the options. Therefore particular attention must be paid to its definition.

In general, the definition of mission success should reflect the overall mission objectives. While various aspects of mission success may be difficult to realize or quantify in some instances, as often happens due to limitations imposed by the model used and its ability to record particular types of information, shortcuts made to bypass the determination of key factors should be avoided if at all possible. An improper definition of mission success can lead to biased results that do not actually achieve the intent of the study. In particular, attention should be given to avoid the tendency to simply define mission success in terms of attrition at the expense of other factors that are important for the successful completion of a mission. The more direct the connection is between mission success and the stated mission objectives, the better.

2.3. Measures of Performance

MOPs are useful to capture the performance of the analyzed systems, and can provide supplementary information to MOEs. This supplementary information offers useful input towards making the final decision about the preferred option, especially in situations where the top-ranked options are nearly indistinguishable with regard to aggregate effectiveness. However, it needs to be kept in mind that MOPs characterize system performance rather than the operational effectiveness and therefore they should not be simply used in rankings of effectiveness.

Examples of MOPs can include

- Weapon lethality (number of ammunition rounds needed to eliminate the target);
- Range of engagement;
- Ammunition expenditure; and
- Detections (in a combat scenario); etc.

Note that some of these MOPs may overlap with the selection of MOEs as noted in Section 1.1. It must be emphasized that the classification of the measure, one way or another, is scenario dependent and should be based on alignment with the mission objectives. Those strongly aligned should become MOEs while the ones more focused on characterizing performance of individual systems should become MOPs.

While not appropriate for the ranking of effectiveness for individual options, MOPs can be used to flag options that are not performing well in key performance categories. The flag could be defined as a simple threshold value signalling “caution” (e.g., more than 50% of available ammunition spent), or something more elaborate (e.g., “yellow” for 50% of ammunition expended, “orange” for 70% expended and “red” for 90% expended). In this fashion, if a particular option is flagged on certain MOPs, it could potentially provide valuable information to the client. If, for instance, a particular option is very effective, but expends 90% of its ammunition, then either the amount of ammunition, the scenario assumptions, or both might need to be re-evaluated. Performance flags could also indicate that a given option might turn out to be a logistical nightmare if actually deployed.

2.4. MOE / MOP Summary

In summary, the distinction between MOEs and MOPs is important for a proper assessment of the operational effectiveness of combat systems (weapons, sensors, etc.). MOEs are measures that characterize the operational effectiveness and should be used in rankings. MOPs provide supplementary information about system performance (related to technical parameters and tactical definitions), and allow for flagging inadequate system performance. Effectiveness plus performance together produce an estimate of an option’s sustainability. Both are important in the design of a modern combat system. Separate analysis of the two may lead to more cohesive insights into the relative merits of different options. In some circumstances a measure may sit on the borderline between MOPs and MOEs. There is no hard and fast rule in such cases, and the analyst is left to rely on his or her best judgement.

A final important point to reiterate concerning MOEs is that the number of measures used to characterize system effectiveness should not surpass the number of free parameters in the system; otherwise it is impossible to ensure the independence of measures. Dependence poses potential problems for MCDA. Such a restriction does not apply to MOPs, which are more-or-less provided for reference and flagging purposes.

3. HEURISTICS

3.1. An Overview of Multiple Criteria Decision Analysis

MCDA has been described as [8]

"*… a collection of formal approaches which seek to take explicit account of multiple criteria in helping individuals or groups explore decisions that matter*."

MCDA approaches have in common that a set of options (also called courses of action, strategies, alternatives, etc.) are evaluated based on a set of criteria (also called attributes, aspects, dimensions, etc.) to arrive at a ranking of the options. The criteria cover the relevant parts of the decision problem that need to be considered. They may each cover distinct, unrelated aspects of the domain of consideration (e.g., a decision to spend $5000 on one of a selection automobiles based only on colour and mileage) or be strongly related (e.g., decision to buy one stock from a list of stocks based on the one year trends and the advice of two financial analysts … one favouring secure investments and another that favours potentially high payoff investments). The ranking need not be of the form of a strict less-than or greater-than relationship, but may also consist of a classification of the options into ranked classes or groupings (e.g., {Good, Medium, Bad}).

An important aspect of MCDA approaches is the relative weighting of the various criteria that a decision is to be based on. Weights reflect how the decision maker values a particular criterion compared to the others. They are usually at least somewhat subjective, however methods exist to limit subjectivity to the determination of the order of importance of the criteria only (see Section 3.3, below). Those criteria of high importance receive high weight values, and those with low importance receive low values. The weights must be determined with respect to some common reference scale, in the sense that relative weightings have to compare sensibly across the various criteria. However, they need not be fixed. In general, for a given criterion, the weight can be constant, a function of the various criteria, a function of the state of the environment (past, present, or projected), or of some other exotic variety. Another aspect of weighting schemes concerns uncertainty in the decision-making preferences, which affects the accuracy of choice. If the decision maker assigns weights subjectively to, say, within 10% accuracy, then that ambiguity propagates to a kind of error term or 'fuzziness' in the rank ordering. If the error is small enough, then the rank order is not affected. But if the error is too high, two or more options may overlap in rank order such that the better of them cannot be deduced using the given criteria and weighting scheme. Note that even if errors are very small it is possible that options come so close in rank value they are statistically indistinguishable in rank order.

Weighing criteria into a decision is fine in the abstract sense. However, there are important subtleties that come into play when one considers how to compare options across different kinds of criteria, especially when there are significant differences in performance on different criteria. The technique employed determines the precise meaning of the weights and also exactly how contributions from the various decision factors filter into the ranking or rating of alternatives.

MCDA is divided into several schools of practice. Some of the main techniques are listed below. Key aspects of the various methods are then stated briefly.

- Analytic Hierarchy Process (AHP) [9]
- Multi-attribute Global Inference of Quality (MAGIQ) [10]
- Simple Multi-Attribute Rating Technique Exploiting Ranks (SMARTER) [11]

- Data Envelope Analysis (DEA) [12]
- Goal Programming (GP) [13,14]
- Dominance-based Rough Set Approach (DRSA) [15]
- ELimination Et Choix Traduisant la REalité (ELECTRE) [16]
- Preference Ranking Organisation METHod for Enrichment Evaluations (PROMETHEE) [17]
- Evidential Reasoning Approach (ERA) [18,19]
- Technique for Order Preferences By Similarity to the Ideal Solution (TOPSIS) [20]

AHP, MAGIQ and SMARTER are hierarchical methods whereby decision problems are broken down into a hierarchy of sub-problems that can be analyzed independently. These methods are mainly applied to large-scale, multi-party problems and presented as interactive group activities. DEA and GP generalize on linear programming to solve optimization problems with multiple and possibly conflicting objectives. In essence, a decision is 'located' that is as 'close' to satisfying the objectives as possible. The objectives are actually coded as linear constraints defined by inequalities in the decision variables. A penalty function weighs the objectives relative to one another. DRSA is a linguistic rule-based approach to MCDA that builds on rough set theory[1] to solve decision problems. A rough set approximates the original set, and the DRSA extension of rough set theory, namely the substitution of an indiscernibility relation[2] on sets with a dominance relation[3], allows the method to deal with inconsistencies commonly encountered in MCDA problems. Being rule-based rather than strictly numeric, it is appropriate for decisions that are best justified according to the rules that were followed/violated in making the decision. ELECTRE and PROMETHEE are referred to as 'outranking' methods. Both make use of binary comparisons of alternatives. In general, they consist of two main parts: 1) constructing a series of outranking relations aimed at comparing each pair of actions; and 2) an exploitation procedure that elaborates on the recommendations obtained in the first phase. The nature of the recommendation depends on the problem being addressed. ERA assesses options based, in particular, on the theory of evidence [19]. A belief structure (or matrix) and evidential reasoning algorithms incorporate uncertainty and randomness aspects of decision making. Both qualitative and quantitative criteria are supported. TOPSIS (sometimes referred to as TOPSYS) is a popular *ideal point* method. In this method options are ranked according to their separation from an ideal point defined as the the most desirable, weighted, hypothetical option. The separation is measured via a metric distance.

The above methods are fairly labour intensive, and the outcomes might not provide sufficient justification for the extra effort given the usual uncertainty in the outcomes caused by the stochasticity of the used models. Therefore, in what follows, we limit our focus to methods that fall under the general classification of additive multi-attribute value or utility models. The main reason for the limitation is that past studies in combat modeling by the authors have made extensive use of them, and in these studies certain issues have surfaced that are the subject of investigation later in this paper.

[1] A rough set approximates a conventional set (or *crisp set*) by utilizing two sets representing lower and upper approximations to it.

[2] An indiscernibility relation on a set is another term for an equivalence relation, that is, when describing elements of the set by a selection of attributes *P*, if the elements are indistinguishable judging only from the attributes in *P*, then they are equivalent with respect to *P* (i.e., *P*-indiscernible).

[3] For criteria *P*, a set *X* *dominates* a set *Y* (wrt *P*) if *X* is better than in *Y* on every criterion in *P*.

In multi-attribute value or utility models, the general form of the value function (the function used for ranking) for given criteria $\{x_i\}$ with corresponding weights $\{w_i\}$, $i=1..N$, is:

$$V(x_1, x_2, ..., x_n) = \sum_{i=1}^{N} w_i v(x_i) ,$$

where $v(x_i)$ denotes the value of the *i*th criterion (i.e., the scaled MOE for criterion x_i in the case of combat modeling). Typically, $0 \le w_i \le 1$ and $\Sigma_i \, w_i = 1$. The above formula corresponds to the common notion of ‘weighted sums’.

An alternative method for option ranking using valued criteria and their respective weights is referred to as the *weighted product method* [21-23]. As the name suggests, options are compared to one another by examining a product of criteria ratios, each ratio raised to the power of its respective weight. The ratio formula for two options A_K and A_L is written as:

$$R_{KL} \equiv R(\frac{A_K}{A_L}) = \prod_{i=1}^{N} \left(\frac{v(x_{K,i})}{v(x_{L,i})} \right)^{w_i} .$$

where $v(x_{Ji})$ denotes the value of the i^{th} criterion of the J^{th} option. If $R_{KL} \geq 1$ then A_K is preferred over A_L. The highest ranked option is then the one that is better (or equivalent to) all other options.

In the next section, various methods for weighting decision criteria are discussed in the context of the weighted sums technique, which is the simplest and most common technique used.

