Best Practice for using Assessment Hierarchies in Operational Analysis – Principles and Practical Experiences

Graham Mathieson
Centre for Defence Analysis*
Defence Evaluation and Research Agency
Portsdown West, Portchester Hill Road, Fareham, Hants., PO17 6AD, UK
+44 (0) 2392 336359
glmathieson@dera.gov.uk

Abstract

The bedrock of military operational research for many years has been the use of combat models to convert measures of system performance into measures of force effectiveness. Quantifying effectiveness in the context of operations other than war, and taking account of the human and organisational aspects, has proved difficult using conventional modelling techniques. The need for multiple measures of merit and multiple decision criteria makes the use of assessment hierarchies very attractive to hard-pressed executives. There is also a trend, in these cost conscious times, to want cheaper and more common OA tools and methods across the full range of investment decision-making, from requirements capture, through design to investment appraisal. In all of these application areas assessment hierarchies appear to offer a relatively simple, highly visible and low cost means of assessing the value of complex investments. However, this appearance is dangerously deceptive.

The relatively uncontrolled and unrigorous use of assessment hierarchies, combined with the self-reinforcing features of facilitated judgemental methods, can lead to questionable advice to decision-makers. Many previous treatments of this subject have focussed on the details of judgement elicitation or mathematical manipulations, without fully addressing the larger issues of appropriateness and validity. This paper will discuss the principles and practice of the application of assessment hierarchies more rigorously. Drawing on recent study experiences in the areas of Intelligence and Information Systems, it will distinguish between estimating effectiveness and valuing performance, set out conditions for appropriate (and inappropriate) use of assessment hierarchies, and offer practical elements of good practice.

1. Introduction

The bedrock of military operational research, or operational analysis (OA), for many years has been the use of dynamic combat simulations to convert measures of system performance into measures of force effectiveness. The increasing challenge for OA has been to study the

effectiveness of investments in the context of operations other than war (OOTW), especially for integrated operational and non-operational information system investments, taking account of the human and organisational elements of investment. This challenge has emphasised the difficulty, if not impossibility, of quantifying effectiveness in these circumstances using conventional modelling techniques. Consequently, the use of static scoring systems such as multi-criteria assessment hierarchies has blossomed.

Command and Control (C2) problems, in particular, frequently require rich, multi-dimensional assessment covering both functional and non-functional aspects of systems. Decisions about information technology investment need to be taken in the context of consequential organisational change. This introduces social and political issues, which have to be synthesised with the technical aspects of the problem to achieve an adequately balanced assessment. The need for multiple measures of merit and multiple (often unquantifiable) decision criteria makes the use of assessment hierarchies and other multi-criteria analysis methods very attractive to hard-pressed executives.

Similar issues arise in the use of OA to support operational C2 in complex non-war-fighting operations. In these circumstances there is a need for OA to produce comprehensive, but also comprehensible, advice on the consequences of interventions which may be considered by Operational Command.

There is also a trend, in these cost conscious times, to want cheaper and more common OA tools and methods across the full range of investment decision-making, from requirements capture, through design to investment appraisal.

In all of these application areas multi-criterion analysis (MCA) appears to offer a relatively simple, highly visible and low cost means of assessing the value of complex investments. However, this appearance is dangerously deceptive, of which more later. The Centre for Defence Analysis has considerable experience in using MCA in a variety of contexts and this paper draws, in general, from this experience.

For the purposes of this paper the term MCA is used as a generic class of analysis including multi-criteria decision analysis (MCDA) and assessment hierarchies. The term 'assessment hierarchy' is used in this paper to cover a broad subset of possible formulations and uses of MCA. Simplistically, it covers off-line assessment as contrasted with on-line decision support. The distinction here is important, because it speaks to the validity of the judgements used to generate scores and weights within the analysis.

2. The distinction between decision support and off-line assessment

Classically, MCA has been seen as a device for decision-makers themselves to use (with analyst support and facilitation) to capture their preferences in terms of multiple, often conflicting, criteria [Goodwin & Wright, 1999]. The archetypal MCA is a hierarchical structure taking quantitative attributes at the bottom level and, through a series of judgementally derived value functions and

combination weights, providing an integrated view on the overall implications of options for sets of attribute values.