3.2. MARCUS

MARCUS, the Multi-criteria Analysis and Ranking Consensus Unified System was developed to aid decision-makers in reaching group consensus regarding the relative merit of several options evaluated against a set of criteria (i.e. the consensus ranking problem). According to the MARCUS methodology, each decision-maker ranks a set of options in order of preference (e.g. Option A is preferred over Options B and C, and Option B is preferred over Option C, etc), making no specification regarding the strength of preference. Taking the rankings of all decision-makers as input, MARCUS returns a single ranking that represents a consensus of the group. By design, MARCUS overcomes many of the problems inherent in other scoring and ranking systems, such as susceptibility to biases introduced by a malevolent voter or voter inconsistency, and satisfies the fundamental requirements of a voting system.

However, MARCUS was not devised to deal specifically with rankings of options in the context of combat modeling. Since it only considers ranks of the evaluated options for individual MOEs, it does not allow capturing the magnitude of the difference in performance between options for the individual MOEs, nor does it enable the identification of how much is one option better than the other.

3.3. Criteria Weighting Methods

3.3.1. Background

Criteria weighting methods, also called attribute weighting methods, specify the scheme used to assign numerical weights to criteria. These weights quantify the relative importance of each criterion for the decision at hand. As mentioned previously, they can simply be numbers, or they can be functions with any number of variables. The way criteria are weighted affects the final rank-order of the available options by suppressing, amplifying, or equating specific decision variables (for combat modeling, this typically means the MOEs). Some of the methods available are statistical, while others are purely subjective. A few of the more popular criteria weighting methods are described below, with the scope limited to those that do not involve functions of one or more variables. Note that there are some authors who argue that the weights do not matter in many practical applications, apart from extremes [see for example references 24,25], but most authors would disagree with the statement. In general, it suffices to say that the specific values of the weights matter more for some decisions than for others.

3.3.2. Equal Weight

The simplest of the criteria weighting methods is equal weight. It is often used as a baseline to compare and contrast performance against other methods such as those presented below. It can also be used for human decision-making in the absence of MCDA aids.[4] Furthermore, the equal weight method is of use when no information is available concerning the relative importance of the criteria (e.g., a fallback or default). As the name suggests, in this scheme all attributes are given the same weight, that is, the ranking is determined as if all factors were considered equally relevant. For N criteria, we can simply write the weights w_i as:

$$w_i = \frac{1}{N}, \text{for all } i = 1..N.$$

In references [24,25]it is suggested that the equal weight method frequently produces decisions that are at least as good as those using more complicated schemes to assign the weights. However, this assertion need not apply to any particular decision. In any event, the method holds value as a point of reference to compare how dramatically the overall ranking changes when all criteria are considered equal. Below are several examples of some of the common methodologies. However, they are presented only as examples, and the authors do not argue in favour of any of these methods.

3.3.3. Random Weight

In this case some or all of the criteria are assigned either completely random weights or partially random weights. The weights may be drawn from prescribed distribution functions appropriate to the various criteria. In the partially random case, they might be random but with a preferred directional bias representing a tendency for or against any given criterion. Random weighting schemes are useful when little is known about the weights or if they are known within certain limits that sensibly can be covered through randomization. They are also

[4] Note that random or near-random weighting of the criteria has been used in this manner as well.

used to analyze the sensitivity of a decision with respect to the criteria weights and furthermore to compare decision fidelity against other weight assignment mechanisms, especially when trying to assess if the weights even matter for the case at hand.

3.3.4. Ratio Scale

Ratio Scale methods aim to quantify the relative weight of a criterion by considering the relative importance of moving a criterion from its worst value to its best value [26]. There are many ways of arriving at ratio scale weights. What is common to all methods is that they preserve the relative scaling properties of the decision-maker's preferences [26]. Two common procedures are described below:

- Swing Weight Method [8,26,28]

  The 'swing' captures the relevance of moving a criterion from its worst value to its best value. Steps for this method are as follows:

  1. The decision-maker orders the criteria by importance in terms of their associated value ranges.
  2. Assuming that each criterion is at its worst possible value, the decision maker is asked which is the preferred criterion to move from its worse value to its best value (the 'swing'). This criterion is deemed the most important one and is given the highest weight on the chosen scale (say 100 on a scale of 1 to 100).
  3. The next most important criterion is found by following the same procedure as above for the criteria that remain, and is assigned a weight in relation to the highest (say 75), and so forth.
  4. The weights can then be divided by their total sum to normalize them so that the sum of the adjusted weights is unity. The last step is not necessary.

- Direct Tradeoffs [27,28]

  The decision is based on making tradeoffs among competing objectives. Weights are assigned by equating direct, decomposed tradeoffs between the criteria.

  1. The decision maker is presented with binary choices and asked to fill in a missing entry in one of the two choice vectors so as to make them equally attractive.
  2. This allows one to progressively narrow down the range of the decision weights.

3.3.5. Rank-order

Rank-order methods fall under the category of approximate weighting schemes and preserve only ordinal properties of the decision maker's judgements [26], that is, the order of the weights (importance) is known. The relative separation in magnitude between how important attributes are to the final decision is not captured – only the order of importance is of concern. Such methods are especially useful when more exact methods of determining weights are not feasible or the response error is high. Two common procedures are described below.

- Rank-order Centroid Method [29]

  Using the Rank-order Centroid (ROC) method, the weights $w_1 \geq w_2 \geq \ldots \geq w_N$ for N ranked criteria are assumed to be uniformly distributed on a unit simplex[5] having N-dimensional vertices $v_1 = (1,0,0,\ldots)$, $v_2 = (1/2,1/2,0,0,..)$, $v_3 = (1/3,1/3,1/3,0,0,..)$, … $v_N = (1/N,1/N,\ldots,1/N)$. The weights are selected as the coordinates of the mid-point (centroid) of these vertices, found by summing them (akin to vector summation) and dividing by N. Then w_1 is the first component of the centroid position, w_2 the second, etc. The formula for the i^{th} weight simplifies to:

  $$w_i = \frac{1}{N}\sum_{k=i}^{N}\frac{1}{k}, \quad i = 1..N.$$

  The weights sum to unity and $0 \leq w_I \leq 1$. The formula represents the expected value of the 'true' weights under the assumption of uniformity (above), which is not overly unreasonable in the absence of any quantitative constraints about decision-maker preferences. As an example, if we are given three criteria with unknown weights $w_1 \geq w_2 \geq w_3$, the computed values are $w_1 = 11/18$, $w_2 = 5/18$ and $w_3 = 1/9$. Note that $w_1 + w_2 + w_3 = 1$.

- Rank-sum Method [30]

  Again, as with ROC, in the Rank-sum method (RS) attributes are ranked according to their relative importance towards the decision at hand. The most important criterion is assigned a weight of N/(sum-of-ranks) and the least important criterion 1/(sum-of-ranks) [24]. The formula for the i^{th} weight, corresponding to the criterion of i^{th} importance, is thus given by

  $$w_i = \frac{N+1-i}{\sum_{k=1}^{N} k} = \frac{2(N+1-i)}{N(N+1)}, \quad i = 1..N.$$

  These weights also sum to unity and $0 \leq w_i \leq 1$. Repeating the previous example for the ROC method, if we are given three criteria with unknown weights $w_1 \geq w_2 \geq w_3$, the computed values are $w_1=1/2$, $w_2=1/3$ and $w_3=1/6$. In general, RS weights are more flatly distributed than ROC weights [13].

3.3.6. Subjective Weights

Another common approach is the subjective assignment of weights. In this instance the decision-maker (often guided by the analyst) assigns weights to the individual criteria on the

[5] An m-simplex is the m-dimensional analogue of a triangle. For example, a 0-simplex is a point, a 1-simplex is a line, a 2-simplex a triangle, and a 3-simplex is a tetrahedron. N>0 vertices define an (N-1)-simplex.

basis of his or her personal preferences (or feelings), without any regard for any particular weighing method.

In this particular instance, the sensitivity analysis is vital. While it is recommended in general, since in the case of the subjective weights there is no particular objective foundation for the weights, the sensitivity analysis becomes a must. It is recommended that some of the above mentioned methods (or their combination – for instance equal weights and Swing weight method) be used to introduce more rigour into the rankings.

3.4. Valuing Criteria

Determining whether one option is superior to another requires that a consistent method for scoring the options be defined. In a multi-criteria decision problem, this in turn necessitates the development of some kind of aggregation scheme to combine the separate scores assigned to the various criteria (i.e., MOEs) used to evaluate the option. The result is a single, all-encompassing score that, to the analyst's knowledge and capabilities, best represents how well the option fared relative to the weighted criteria. One such aggregation scheme is the weighted sums method (see Section 3.1, above). However, for this scheme to make sense, as mentioned above, the raw MOE values computed for individual options need to be standardized to a common scale[6]. Otherwise, since the weights are dimensionless, the aggregation would involve summing quantities having different units and/or scales. Converting to a common scale can be as simple as rating each attribute on a scale from 1 to 5: 1 being terrible, 2 poor, 3 moderate, 4 good and 5 exceptional. There are no fixed rules for determining the rescaling mechanism. A reasonable expectation is that it should follow whatever best represents the intentions of the decision maker. For instance, if a score of 'terrible' for a particular criterion was completely unacceptable and at least some options were expected not to be terrible for the same criterion, then it might make sense to use a slightly altered scale for that attribute instead. For instance, assign -1000 (or another large negative value) to whatever constitutes 'terrible' for that criterion only and keep the remainder the same as before. That way, the targeted options are destined to fall to the lowest levels of the option ranking scheme as long as the chosen weighting scheme is not so unbalanced as to counter the intended effect (in the latter case the magnitude of the negative value can be further increased).

The example in the previous paragraph highlights a crucial aspect of multi-criteria decision problems – they are not generally well defined [21]. The reason is that there is no universally valid way to quantify all pertinent data if it is derived from measures having different units. Nonetheless, such an obstacle does not prohibit decision makers from doing exactly that which is most difficult to define. For instance, in economic analyses or damages lawsuits, 'equivalent' dollar figures for a wide range of quantities that are beyond price can factor into a decision process (e.g., mental suffering).

At this juncture, it seems clear that the rescaling method that is selected can have a large effect on the ranking of options. Furthermore it is usually arbitrarily defined. This provides one of the motivating factors for Section 4, wherein guidelines for MOE rescaling applicable to the evaluation of a wide range of options pertaining to combat modeling are explored.

[6] Note that this does not apply to all methods of aggregation. For example, the *weighted product method* mentioned in Section 2.1 is a form of dimensionless analysis for comparing and ranking options [20-22].

3.5. Additional Comments on MCDA

MCDA techniques have been applied to a wide variety of decision making problems [8]. Crucial aspects contributing to the success of a technique seem to be: 1) deciding on the right set of criteria, 2) valuing the criteria on comparable scales, and 3) weighing them appropriately for the decision at hand. Yet another is deciding on a suitable MCDA method to follow for the given decision problem. Generally speaking, the chosen method should make optimal use of the available information and produce good results in all testable and foreseeable circumstances. When aiding a decision to be made by humans, the process should also be transparent and easy to understand by the decision-maker(s) – a decision-aid that is simply a 'black box', void to the viewer of context and the intricate balancing and mutability of decision factors, is not likely to help convince anybody of anything.