In this form of use, illustrated by figure 1(a), the model explicitly created by the analyst/facilitator need not be complete since the decision-maker is 'in-the-loop' of the analysis and will introduce reasoning and other factors in parallel with the explicit representation. The expressed model merely reflects the decision-maker's mind sufficiently to complement and enhance his own thought processes. Also, the explicit "answer" produced by the model need not be correct provided the process of model building has been sufficiently insightful for the decision-maker to be better placed to make his decision. The analyst's model is effectively incorporated into the decision-makers internal model of the problem rather than being used for problem solving directly.

Whilst not ideal from the analyst's point of view, this is not necessarily an issue for validity since the purpose of the analysis is to provide the decision-maker with evidence and insight rather than to obtain an answer. Ideally, the empowered decision-maker will file the MCA model at the end of the session and then make a decision (the MCA model forms part of the audit trail rather than being an automation of any part of the decision process itself). In this context, multi-criterion decision analysis (MCDA) is a powerful and effective method [Goodwin & Wright, 1999].

In providing analysis support to large, hierarchical organisations or to complex technical areas it is frequently the case that access to real decision-makers is limited. Instead analysis is included in a staffing process whose purpose is to produce evidence for presentation to decision-makers rather than directly supporting the decision-making itself. Within some (especially publicly funded) organisations there is also and institutional desire to base decision-making on evidence which is independent of advocacy and individual prejudice (even the prejudice of the decision-makers themselves). In such contexts, MCA still has a role to play, but in a different way from on-line decision support. Assessment hierarchies, as used in this paper, represent the form of MCA used in this off-line, evidence generation context. In place of decision-makers, assessment hierarchies typically use subject matter experts (SME) to judge the parameters of the hierarchy (e.g. relative weights) and, often, to select the categories which are scored.

As illustrated in figure 1(b), the model created to support off-line analysis must be complete, in the sense that it explicitly captures the decision-maker's value system. This is a difficult thing to do, and it is often the case that SME are used on the assumption that they share the eventual decision-maker's value system and, hence, can be relied upon to express preferences. The model created for off-line analysis must also be subject to questions of validation (i.e. testing to demonstrate fitness for purpose). Indeed, the SME who are used to populate the model must also be treated as input data sources, which need to be validated or certified as fit for purpose. Ideally, SME judgements should be elicited fully, subject to independent scrutiny and, if possible, validation through independent modelling.

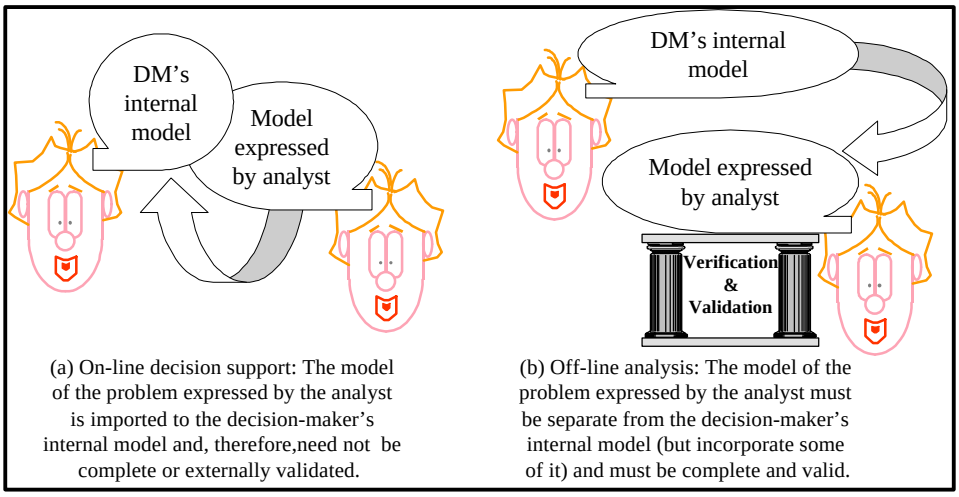

Figure 1: Illustration of on-line decision support and off-line analysis

In the on-line case, the decision-maker is being used as a source of preferences. His fitness for purpose is embodied in his empowerment to make decisions based on his preferences and his role-based authority. In the off-line case, the SME is being use because of his ability to assess the consequences of variation in inputs to the assessment hierarchy. SME preferences are not relevant; their fitness for purpose is derived from their expertise rather than their opinions or official role.

The boundary line between assessment hierarchies and MCDA is neither sharp nor unambiguous. SME are frequently used to make initial judgements, which are then reviewed and refined by decision-makers. However, this complication does not change the principle that an important boundary exists, one which impacts on the validity of the whole analysis.