Different MCDA methods can lead to different rankings of alternatives. Thus one not only has to decide on the best option, but also *choose how to decide* on the best option. This can further confound decision-making, especially if the decision maker is presented with varying sets of ranked options for the same problem. Thus it is instructive, where possible, to settle on one particular technique for reporting option ranking results for a given decision problem or similar family of problems. Other methods may enter in as part of a sensitivity analysis of the rank-order, and discrepancies must be dealt with in that light.

In the section that follows, MCDA and criteria valuing schemes are discussed in the context of evaluating and ranking options through the use of MOEs in combat modeling and simulation. Suggestions concerning appropriate techniques applicable to the majority of practical scenarios encountered in this regime, as per the authors' experiences, are provided.

4. MULTI-CRITERIA DECISION ANALYSIS IN COMBAT MODELING

By definition, the question of determining which of several options should be preferred over the others, taking into account several measures (the criteria), is addressed – not necessarily answered - by adopting an MCDA approach. Central to any such approach is defining the nature and content of the decision-making criteria. In what follows, we take a simplistic perspective and consider the typical MOE to be a numerical rating that captures the value of a fundamental aspect of the decision at hand. As per Section 3.3, the scale and limits for each MOE are selected such that they are comparable to one another in a consistent manner (i.e., on some {min..max} cardinal scale). For example, each MOE might be rated on a scale of 1 to 10, 1 being the worst value and 10 the best value. An MOE may result from a single measurement, or it may represent an ad-hoc aggregation of rated or ranked criteria. As alluded to in Section 1, MOEs feed directly into the goal that the decision aims to achieve. In the context of tactical-level modeling, they should represent the operational effectiveness of the analyzed systems. Specifically, in the simplest form options for attaining the goal are ranked based on how well they perform relative to a weighted sum of the chosen MOEs.

MOPs are follow-on[7] measures and are more loosely constrained. They generally stand independently and there is no general need to convert them to a common scale. They represent an added value quantifying what to expect once a course of action is adopted. For instance, if the sustainability of an option is in question, MOPs that keep track of resources could play a significant role in flagging potential problems. Occasionally, the assessment focus may shift to MOPs in case the MOEs do not provide sufficient distinction between analyzed options. In such cases, MOPs may cross a threshold and enter into the decision-making process of subsequent analyses, at which time they must be treated as MOEs. It needs to be noted here that the decision whether a particular characteristic is an MOP or MOE is scenario dependent, and what is considered an MOP in one study can be an MOE in another study.

Currently in LFORT, the primary MCDA tool used is MARCUS[8]. As discussed earlier, MARCUS accepts ordinal information regarding the performance of options (their rankings), thereby losing information about the actual differences in performance (some other tools might accept both ordinal and cardinal information). It can also handle a relative quantitative weighting of the measures. This method is very useful for determining which option is better, but does not indicate how much better.

Whether or not the numerical difference between the performance values of options relates to how much better one option is than another depends on the aggregation scheme used.

There are a number of aggregation methods, each having unique assumptions and characteristics. The well-known weighted sums method assumes full compensation of performance values (trade-off of bad and good performance) and independence between the measures. While the latter (independence) is usually assumed, it is not always justified. The main problem arises when several measures (or perhaps even most of the measures) rely on a common source of information.

[7] These could also be called ‘tertiary’, ‘threshold’ or simply ‘extra’ measures. Follow-on is chosen since they represent variables that typically represent the efficiency and performance of the compared systems.

[8] MARCUS [31] is an in-house tool developed by DRDC CORA’s Central Operational Research Team. There are other commercial MCDA tools available, such as TOPSYS, used by the TNO.

In the next section some methods for valuing MOEs and their impacts on option ranking through MOE aggregation are discussed.

4.1. The Impact of Rescaling MOEs on Option Ranking

A key process in the assessment of multiple options using MCDA to obtain relative option effectiveness is assigning values to MOEs that allow one to compare them on equal footing. We begin with an example that illustrates how the MOE valuing system that is chosen can impact option ranking. ‘Solution 1’ details a simple, artificial example involving at first three options (e.g., weapon systems A, B and C), and later only two (Option C is subsequently dropped). There are two MOEs used in the analysis - the number of expended rounds and the number of BLUE casualties. For both MOEs, it is assumed that 'less is better' and the minimum is zero. How can the results be combined so that it is possible to state, for instance, that Option A is N% better (or worse) than Option B?

4.1.1. Solution 1: Relative to Best

The initial proposed solution could be summarized as follows.

i) Determine the raw values of the MOEs for all options (raw values must not be negative[9]).

ii) Determine the relative effectiveness $RE_{i\,j}$ of the i^{th} option with respect to the j^{th} MOE, taking the best (lowest) value across options as a baseline of 100%. If the best value of MOE j happened to occur in option k, then $RE_{kj} = 100\%$ and

$$RE_{ij} = \begin{cases} \dfrac{v_{kj}}{v_{ij}} \times 100\%, & v_{ij} > 0 \\ 100\%, & otherwise \end{cases} ,$$

where v_{ij} represents the raw value of the MOE j computed for option i.

iii) Apply the relative weights of the MOEs and compute the weighted sum of MOEs

iv) Compare the resulting numbers for different options and rank the options.

Continuing with the example described above, it is assumed that the results of the simulation led to values shown in Table 1 for the two MOEs applied to all three options. The relative effectiveness is calculated using the best option as 100% (e.g., 50 rounds within the Ammunition Expenditure MOE. Note that Option C has BLUE casualties rated significantly better (by a factor of four) than casualties reported from the other options.

[9] If the raw MOE values have zero (0) as a minimum, the method can still be applied, however all nonzero values will have 0% relative effectiveness, which is not particularly informative. Thus, relative to 0, all values are considered equally bad using this method.

Table 1. *Results for MOEs for the notional example of three Options (A,B,C) and two MOEs.*

	Ammunition Expenditure		**BLUE Casualties**	
Option	***Value***	***Relative Effectiveness***	***Value***	***Relative Effectiveness***
A	100	50%	2	25%
B	50	100%	4	12.5%
C	75	75%	0.5	100%

It is further assumed that the ammunition expenditure is assigned a weight of 0.25 and the BLUE casualties 0.50. Options are scored by summing the MOE relative effectiveness values multiplied by the appropriate MOE weights. Calculating the aggregate option scores (Table 2) and arranging the results in descending order leads to a ranking of C>B>A: Option C is the best option, then Option B and last is Option A. Naively, it can be concluded that, for instance, since B>A by 6.25%, then Option B is 25% better than Option A (the difference of 6.25% is 25% of Option A's score of 25%).

Table 2. *Aggregate values for the notional example of three Options (A,B,C) and two MOEs.*

		Relative Performance			Weighted Score		
MOE	**Weight**	**Option A**	**Option B**	**Option C**	**Option A**	**Option B**	**Option C**
Ammunition Expenditure	0.25	50%	100%	75%	12.5%	25%	18.75%
BLUE Casualties	0.50	25%	12.5%	100%	12.5%	6.25%	50%
Total Score					**25%**	**31.25%**	**68.75%**

Now consider the same problem, except with Option C removed. In a realistic scenario, this could happen in the post-analysis stage for several reasons. For example, Option C might become infeasible due to unexpected budgetary constraints, equipment unavailability, or logistical factors alluded to by existing MOPs. Fixing all other aspects, it can be seen that the relative effectiveness changes, as shown in Table 3.

Table 3. *Results for MOEs for the notional example of two Options (A,B) and two MOEs.*

	Ammunition Expenditure		**BLUE Casualties**	
Option	***Value***	***Relative Effectiveness***	***Value***	***Relative Effectiveness***
A	100	50%	2	100%
B	50	100%	4	50%

The ordinal results are presented in Table 4, which now shows a ranking of A>B by 12.5%. Naively, it can be concluded that Option A is now 25% better than Option B (the difference of 12.5% is 25% of Option B's score of 50%) compared to the exact opposite notion found in the previous example. The meaning of this 25% is somewhat abstract since it is an aggregated figure composed of scores and weights. However, one could state that it means that the customer is willing to pay 25% more for Option A than for Option B. The expressed difference between the options is only meant to be informative, and should not be taken as an absolute justification for selection.

Table 4. *Aggregate values for the notional example of two Options (A,B) and two MOEs.*

		Relative Effectiveness		**Weighted Score**	
MOE	**Weight**	**Option A**	**Option B**	**Option A**	**Option B**
Ammunition Expenditure	0.25	50%	100%	12.5%	25%
BLUE Casualties	0.50	100%	50%	50%	25%
Total Score				**62.5%**	**50%**

The main point is that Option A now outscores Option B in the pair-wise comparison of attributes, despite the fact that when Option C was included in the mix, one would come to the opposite conclusion. Thus, in general, the relative performance of options can change with the deletion or introduction of options under this valuing scheme. This has implications for the scenario outlined above, wherein a forerunning option had to be dropped by the sponsor post-analysis. The sponsor would not be able to use the remaining rankings with confidence. A new analysis would have to be performed to determine the new rankings using the reduced set of options. In the next section, one possible method for avoiding this problem is outlined.

To help understand this phenomenon, the overall impact of the rescaling introduced by the exclusion of Option C is visualized in Figure 1. In Figure 1, options for the two cases appear as vertices on a graph with three axes: AMMO, CAS, and SCORE. The (weighted) SCORE axis is the vertical axis, so higher points correspond to higher scoring options on a scale of 0 to 1. The AMMO and CAS axes also range from 0 to 1. The first case with 3 options is labelled by vertices A1, B1, and C1 for options A, B, and C respectively. The three vertices form a triangle. Similarly, the second case with 2 options has vertices A2 and B2, forming a line.

The vertices of the red triangle shows how options A, B, and C score relative to one another. The blue line near the top of the graph shows how options A and B score relative to one another in the absence of Option C. For the 2-option case, A2 clearly scores higher than B2, illustrating why A2 is superior in this instance. For the 3-option case, the rescaling of the casualty (CAS) axis clearly demonstrates how Option C (vertex C1) emerges as the better option, and furthermore shows how options A and B (now vertices A1, B1 are 'down-sized' to a level well below that of Option C in terms of the weighted score (compare vertices A1 and B1 relative to vertices A2 and B2, respectively). Notice that when all three options are considered together, Option A (vertex A1) falls below Option B (vertex B1) on the MOE axis[10].