A key test of when that boundary has been crossed is to consider the logical combinability of the criteria being evaluated. Where those criteria are logically combinable, then it is reasonable to call upon SME to assess that combination and to generate more aggregate criteria. There comes a point, however, when the criteria cannot be logically combined and one is forced to trade them against each other on the basis of personal (or institutional) preferences. At this point one has crossed over from assessment hierarchy to MCDA and SME are no longer the appropriate judges to use in subsequent aggregation. This division is illustrated in figure 2.

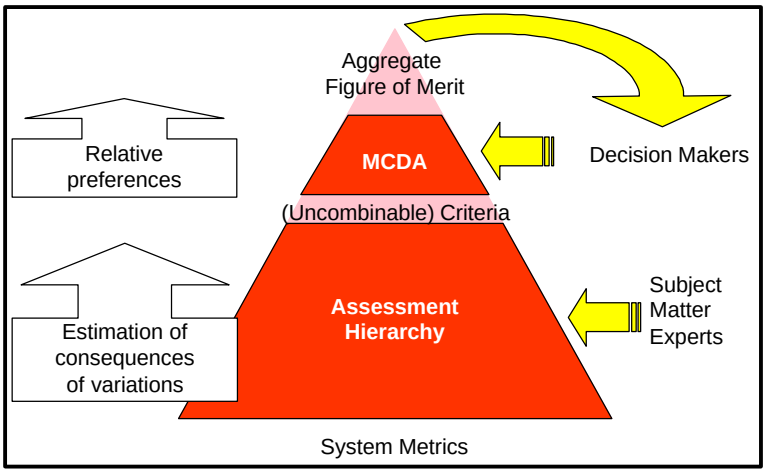

Figure 2: Illustration showing distinction between Assessment Hierarchy and MCDA

Just as SME and assessment hierarchies are not appropriate judgement sources and tools for dealing with logically uncombinable criteria, so decision-makers and MCDA are not appropriate for the treatment of criteria that can be logically combined. By implication, trying to use MCDA with SME is also inappropriate, although this is the format in which much application is done.

Using a simplification of the decision to purchase a car as an example, the distinctions between assessment hierarchy using SME and MCDA using decision-makers can be made clear. Consider that one wishes to buy a car and one is concerned with the following (simplistic) criteria:

- Fuel consumption, because of a need for economy;
- Bumper design, because of a need to avoid damage in a bump;
- Braking performance, because of a need to avoid bumps, and hence damage;
- Paint colour, because of a need for aesthetic appeal.

Since two of the decision criteria (bumper design and braking performance) are servicing the same need (damage avoidance) there is a logical way to combine them. Experts in road traffic accidents could be brought in to assess the likely severity of damage given certain levels of each criterion. However, there is no logical way for experts to combine a damage avoidance criterion with the aesthetic appeal of a given colour. This must be a matter for decision-maker preference.

Even this simple example has room for ambiguity, however. The combination of damage avoidance criteria with the economy criterion (fuel consumption) could either be seen purely as a matter for decision-maker preference, or it may be possible to create a logical combination based upon the economic value of crash damage in the context of whole life ownership costs. Figure 3 illustrates the example and indicates the two possible variants of the division between assessment hierarchy and MCDA.