The reason for the switch in rankings for A and B between the two cases can be explained as follows: in the 2-option example, ammunition and casualty values for Options A and B were of the same order of magnitude and ranking was dominated by the value of the 'BLUE casualties' MOE wherein A outperformed B (Figure 1, blue line from A2 to B2: A2 is higher). In the original example with three options, the leading casualty value (Option C) was sufficiently better than its competitors. This diminished the importance of the difference between Options A and B for this category to the point of being insignificant (at the scale dominated by C). However, given that the raw ammunition expenditure value for Option C is in line with values for options A and B in the 3-option case, differences in that category were still significant. Since Option B scored notably better in ammunition expenditure than Option A, this difference clearly was the dominating factor in determining the original relative ordering of A and B in the presence of C (Figure 1, triangle edge from A1 to B1: B1 is higher). It needs to be noted here that the differences in casualty magnitudes in the two examples are comparable to MOE values found in realistic games. Thus it can expected that using the *Relative to Best* method of transforming MOEs leaves one susceptible to this effect.

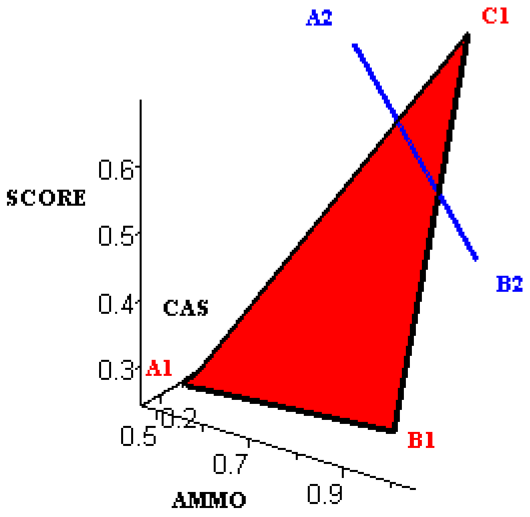


Figure 1. *A Geometric Visualization of Option Ranking.*

[10] Note that all possible MOE values must lie in the plane defined by the equation 0.25 * X + 0.5 * Y = Z, where X is the AMMO axis, Y the CAS axis and Z is the MOE axis.

In cases where Option C has the best raw casualty MOE value (denoted v_{C2}) and all other MOE values are fixed, the 'crossover point' occurs at $v_{C2} = 1$, such that

- A<B when $0 < v_{C2} < 1$;
- A=B when $v_{C2} = 1$; and
- A>B when $1 < v_{C2} < 2$.

Indeed, testing values $v_{C2} = \{0.5,1,1.5\}$, the option scores for A and B are found to be consistent with the above assertion (note that the score for Option C is unchanged at 68.75%):

- $v_{C2} = 0.5$: Score A = 25%, Score B%=31.25;
- $v_{C2} = 1.0$: Score A = 37.5%, Score B=37.5% and;
- $v_{C2} = 1.5$: Score A = 50%, Score B=43.75%.

Figure 2 depicts how the overall scores for options A and B vary as a function of v_{C2} in the interval $0 < v_{C2} < 2$ within the 3-option example, above.

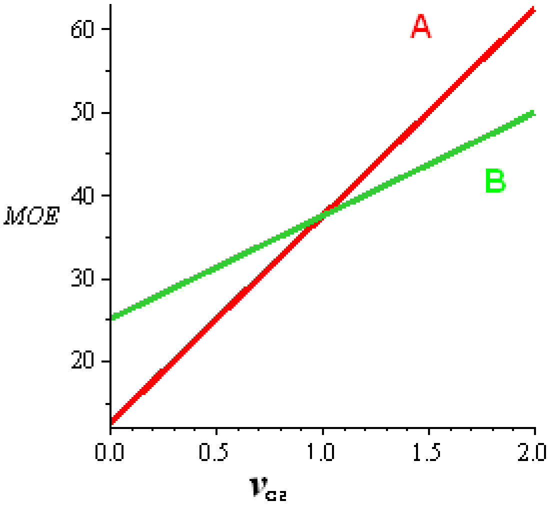


Figure 2. *Overall Scores for Options A and B as v_{C2} Varies*

It is interesting to note that the phenomenon described also applies to the selection of an employee from a pool of candidates. If the candidates were evaluated relative to one another based on the extent of their job qualifications (the criteria), then, for instance, if the top candidate refused a job offer then the relative ranking of the remaining candidates could switch. Thus 'number 2' on the original list, the next person you would normally call, actually might be 'number 3' now that 'number 1' is no longer involved in the competition.

4.1.2. Solution 2: Objective Scales

One possible solution to the problem outlined above is to make the scoring mechanism independent of the effectiveness of individual options (it would be determined prior to the modeling). That means that at the time of scenario and MOE definition the scales for scoring

MOEs need to be determined as well. Thus a fixed scale is created. The score of an option is determined based on the position on this fixed scale and the weight of each criterion, rather than based on the relative effectiveness derived via measurements made for the individual options. The introduction of a new option cannot change the relative effectiveness of the other options. Figure 3 shows an example of a scoreboard for casualties. The minimum score is 0.1 (10%) for 16 and more casualties, and the maximum value is 1.0 (100%) for 1 and fewer casualties. Note that other types of dependencies (linear, nonlinear) could have been chosen instead.

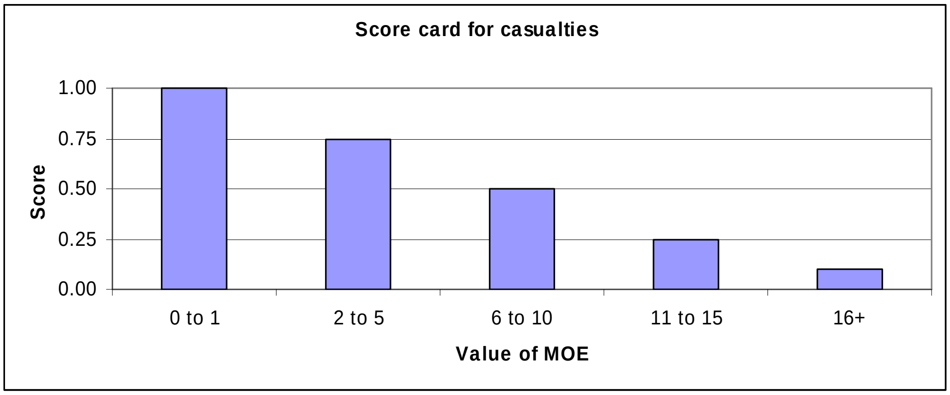


Figure 3. *Example of a scoreboard for casualties.*

With the introduction of fixed scores, the end result is a grade for every option. This prevents variation in the assigned score with respect to a presence or absence of a particular option, while still maintaining the significance of any differences.

A potential drawback of this method is that the number and span of the intervals will influence the scoring system. For instance, numerically proximal values lying near a categorical boundary can become separated into distinct categories, thus magnifying a small difference. One possible method of alleviating this is to determine the scores after the results have been recorded and any natural clusters identified. Categorical boundaries would then be chosen so as to avoid such circumstances, if possible. Note that the intention here is not to allow the distribution of the results to drive the setting of categorical boundaries within the scoring system, but rather to provide a means to take advantage of any flexibility resident in the high-level interpretation or abstraction of the results. A disadvantage of this approach is a potential lack of consistency across different studies (e.g., a follow-up study); each study might yield a new set of categorical boundaries.

As with many other MCDA schemes, a possible setback of using objective scales is the sensitivity of the results with regards to the weights assigned to individual MOEs. Therefore it is recommended to perform a sensitivity analysis for every study to ensure that the results are robust with respect to the weights chosen[11]. This will improve the confidence in the results. If small deviations in the weights lead to different rankings, especially within front-running options, then the rankings are not robust and distinctions in ranking between the options are likely to be of less value than desired.

[11] As mentioned earlier, the above statement about the sensitivity of results is true of any weighted system, not just this one.

4.1.3. Other Issues

There is a new problem introduced by the predefined score. Suppose a scoring mechanism has been introduced stating that four or more casualties are "unacceptable". Options incurring four or more casualties therefore receive the same low score (zero). How might one deal with a situation when two options result in significantly different numbers of casualties, but they are both more than four? For example Option A results in five casualties and Option B in fifteen casualties. Is Option A better than Option B? This problem touches on the definition of the criteria for success. The natural response would be that if it was determined that more than four casualties are unacceptable, both options should fail. However, for study purposes it might be still desirable to maintain the distinction between the two, especially since the number of casualties is often a function of the input parameters (e.g. kill probability, personal protection characteristics) and assumptions (e.g. behaviour, risk tolerance) made to formulate the model. In addition, the combat models typically dealt with are stochastic, and thus the number of casualties are often distributed within a certain range.

While this issue can be managed by means of a post game assignment of the scores as mentioned above, care must be taken to avoid assigning the scores in such a manner as to create the appearance of favouring one option over another. This could be achieved by removing the option labels from the results so that the body responsible for assigning the score does not know which option achieved a particular result. However there is no easy answer to this question. In this instance, the scoring system might be too severe. Also, at the definition stage it needs to be born in mind that one deals with modeling for study purposes and not real life operations. While a particular outcome might not be acceptable in real life, the definition of measures needs to reflect the fact that sufficient variability in the results is necessary to allow for the statistical analysis. Therefore the definition of the score should provide a fine enough scale to capture such differences. If a situation such as the one above arises, it might be necessary to go back to the definition of the measures and redo the analysis with a more detailed definition. This can lead to additional problems, however, since the relative rankings of all options may change. Therefore, rescoring should be used carefully and any changes as well as the original rankings should be documented. As mentioned earlier, performing sensitivity analysis is recommended to indicate how the study's conclusions may have been influenced by such changes.

The scoreboard also provides a method for capturing a more complicated issue, illustrated in the following example. Assume that BLUE uses weapon B1 and RED has weapons R1 and R2. Ranges of the individual weapons are shown in Table 5.

Table 5. *Example of notional weapon ranges for RED and BLUE.*

Weapon	B1	R1	R2
Range (m)	300	100	200

It is further assumed that it is better for BLUE to neutralize RED while they are still outside of their range. The question is how to define a measure that takes this assumption into account? A scoreboard provides the means to do it in a straightforward fashion.

For example, experts determined that neutralising within the range of both of RED's weapons is worth a score of 0 ('bad'), outside RED range1 is worth a score of 0.5 ('good') and outside of RED range 2 is worth a score of 1 ('best'). The devised scoreboard is shown in Table 6.