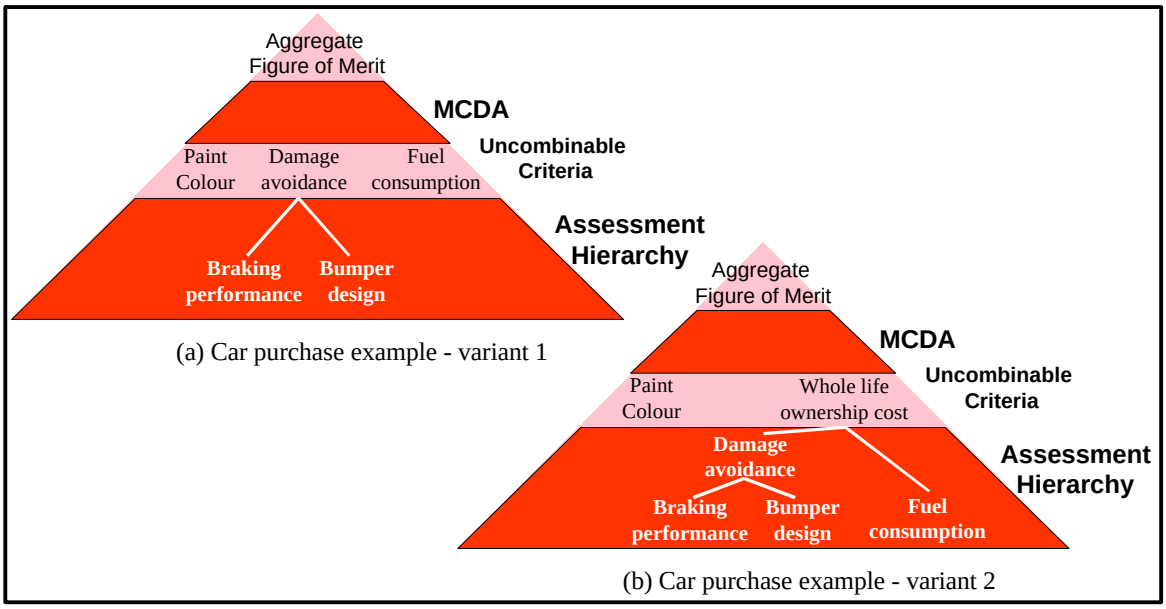

Figure 3: Simple example based on car purchase

The relatively uncontrolled and unrigorous use of assessment hierarchy methods can lead to very suspect OA and questionable advice to decision-makers. The positive and self-reinforcing features of facilitated judgemental methods often blinds participants to fundamental flaws in those methods when used to produce evidence for subsequent decision making by others. This paper will try to expose these flaws, their implications for assessment and appropriate responses by the analyst.

Many previous treatments of this subject [Goodwin & Wright, 1999; Abi-Zeid et al, 1998] have focussed on the technical details of judgement elicitation or mathematical manipulations, without fully addressing the larger issues of appropriateness and validity. This paper will discuss the principles and practice of the application of assessment hierarchies, and alternative analysis techniques, for dealing with problems where dynamic campaign modelling is either not available or not applicable. Drawing on recent study examples in the areas of Intelligence and Information Systems, it will distinguish between estimating effectiveness and valuing performance, and set out conditions for appropriate (and inappropriate) use of assessment hierarchies, with particular emphasis on practical alternatives where appropriate.

A key theme of the paper is that the rigour of analysis is fundamental in the off-line context. The next section discusses rigour and its importance for assessment studies.

3. The nature of rigour in assessment

Operational analysis, like all scientific disciplines, depends upon rigorous application to ensure reliability and to avoid the generation of misleading advice to decision-makers. Rigour is the *sine qua non* of scientific work and its importance cannot be overstated. Arguments that OA is not a truly scientific discipline, whilst having some merit, should not be used to excuse poor quality thinking or logical weaknesses. Rigour and good OA are inseparable.

However, rigour is often mis-understood in the operational analysis community, as evidenced by the adoption of many unrigorous methods and the unrigorous use of methods. In the author's experience, there is a widespread but mistaken belief that the concepts of rigour apply only to classical quantitative analysis and that such standards can (or indeed must) be relaxed when using so-called "soft" analysis methods. There is also a tendency to relax standards of rigour for studies that are short of time or money, based on an unscientific appeal to "pragmatism".

As a guard against such errors, the author devised a pragmatic definition of rigour, based upon using the word itself as a mnemonic, which was presented at a recent international conference [Mathieson, 1998].

Figure 4 illustrates the RIGOUR mnemonic, each element of which is described and justified below. Clearly, perfect rigour is an ideal not always achievable in real life, even in the hardest of scientific disciplines. However, the extent to which the ideal of rigour is achieved in an operational analysis study is strongly correlated with the extent of fitness of that study as a source of reliable advice to decision making.