Table 6. *Example of notional scoreboard for range of engagement.*

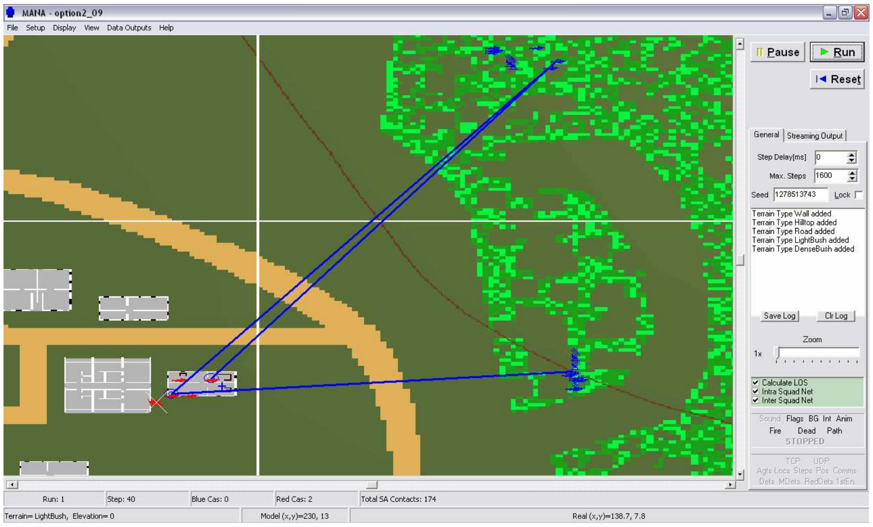
Distance of RED from B1 in meters	<100	100-200	> 200
B1 destroyed RED	0.0	0.5	1.0

4.2. Examples of the Scored MOEs

To test the proposed methodologies (Sections 4.1.1 and 4.1.2) a comparison was performed ranking the results from two previous studies for a subset of measures. The rankings were at first obtained using MARCUS, and then using both of the proposed methodologies. The results follow below.

4.2.1. Scenario 1: Infantry Assault into Urban Terrain

The first scenario consisted of a single BLUE section assaulting a RED team situated in a building at the edge of a village. The RED force had the support of a single armoured personnel carrier (APC) that moved around the house. BLUE had a support of a FIREBASE located northeast of the building. The initial deployment is shown in Figure 4. Three different options were investigated.



Step Delay[ms] 0
Max. Steps 1600
Seed 1278513743 Lock □

```
Terrain Type Wall added
Terrain Type Hilltop added
Terrain Type Road added
Terrain Type LightBush added
Terrain Type DenseBush added
```

☑ Calculate LOS
☑ Intra Squad Net
☑ Inter Squad Net

Run: 1 | Step: 40 | Blue Cas: 0 | Red Cas: 2 | Total SA Contacts: 174

Terrain= LightBush, Elevation= 0 | Model (x,y)=230, 13 | Real (x,y)=138.7, 7.8

Figure 4. *Initial deployment for the force-on-force scenario.*

To compare the effectiveness of each option, five measures of effectiveness (MOEs) were used. One was the overall mission success with possible values 0 to 4, four being the best. The other MOEs were:

- BLUE residual combat strength (RCS) expressed as the ratio of BLUE soldiers remaining to the original number of BLUE soldiers;
- Extraordinary RED casualties (ERC) expressing difference of RED casualties from the mean loss exchange ratio (LER) over all the options, a derivative of the loss exchange ratio (LER), which is valid even in the case of no BLUE casualties (wherein the LER is infinite)[12];
- Time to mission success; and
- Fratricide.

The results are shown in Table 7. Mission Success varied minutely between the different options. For other MOEs there was an apparent separation between Option 1 and Options 2 and 3.

Table 7. *Results of the force-on-force scenario.*

	Mission Success	**BLUE RCS**	**ERC**	**Time (sec)**	**Fratricide**
Option 1	1.00	0.73	-4.8	124	0.01
Option 2	1.00	0.93	2.0	58	0.00
Option 3	0.98	0.95	2.8	84	0.00

At first, the results were analyzed using the BRANDO [32] tool developed by DRDC CORA to obtain rankings of the options for individual MOEs. These are shown in Table 8. Afterwards, the ranks were used to obtain the final rankings of the three options using MARCUS. The final rankings were Option 2 ranked first, Option 3 ranked second, and Option 1 ranked third.

[12] ERC is calculated as follows. At first the overall mean LER (a ratio of all the RED casualties to all the
BLUE casualties) is calculated, considering all of the options that were modeled. Therefore, it is sufficient if there was at least a single BLUE casualty over all the options. Then the theoretical red casualties are calculated for each replication as a product of the average LER and the BLUE casualties for that replication. Finally, the difference between this value and actual RED casualties is calculated. If the number is negative, it implies that fewer RED were killed than was to be expected; if it was positive, more were killed. Since this method eliminates the need to divide RED/BLUE casualties for every replication, the potential for division by zero is eliminated.

Table 8. *Rankings of options for the force-on-force scenario.*

	Mission Success	BLUE RCS	ERC	Time	Fratricide	Total
Weight	0.470	0.200	0.055	0.070	0.045	**---**
Option 1	1	2	3	3	2	**3**
Option 2	1	1	2	1	1	**1**
Option 3	1	1	1	2	1	**2**

Figure 5 shows the relative distance of the final ranks of the three analyzed options. This distance is based solely on the ranks and weights of individual measures, and it does not consider the actual values of individual MOEs. The distance between Options 2 and 3 is much smaller than the distance of either of them from Option 1.

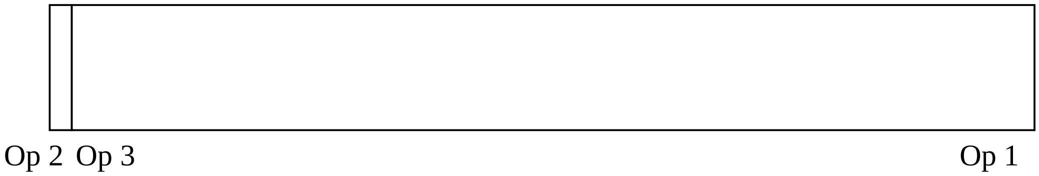


Figure 5. *Distance between ranks obtained from MARCUS for the force-on-force scenario.*

Using the *Relative to Best* scoring method (Section 4.1.1), all the values were normalized in terms of the best performance, which was assigned a score of 1.00. In the case of the ERC, the values were at first transformed using an affine transformation (adding a constant value) in order to make all the values positive. The scores are in Table 9.

Table 9. *Scoring method 1 (Relative to Best) for the force-on-force scenario – normalized scores.*

	Mission Success	BLUE RCS	ERC	Time (steps)	Fratricide
Option 1	1.00	0.77	0.00	0.47	0.00
Option 2	1.00	0.98	0.89	1.00	1.00
Option 3	0.98	1.00	1.00	0.69	1.00

These scores were then combined with the appropriate weights to obtain the final scores. The outcome is in Table 10. The last column contains the total scores for the individual options. The results lead to Option 2 ranking first, Option 3 ranking second, and Option 1 ranking last, far behind the first two. There is only a slight difference between Options 2 and 3 (scores 0.83

and 0.81 respectively). Option 1 scored 0.66. Overall, the rankings are consistent with the results obtained from MARCUS.

Table 10. *Scoring method 1 (Relative to Best) for the force-on-force scenario – weighed scores.*

	Mission Success	BLUE RCS	ERC	Time	Fratricide	Total
Weight	0.470	0.200	0.055	0.070	0.045	**---**
Option 1	0.470	0.154	0.000	0.033	0.000	**0.656**
Option 2	0.470	0.196	0.049	0.070	0.045	**0.830**
Option 3	0.461	0.200	0.055	0.048	0.045	**0.809**

The second scoring method, *Objective Scales* (Section 4.1.2), requires developing absolute scales for individual measures. While this should be done in advance without knowledge of the actual results to allow for capturing the relevant difference between options on the basis of the sponsor's judgement, in this instance the scales had to be developed *a posteriori*. These scales were developed incorporating the significant differences obtained using BRANDO. For simplicity, the scores were selected such that for each of the MOEs there would be six possible values between zero and one (0, 0.2, … , 1.0).

For Mission Success, 100% success was assigned score 1.0, and 50% success was assumed unacceptable (score of zero). For RCS, it was assumed that losing eight or more soldiers was unacceptable (entire section lost). Therefore RCS equal to 0.90 was assigned a score of 0. An RCS of 1.0 was assigned a score of one (1). For the ERC, it was assumed that the value of 5 was assumed to have a score of one (1) (five more than the "expected" RED casualties). ERC less than zero (less than "expected" number of RED casualties) was assigned a score of zero (0). Time less than 50 was assigned a score of one, and time more than 130 was assigned a score of zero (0). Fratricide was the exception. It was assumed that it was either one (if there was no fratricide), or zero (if there was fratricide). The scores are in Table 11.

Table 11. *Scale of scores assigned to individual MOEs for the force-on-force scenario.*

Score	Mission Success	BLUE RCS	ERC	Time (steps)	Fratricide
1.0	0.91-1.00	0.96-1.00	4.1-5.0	<50	0.0
0.8	0.81-0.90	0.91-0.95	3.1-4.0	51-70	--
0.6	0.71-0.80	0.86-0.90	2.1-3.0	71-90	--
0.4	0.61-0.70	0.81-0.86	1.1-2.0	91-110	--
0.2	0.51-0.60	0.76-0.80	0.1-1.0	111-130	--
0.0	<0.50	<0.75	<0.0	>130	>0.0

These values were then used to score the individual options. The outcome is in Table 12.

Table 12. *Scoring method 2 (Objective Scales) for the force-on-force scenario – normalized scores.*

	Mission Success	BLUE RCS	ERC	Time (steps)	Fratricide
Option 1	1.0	0.0	0.0	0.2	0.0
Option 2	1.0	0.8	0.4	0.8	1.0
Option 3	1.0	0.8	0.6	0.6	1.0

Then the appropriate weights were used to obtain the weighted scores (Table 13). The resulting ranks were consistent with the results of the other two methods. Like in the first scoring approach, Options 2 and 3 were close to each other (scores 0.753 and 0.750), and Option 1 was far behind (0.484).

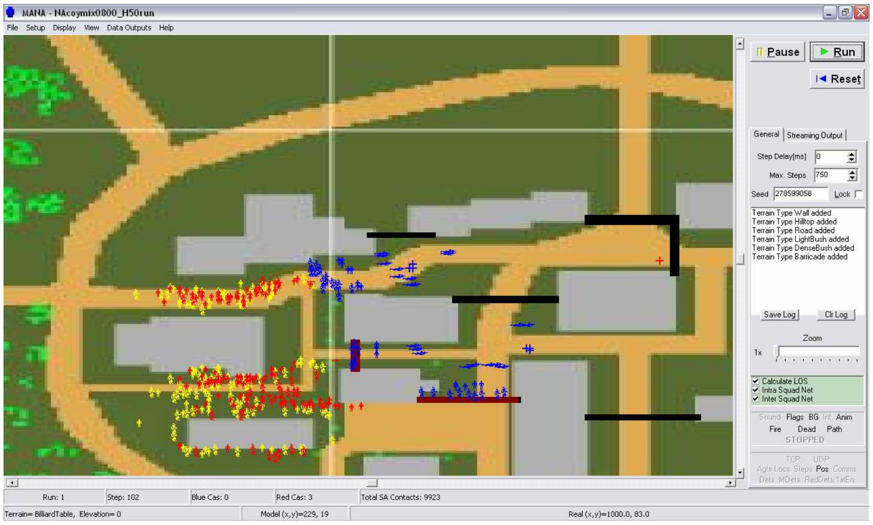
Table 13. *Scoring method 2 (Objective Scales) for the force-on-force scenario – weighted scores.*

	Mission Success	BLUE RCS	ERC	Time (steps)	Fratricide	Total
Weight	0.470	0.200	0.055	0.070	0.045	**---**
Option 1	0.470	0.000	0.000	0.014	0.000	**0.484**
Option 2	0.470	0.160	0.022	0.056	0.045	**0.753**
Option 3	0.470	0.160	0.033	0.042	0.045	**0.750**

Thus, for the first test scenario, both scoring methods yielded results consistent with the outcome of MARCUS. A caveat has to be included here. Since the score scale in the *Objective Scales* method was devised after the modeling (and using the knowledge of the results), the findings are not entirely objective. If different score-scale was used, the outcome might have been different (e.g., the options might have ended up tied). Nevertheless, the results would have been similar in any case.

4.2.2. Scenario 2: Crowd Confrontation

The second scenario consisted of a BLUE company tasked to confront an aggressive crowd and deny the crowd access to a particular area of a town. The crowd consisted of a mixture of very aggressive gangs together with a less aggressive complement of women, children, elderly people and middle-aged men. A screenshot of the scenario is shown in Figure 6. Five different options were investigated.



Step Delay[ms] 0

Max. Steps 750

Seed 278599058 Lock □

```
Terrain Type Wall added
Terrain Type Hilltop added
Terrain Type Road added
Terrain Type LightBush added
Terrain Type DenseBush added
Terrain Type Barricade added
```

☑ Calculate LOS
☑ Intra Squad Net
☑ Inter Squad Net

Run: 1 | Step: 102 | Blue Cas: 0 | Red Cas: 3 | Total SA Contacts: 9923

Terrain= BilliardTable, Elevation= 0 | Model (x,y)=229, 19 | Real (x,y)=1000.0, 83.0

Figure 6. *Screen-shot of the crowd confrontation scenario.*

To compare the effectiveness of each option, a subset of five MOEs, selected from the complete set, was used (ten measures were used for the actual study). One was the overall mission success with possible values from zero to three, three being the best. The other MOEs were:

- Number of incapacitations by non-lethal launchers;
- Number of fatalities caused by the use of lethal firepower;
- Time to influence the crowd; and
- BLUE RCS (ratio of BLUE soldiers remaining at the end of the mission to the original number of BLUE).

Five of the options are ranked below. The results are shown in Table 14. Option 5 dominated on all MOEs except for RCS. Therefore it was reasonable to expect that that option would rank first. Similarly, Option 1 performed the worst except for RCS. There was a sufficient variability between individual options to proceed with the comparison between rankings obtained using MARCUS and using the scoring methods.

Table 14. *Results of the crowd confrontation scenario.*

	Mission Success	Incapacitations		Time (min)	BLUE RCS
		Non-lethal	Lethal		
Option 1	0.55	10	11	50	0.91
Option 2	0.65	21	6	18	0.84
Option 3	0.80	29	0	36	0.96
Option 4	0.53	24	0	41	0.98
Option 5	0.85	34	0	6	0.94

As for the previous scenario, the results were analyzed using BRANDO to obtain rankings of the options for the individual MOEs. The rankings are shown in Table 15. Afterwards, these ranks were used to obtain the final rankings of the five options using MARCUS. The final rankings were Option 5 ranked first, Option 3 ranked second, Option 2 third, Option 4 fourth, and Option 1 ranked fifth.

Table 15. *Rankings of options for the crowd confrontation scenario.*

	Mission Success	Incapacitations		Time	BLUE RCS	Total
		Non-lethal	Lethal			
Weight	0.35	0.20	0.15	0.07	0.05	**---**
Option 1	2	5	3	5	4	**5**
Option 2	1	4	2	2	5	**3**
Option 3	1	2	1	3	2	**2**
Option 4	2	3	1	4	1	**4**
Option 5	1	1	1	1	3	**1**

Figure 7 shows the distance between the final ranks obtained from MARCUS (based solely on the ranks on individual measures. The distance between Options 3 and 5 was negligible compared to the distance between other options. Also Options 2 and 4 were rather close to each other. Therefore potential reversal between these two pairs of options would not be surprising when the actual distance between options on individual measures is considered.

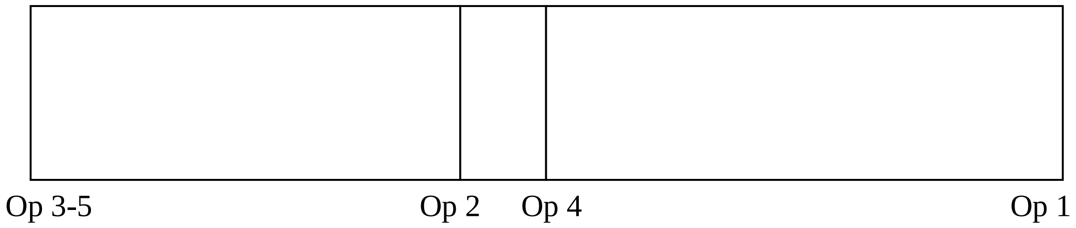


Figure 7. *Distance between ranks obtained from MARCUS for the crowd confrontation scenario.*

Using the first scoring method, *Relative to Best*, all of the values were normalized in terms of the best performance that was assigned score of 1.00. The scores are listed in Table 16. Notice that the Lethal Incapacitations MOE has zero (0) casualties as a minimum value. Therefore, nonzero raw values for this MOE received a score of 0%, or in other words, relative to zero all positive values appeared equally inferior.

Table 16. *Scoring method 1 (Relative to Best) for the crowd confrontation scenario – normalized scores.*

	Mission Success	**Incapacitations**		**Time**	**BLUE RCS**
		Non-lethal	**Lethal**		
Option 1	0.65	0.29	0.00	0.12	0.93
Option 2	0.76	0.62	0.00	0.33	0.86
Option 3	0.94	0.85	1.00	0.17	0.98
Option 4	0.62	0.71	1.00	0.15	1.00
Option 5	1.00	1.00	1.00	1.00	0.96

These scores were then combined with the appropriate weights to obtain the final scores. The outcome is in Table 17. The last column contains the total scores for the individual options. The rankings are slightly different from the results provided by MARCUS. Options 5 and 3 ranked first and second respectively just like in MARCUS (with scores 0.82 and 0.71, respectively), but Options 2 and 4 swapped their places (with scores 0.46 and 0.57, respectively). It was caused by the closeness of these two options on most of the measures, except for the number of lethal incapacitations. On this particular MOE Option 4 far outscored Option 2, which lead to a slightly better overall performance. This large difference in performance on this particular MOE was not captured in MARCUS (using only the ranks). This outcome actually highlights the relevance of capturing the magnitude of the difference between the options for the individual MOEs. In this case the two options performed comparably well on all the other MOEs, and if the Lethal Incapacitations were not considered, Option 2 would actually outperform Option 4 by a narrow margin (0.007). However, since the fatalities are an important consideration in this type of scenario, much poorer performance of Option 2 on the relevant measure should be reflected, as was the case using the *Relative to Best* scoring method. Note, however, that any deviation from the best raw MOE value of 0 translates into a major difference using this scheme, even if the raw value was, for instance, 0.000001 instead of 6. This again suggests that the scoring using fixed scales is better than the comparison with the best option. Option 1 ranked last (score 0.34). Overall, the difference in rankings compared to MARCUS is minimal.

Table 17. *Scoring method 1 (Relative to Best) for the crowd confrontation scenario – weighed scores.*

	Mission Success	Incapacitations		Time	BLUE RCS	Total
		Non-lethal	Lethal			
Weight	0.35	0.20	0.15	0.07	0.05	**---**
Option 1	0.226	0.059	0.00	0.008	0.046	**0.340**
Option 2	0.268	0.124	0.00	0.023	0.043	**0.458**
Option 3	0.329	0.171	0.150	0.012	0.049	**0.711**
Option 4	0.218	0.141	0.150	0.010	0.050	**0.570**
Option 5	0.350	0.200	0.150	0.070	0.048	**0.818**

Again, for the second scoring method, *Objective Scales*, scoring scales for the individual measures had to be developed. Like in the previous case, the scales were developed *a posteriori*, incorporating the significant differences obtained using BRANDO. For simplicity, the scores were selected such that for each of the MOEs there would be six possible values between zero and one (0, 0.2, …, 1.0). The exception was Mission Success for which only four possible scores were assigned. For the Non-lethal Incapacitations it was impossible to develop a uniform scale that would assign different values to each of the options with 6 values. Therefore two of the close values, while ranked different by BRANDO, were assigned the same score.

For Mission Success, a score of 1.0 was assigned for values over 0.9, and 0 for values under 0.4. For non-lethal incapacitations, more than 30 were assigned a score of 1, and less than 10, score 0. For lethal incapacitations, 1 or 0 were assigned a score of 1 and more than 9 scored 0. Time-wise, a score of 1 was given to the duration of 10 minutes or less, while 0 was given to duration over 50 minutes. BLUE RCS was assigned 1 if the result was 1, and 0 if it was below 0.9. The scores are in Table 18.

Table 18. *Scale of scores assigned to individual MOEs for the crowd confrontation scenario.*

Score	Mission Success	Incapacitations		Time	BLUE RCS
		Non-lethal	Lethal		
1.0	0.91-1.00	>30	0-1	0-10	1.00
0.8	0.65-0.89	25-29	2-3	10-20	0.98-0.99
0.6	--	20-24	4-5	20-30	0.96-0.97
0.4	0.40-0.64	15-19	6-7	30-40	0.94-0.95
0.2	--	10-14	8-9	40-50	0.92-0.93
0.0	<0.40	<10	>9	>50	0.90-0.91

These values were then used to score the individual options. The outcome is in Table 19.