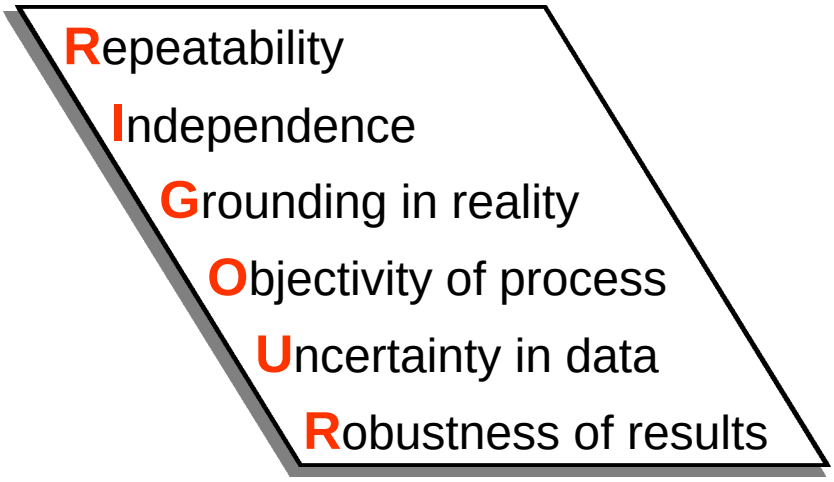

Figure 4: Illustration of RIGOUR mnemonic

Repeatability is at the heart of rigorous scientific method. If an analysis is not repeatable, given effectively equivalent initial conditions, then the decision-maker cannot rely upon its products. Repeatability, in this context, refers to the complete study, not necessarily its individual parts. For example, any analysis step involving the use of judgement is unlikely to be repeatable, even where the same judges are used a second time. However, where unrepeatable judgement is used as part of an analysis, the advice delivered to decision-makers can be made repeatable by managing the uncertainties the judgemental step introduces.

Independence of bias and prejudice is fundamental requirement for reliable analysis. Although difficult to achieve in a judgemental analysis, independence should not be left to chance. Even without deliberate intent, biased judgement is a constant threat to the reliability of assessment hierarchies.

Whilst much useful analysis can be done based on theoretical considerations, it is important that there be a significant **grounding in observed reality**, through experimental or field data, to ensure that conclusions are soundly based. It is interesting to note that the use of expert judgement is one commonly used way to ground operational analyses in reality.

Objectivity is a core principle of science. This aspect of rigour emphasises objectivity of the analysis process rather than data. Objective analysis of objectively derived data is an ideal. However, the presence of subjectively derived data (such as expert judgement) does not prevent rigorous analysis provided the subjectivity is documented and the treatment of the data is objective.

All practical analysis projects have to deal with a variety of **uncertainties** in observed or other input data, and in the analyst's understanding of the systems under study or in the study assumptions made. The use of judgemental data introduces additional uncertainty, often difficult to capture. Rigorous analysis requires the explicit treatment of such uncertainties so that consequent uncertainty in conclusions can be objectively assessed.

Given uncertainty in the conclusions of analysis, the advice given to decision-makers needs to be made **robust**, i.e. insensitive to the uncertainty. Reporting trends or differences which are not robust against changes in uncertain assumptions is unhelpful and potential dangerous for decision-makers.

The RIGOUR mnemonic can be used as a checklist to test the rigour of a proposed analysis. Appendix A provides a more detailed checklist with appropriate questions to ask and possible solutions to introducing rigour. Whilst there are no "magic bullets" for ensuring rigour, there are many practical steps to improving the rigour of assessment hierarchies. Some of these are discussed below, after a brief treatment of quantification, validation and cost issues.

4. Objectivity and quantification

It is a common myth that objectivity and quantification go hand in hand, and that qualitative analysis is intrinsically subjective. Clearly, there is Subjective Qualification, e.g. obtaining an opinion from SME. Also, there is Objective Quantification, e.g. generating a mathematical model of the problem from first principles, which can be independently verified. This is an ideal for OA, although it is rarely possible to achieve full objectivity because of the impact of doctrine and tactics on operational behaviour.

However, objectivity and quantification are not synonymous. Rather, they are orthogonal, as illustrated in figure 5. This allows one to consider Subjective Quantification, e.g. assessment hierarchies and MCDA, and Objective Qualification, e.g. rigorous reasoning (not just debate or invective!), graphical analysis where shape and relationship are the issue rather than position on scales, or data mining, where patterns of relationships between data are the issue. Whilst there is no reason in principle why the quantification in assessment hierarchies cannot be objective, they are typically used where objective numbers are not available and they are, therefore, an example of subjective quantification.