Table 19. *Scoring method 2 (Objective Scales) for the crowd confrontation scenario – normalized scores.*

	Mission Success	Incapacitations		Time	BLUE RCS
		Non-lethal	Lethal		
Option 1	0.4	0.2	0.0	0.0	0.2
Option 2	0.8	0.6	0.4	0.8	0.0
Option 3	0.8	0.8	1.0	0.4	0.6
Option 4	0.4	0.6	1.0	0.2	0.8
Option 5	0.8	1.0	1.0	1.0	0.4

Afterwards the appropriate weighted scores were calculated (Table 20). The resulting ranks were consistent with the results obtained from MARCUS. The large difference in the lethal incapacitations was somewhat reduced by the linear fixed scale leading to the results consistent with using solely ranks of individual options like in MARCUS. Options 2 and 4 (which ranked 3rd and 4th) were close to each other. However, if it was desirable to better capture such a wide range of values, a non-linear (e.g. logarithmic) scale could be employed.

Table 20. *Scoring method 2 (Objective Scales) for the crowd confrontation scenario – weighted scores.*

	Mission Success	Incapacitations		Time	BLUE RCS	Total
		Non-lethal	Lethal			
Weight	0.35	0.20	0.15	0.07	0.05	---
Option 1	0.140	0.040	0.000	0.000	0.010	**0.190**
Option 2	0.280	0.120	0.060	0.056	0.000	**0.516**
Option 3	0.280	0.160	0.150	0.028	0.030	**0.648**
Option 4	0.140	0.120	0.150	0.014	0.040	**0.464**
Option 5	0.280	0.200	0.150	0.070	0.020	**0.720**

As with the first scenario, in the second scenario both scoring methods yielded results more-or-less consistent with the outcome of MARCUS. The caveat mentioned in the discussion of Scenario 1 applies here as well. Since the scoring scale in Method 2 (*Objective Scales*) was devised after the modeling (and using the knowledge of the results), the findings are not entirely objective. Utilizing a different scoring scale might have resulted in a different outcome (e.g., the options might have ended up in a tie). Nevertheless, the results would have been similar in any case.

Overall, the *Objective Scales* method (second) appeared to be more consistent with MARCUS. The fixed scales, independent of the actual obtained values, seemed to slightly reduce the possibility of capturing the large differences in outcomes for the individual MOEs, but this could be easily alleviated by using non-linear scales. In the scale derived from the actual obtained numbers, differences between the values of individual measures are preserved. In any case, scoring methods for the assessment of simulation results provided additional information about the relative performance of the available options. Thus these methods show potential for enhancing the value of information obtained in a study, and thus provide a viable supplement, or even alternative to MARCUS.

5. SUMMARY AND RECOMMENDATIONS

5.1. Summary

In evaluating the results of combat simulations, some kind of MCDA typically needs to be used to provide rankings of individual options. Currently, LFORT uses the MCDA tool called MARCUS, and TNO uses TOPSYS methodology to obtain the final rankings of analyzed options. However, it was identified that MARCUS and similar tools does not provide sufficient capability of capturing the magnitude of differences in the performance of individual MOEs, and therefore two alternative MCDA methods based on scoring the MOEs (*Relative to Best* and *Objective Scales*) were proposed.

It was demonstrated that by using predefined common scales (O*bjective Scales* method), for different MOEs it is possible to develop a scoring system that allows for the determination of how much one option is better than another without concern for the possibility of rank-order switching due to post-analysis elimination of one or more options. The scoring system presented here incorporates both the results of individual MOEs as well as their relative weights. Care needs to be taken to define proper scales (often the best scales might be non-linear) to enable capturing the magnitude of differences between options. The *Relative to Best* method generally captures the magnitude of the difference well, but may be susceptible to rank reversal in case of omission of some options in the post-analysis stage, since it is dependent on the values obtained by the best option for each individual MOE. This can lead to misleading rankings.

Comparison with rankings generated using LFORTs' traditional MCDA tool (MARCUS) showed that there is a high degree of consistency between the two approaches for the cases considered. While the scoring methodology outlined herein lacks the mathematical rigour of MARCUS, it appears to be working well for practical problems that are likely to be encountered during analysis of combat model outcomes.

Thus, the scoring methodologies have the potential to provide a viable supplement, or perhaps even alternative to the currently used MCDA tool (MARCUS, in the case of LFORT). In the event that the two methods produce different rank orders, i.e., MARCUS ranks options A, B and C in that order while the scoring method ranks the same options as B, A and C, a careful examination of why that occurred must be conducted to then determine what the appropriate ranking should be. In such a case, MOPs may provide the additional insight needed to tip the scale towards choosing one option over another. Such comparison of two or more MCDA approaches could be a part of sensitivity analysis.

It was noted that a proper definition of measures, as well as proper distinction between MOEs and MOPs, are important aspects of a well-balanced statistical analysis of the system measurement results, supporting informed interpretations and decision-making. A number of *primary* MOEs corresponding approximately to the number of degrees of freedom in the scenario is most desirable. Furthermore, these measures should be defined as independent from one another as is possible in order to facilitate a simple and clear MCDA. As a rule of thumb, fewer distinctive measures are better, but not too few. Three to four MOEs should be sufficient for most scenarios. MOPs, characterize system performance rather than the operational effectiveness, and as such they should be considered separately. Since they normally do not enter into the rankings directly, there is no general need to restrict the number of MOPs for a given scenario. Together, MOEs and MOPs can provide an estimate of option sustainability, which is an important consideration for modern combat systems.

5.2. Recommendations

Based on the findings of this study, the following recommendations are being made with respect to the best practices for evaluating the results of combat simulations:

- A scoring system can be used to supplement or perhaps even replace the traditional MCDA methods employed to determine which option is the best and by how much;
- When considering sensitivity to option rank-order switching under a post-analysis elimination of options, and the possibility of measures for which the best result is zero, *Objective Scales* method is a better choice for valuing (or rescaling) MOEs than the *Relative to Best* method.
- The number of MOEs used to rank options should be reasonable (three to four MOEs should be sufficient for most studies) and should not exceed the number of the degrees of freedom;
- Care should be taken to ensure independence of the MOEs;
- Proper distinction between MOEs and MOPs should be maintained. MOPs can be used in the final decision-making, but not to rank the option effectiveness of options;
- Particular attention needs to be given to the definition of mission success since it is typically the weightiest measure. It needs to be well aligned with the stated mission objectives;
- A scenario-specific combination of MOEs and MOPs for leading options should be considered as a functional indicator when operation sustainability is at issue;
- Sensitivity analysis should be performed whenever possible. At a minimum it should consist in varying the weights assigned to the individual MOEs, but it can include varying the MCDA method as well; and
- When high-ranking options are close in the value of the ranking parameter, a separation ‘distance’ between options that measures how much one option is better than another should be provided, or if appropriate the options should be presented as equivalent in rank.

6. REFERENCES

1. Dexter, R.M., *The Close Action Environment (CAEn) V9.2 Training Manual, Part One: Graphical User Interface Overview*. DSTO-GD-0404. Edinburgh, Australia: Land Operations Division, Defence Science and Technology Organisation, 2004.

2. Bernier, M.Y., Jensen, G.K., *CAEn^{XP} Training Program*, DRDC ORD Technical Memorandum TM 2005-16, 2005.

3. Lauren, M.K. and Stephen, R.T., Fractals and Combat Modeling: Using MANA to Explore the Role of Entropy in Complexity Science, Fractals 10, 4 (2002): 481-489.

4. Ilachinski,, A., EINSTein: An Artificial-Life "Laboratory" for Exploring Self-Organized, Emergent Behavior in Land Combat , CRM D0002239.A1. Alexandria, Center for Naval Analyses, 2000.

5. Richardson, S.B. and Skinner, D.E., *Getting Started with HiLOCA*, QinetiQ Ltd., June 2007.

6. McIntosh, G.C. and Lauren, M.K., Genetic Algorithms Applied to Course-of-action Development Using the MANA Agent-based Model, Journal of Battlefield Technology, 9, 3, 2006.

7. Ilachinski, A., Artificial War: Multiagent-based simulation of combat, World Scientific (2004).

8. Belton, V. and Stewart, T.J., Multiple Criteria Decision Analysis: An Integrated Approach, Springer (2002).

9. Saaty, T.L., The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation, McGraw-Hill (1980).

10. McCaffrey, J., *Multi-Attribute Global Inference of Quality (MAGIQ)*, Software Test and Performance Magazine, 2, 7 (2005): 28-32.

11. Edwards, W. and Barron, H.F., *SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement*, Organizational Behaviour and Human Decision Processes, 60 (1994):306-325.

12. Charnes, A., Cooper, W., Rhodes, E., *Measuring the efficiency of decision-making units*, European Journal of Operational Research 2 (1978): 429-444.

13. Charnes, A., Cooper, WW., Management models and industrial applications of linear programming, Wiley (1961).

14. Schniederjans, M.J., Goal Programming: Methodology and Applications, Springer (1995).

15. Greco, S., Matarazzo, B., Słowiński, R., *Rough sets theory for multicriteria decision analysis*, European Journal of Operational Research 129, 1 (2001): 1-47

16. Roy, B., *Classement et choix en présence de points de vue multiples (la méthode ELECTRE)*, la Revue d'Informatique et de Recherche Opérationelle (RIRO) 8 (1968): 57-75.

17. Brans, J.P., Mareschal, B., Vincke, Ph., *PROMETHEE: a new family of outranking methods in multicriteria analysis*. In J.P Brans, editor, Operational Research, IFORS 84. North Holland, Amsterdam (1984) : 477-490.

18. Yang J. B. and Singh M. G., *An evidential reasoning approach for multiple attribute decision making with uncertainty*, IEEE Transactions on Systems, Man, and Cybernetics 24/1, (1994): 1-18.

19. Shafer, G.A., *Mathematical Theory of Evidence*, Princeton University Press (1976).

20. Hwang, C.L. and Yoon, K., Multiple Attribute Decision Making: Methods and Applications, A State of the Art Survey, Springer-Verlag, New York (1981).

21. Tryantaphyllou, E. and Baig, K., *The Impact of Aggregating Benefit and Cost Criteria in Four MCDA Methods*, IEEE Transactions on Engineering Management 52, 2 (2005): 213-226.

22. Bridgeman, P.W., *Dimensionless Analysis*, New Haven, CT: Yale Univ. Press (1922).

23. Miller, D.W. and Starr, M.K., *Executive Decisions and Operations Research*, Englewood Cliffs, NJ: Prentice-Hall (1969).

24. Dawes, R.M. and Corrigan, B., *Linear Models in Decision Making*, Psychological Bulletin 81 (1974): 95-106.

25. Wainer, H., Estimating Coefficients in Linear Models: It Don’t Make No Nevermind, Psychological Bulletin 83, 2 (1976): 213-217.

26. Jia, J., Fischer, G.W., Dyer, J., Attribute weighting methods and decision quality in the presence of response error: A simulation study, Journal of Behavioural Decision Making, 11 (1998):85-105.

27. Von Winterfeldt, D. and Edwards, W., *Decision Analysis and Behavioural Research*, Cambridge University Press (1986).

28. Schoemaker, J.H and Waid, C.C., An Experimental Comparison of Different Approaches to Determining Weights in Additive Utility Models, Management Science, 28, 2 (1982): 182-196.

29. Barron, H.F. and Barrett, B.E., *Decision Quality using Ranked Attribute Weights*, Management Science, 42, 11 (1996): 1515-1523.

30. Stillwell, W.G., Seaver, D.A., Edwards, W., *A Comparison of Weight Approximation Techniques in Multiattribute Utility Decision Making*, Organizational Behaviour and Human Performance, 28 (1981): 62-77.

31. Eles, P., *Visualizing Strength of Agreement in Consensus Rankings*, DRDC CORA Technical Note TN 2006-10 (2006).

32. Emond, E.J., A.E. Turnbull, *BRANDO: Breakpoint Analysis with Nonparametric Data Option* DRDC CORA Technical Memorandum, in preparation (2006).

33. Dobias, P., G. Woodill, *Assessment of Suitability of MANA for Combat Modeling in LFORT*, DRDC CORA Technical Report TR 2006-31 (2006).

34. Z. Bouayed, S. Bassindale, P. Dobias, G. Woodill, *Optimum Number of Non-Lethal Weapons Study – Nickel Abeyance Part I*, DRDC CORA Technical Memorandum 2007-11 (2007).

35. P. Dobias, Z. Bouayed, G. Woodill, S. Bassindale, Nickel Abeyance II, Optimum Number of Non-Lethal Launchers Study (Non-Interactive Modeling Using MANA), DRDC CORA Technical Report 2006-18, 2006.

36. P. Dobias, Z. Bouayed, G. Woodill, S. Bassindale, Optimum Number of Non-Lethal Launchers Study – Nickel Abeyance II.5 (Alternative Crowd Behaviour), DRDC CORA Technical Report 2007-06 (2007).

List of symbols/abbreviations/acronyms

AHP	Analytic Hierarchy Process
BRANDO	Breakpoint Analysis with Nonparametric Data Option
CAEn	Close Action Environment
CLS	Chief of Land Staff
CORA	Centre for Operational Research and Analysis
CSAF	Combat Simulation Assessment Framework
DEA	Data Envelope Analysis
DLR	Director of Land Requirements
DND	Department of National Defence
DRDC	Defence Research and Development Canada
DRSA	Dominance-based Rough Set Approach
EINSTein	Enhanced ISAAC Neural Simulation Toolkit
ELECTRE	ELimination Et Choix Traduisant la REalité
ERA	Evidential Reasoning Approach
ERC	Extraordinary RED casualties
HiLOCA	High Level Operations using Cellular Automata
ISAAC	Irreducible Semi-Autonomous Adaptive Combat
LER	Loss Exchange Ratio (RED casualties over BLUE casualties)
LFORT	Land Forces Operational Research Team
MAGIQ	Multi-attribute Global Inference of Quality
MANA	Map-Aware Non-Uniform Automata
MARCUS	Multi-criteria Analysis and Ranking Consensus Unified System
MCDA	Multi-Criteria Decision Analysis
MOE	Measure of Effectiveness
MOP	Measure of Performance
NL	The Netherlands
OR	Operational Research
PROMETHEE	Preference Ranking Organisation METHod for Enrichment Evaluations
RCS	Residual Combat Strength
ROC	Rank-Order Centroid
RS	Rank Sum Method
SMARTER	Simple Multi-Attribute Rating Technique Exploiting
TNO	The Netherlands Research and Technology Organization

Distribution List

Internal

Dr. Peter Dobias (hard copy, CD)
Dr. Kevin Sprague (hard copy, CD)
Mr. Gerald Woodill (hard copy, CD)
Maj Steve Bassindale (hard copy, CD)

DG DRDC CORA (e-mail)
Chief Scientist DRDC CORA (email)
Section Head, Land and Operational Command OR (email)

DRDC CORA Library (CD, hard copy)
LFORT (CD, hard copy)
LCDORT (CD)
CEFCOM ORT (CD)
CANSOFCOM ORT (CD)
CanadaCOM ORT (CD)

External

ADM(S&T) (for distribution) (CD)
DRDKIM 3 (CD)

Director S&T Land (CD)
DG DRDC Valcartier (CD)

CF College Library (CD)
Fort Frontenac Library (CD)

CFANS Library (CD)

DLCD (LCol Rettie) (e-mail)	Rettie.JM@forces.gc.ca
DLR (Col Lanthier) (e-mail)	Lanthier.JM@forces.gc.ca
DLSE (LCol Bassarab) (e-mail)	Bassarab.RR@forces.gc.ca
DLSE 4 (LCol Lefebvre) (e-mail)	Lefebvre.JAA@forces.gc.ca
DLR 5-3 (Maj Dufour) (e-mail)	Dufour.JJS@forces.gc.ca
DLR COORD 3 (Maj Roy) (e-mail)	Roy.S@forces.gc.ca

International

Mr. Patrick Cleophas
TNO Defence, Security and Safety
Information and Operations
P.O. Box 96864, 2509 JG
The Hague, The Netherlands

Document Exchange Manager (CD)
DSTO Research Library
Defence Science & Technology Organisation
PO Box 44
Pyrmont NSW 2009, AUSTRALIA

Michael Gillman (for dist’n and library) (CD)
Chief Technologist
Land Battlespace Systems
Dstl Integrated Systems
Room 31, Bldg A3, Fort Halstead
Sevenoaks, Kent, UK, TN14 7BP

Director, US AMSAA (CD)
ATTN: AMSRD-AMS-S)
392 Hopkins Road
APG, MD 21005-5071

Dr. James T. Treharne (CD)
OCA Division
Center for Army Analysis
6001 Goethals Road
Fort Belvoir, VA 22060-5230

Mr. John Hughes (CD)
HQ, TRADOC Analysis Center (TRAC)
Programs & Resources Directorate (PRD)
255 Sedgwick Avenue
Fort Leavenworth, Kansas 66027-2345

Ms. Carlo Fiamingo (CD)
TNO Defence, Security and Safety
Information and Operations
P.O. Box 96864, 2509 JG
The Hague, The Netherlands

Mr. Wouter Noordkamp
TNO Defence, Security and Safety
Information and Operations
P.O. Box 96864, 2509 JG
The Hague, The Netherlands

Dr David Galligan (CD),
Head Operations Analysis,
Defence Technology Agency,
HMNZ Naval Base Auckland,
Private Bag 32901, Auckland, New Zealand

Dr. Neville J Curtis (CD)
Research Leader Land Operations Research
75 Labs
Land Operations Division
PO Box 1500
Edinburgh SA 5111, AUSTRALIA

Dr. Jason Field (CD)
Land Battlespace Systems
Dstl Integrated Systems
Fort Halstead
Sevenoaks, Kent, UK, TN147BP

Mr. Patrick O’Neill (CD)
Chief, Combat Support Analysis Division
USAMSAA (ATTN: AMSRD-AMS-S)
392 Hopkins Road
APG, MD 21005-5071

Mr. Robert Barrett (CD)
Chief, International Activities
Center for Army Analysis
6001 Goethals Road
Fort Belvoir, VA 22060-5230

Mr. Bob Barbier (CD)
TNO Defence, Security and Safety
Information and Operations
P.O. Box 96864, 2509 JG
The Hague, The Netherlands

UNCLASSIFIED

SECURITY CLASSIFICATION OF FORM

(highest classification of Title, Abstract, Keywords)

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared e.g. Establishment Sponsoring a contractor's report, or tasking agency, are entered in Section 8).

DRDC CORA
Department of National Defence
Ottawa, Ontario K1A 0K2

2. SECURITY CLASSIFICATION (overall security classification of the document, including special warning terms if applicable)

UNCLASSIFIED – Unlimited Release

3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title)

Measures of Effectiveness and Performance in Tactical Combat Models (U)

4. AUTHORS (last name, first name, middle initial)

Dobias, Peter, Sprague, Kevin, Woodill, Gerald, Cleophas, Patrick, Noordkamp, Wouter

5. DATE OF PUBLICATION (month Year of Publication of document)
October 2008

6a. NO OF PAGES 39

6b. NO OF REFS 36

7. DESCRIPTIVE NOTES (the category of document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Technical Memorandum

8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include the address).

DRDC CORA / LFORT

9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)
N/A

9b. CONTRACT NO. (if appropriate, the applicable number under which the document was written.)
N/A

10a. ORIGINATOR's document number (the official document number by which the document is identified by the originating activity. This number must be unique to this document.)
DRDC CORA TM 2008-032

10b. OTHER DOCUMENT NOS. (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification.)
(X) Unlimited distribution
( ) Distribution limited to defence departments and defence contractors: further distribution only as approved
( ) Distribution limited to defence departments and Canadian defence contractors; further distribution only as approved
( ) Distribution limited to government departments and agencies; further distribution only as approved
( ) Distribution limited to defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in 11) is possible, a wider announcement audience may be selected.)
NONE

13. ABSTRACT (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), or (U). It is not necessary to include here abstracts in both official languages unless the test is bilingual).

Computer simulations are often employed by operational research analysts to evaluate the relative effectiveness of various combinations of military equipment and tactics (i.e., options) for specific tasks within a conflict scenario. If the simulation environment is realistic enough, one can rank the options based on how effective they are when used to complete the assigned objective. The ranking process requires that measures of effectiveness (MOEs) be designed to capture the essence of how well the goal was achieved for any particular option. In this paper, it is shown that the relative ranking of options can be disturbed by omitting options or adding options, dependent on the method used for valuing the MOEs. This has implications for those relying on ranked options as part of a larger decision making process – the omission of one option due to, say, post-analysis logistical, political, budgetary or supply concerns can upset the balance of the remaining rankings and lead to an inappropriate decision if left unchecked. We discuss some circumstances under which rank-order switching can occur. Two methods of valuing MOEs aggregated through weighted sums to produce option rankings are compared and contrasted: 1) a simple Relative to Best scheme, and 2) Valuing with Objective Scales. The latter is shown to be a better choice when rank-order switching is at issue. Furthermore, it is argued that, in general, only a few MOEs are necessary and that too many can lead to undesirable consequences. Moreover, measures of performance (MOPs) are put forward to capture secondary characteristics of the options that may come into play. Although they do not explicitly enter into option ranking, flagging potential problems early on can help identify options that might be eliminated post-evaluation.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified . If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Measures of Effectiveness
Measures of Performance
Combat Models
Wargames
Option Rankings
Multi-criteria Decision Analysis

Canada

DEFENCE R&D DÉFENSE

DRDC CORA

www.drdc-rddc.gc.ca