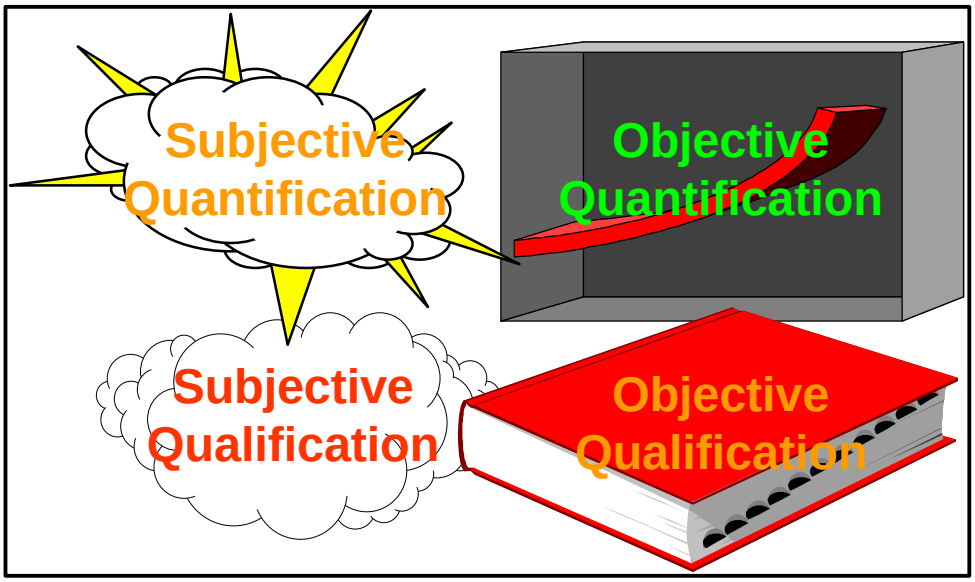

Figure 5: Quantification and objectivity as orthogonal aspects of analysis

The concept of rigour is not limited to quantitative analysis alone nor is it limited to purely objective analysis. The principles of rigour are about the reliability of the advice given to decision-makers. Reliable advice is the only sort a professional operational analyst should offer, whether or not it is based on objective or subjective reasoning.

5. The validation of subject matter experts as data sources

The judgements made by decision-makers or SME in MCA are the principle source of knowledge, and also the principle source of uncertainty and error. Like any element in an operational analysis study, the question of validity must be addressed.

In MCDA the necessary judgements must come from the decision-makers themselves. The criteria for validation of these decision-maker preference judgements are that the decision-makers are properly authorised to decide the issue at hand. The judgements in MCDA are not technical and, hence, the technical or operational expertise of the judges is not a principle concern.

In an assessment hierarchy the judgements required are estimations of the consequences of particular values for lower level criteria. These judgements require technical or operational expertise (depending on the criteria involved) and the appropriate judges are SME. The criteria for validation of SME are to do with their expertise rather than their rank or organisational role.

Given the different natures of decision-maker and SME judgement in MCDA and assessment hierarchies, respectively, it is important to treat the validation of judges differently. It is equally important to understand the scope of validity of SME so that one can ask legitimate questions. Guidance on validation of judges and asking legitimate questions is included below.

6. The real costs of judgemental methods

Assessment hierarchies are frequently used in situations where time and money are in short supply. There is a perception that they offer a quick and relatively cheap way to address complex problems by comparison with so-called "hard" analysis methods based on experiment and simulation. This perception is largely a consequence of two failures: a failure to conduct judgemental analyses in a rigorous way, and a failure to account for the full cost of using SME as data sources.

The true cost of judgemental methods, such as assessment hierarchies, only becomes clear in organisations using full cost accounting methods for all resources. A typical judgement panel comprising 10-15 participants (including facilitators, other stakeholders and technical support) could easily cost £10,000 in manpower charges alone. Experience suggests that at least two, probably more, meetings of a panel are needed to obtain reliable judgements and to achieve consensus and buy-in from SME and stakeholders. On top of this must be accounted the preparation and analysis effort required for assessment hierarchy design, pre-briefing and results interpretation.

Given the fact that SME are typically heavily loaded people, whose participation must be planned and scheduled months in advance, it is clear that rigorously implemented assessment hierarchies are neither quick nor particularly cheap when compared with other methodological options.

The choice of assessment hierarchy, therefore, should be made on grounds of appropriateness to the problem rather than on cost or time grounds alone.

7. Practical examples of how rigour can be improved

The author has been involved in a wide variety of studies over the past few years in which assessment hierarchies have been used. The following ideas are based upon attempts to introduce more rigour to the analysis conducted under these studies. All the ideas and recommendations are based upon practical experience, although the details of the studies involved cannot be released into the public domain. The key purpose in offering these practical examples of how rigour can be improved is to share good practice and to stimulate debate on best practice.

Guidance is offered under the headings of:

- Meaningful metrics;
- Validation of judges;
- Legitimate questions;
- Suitable structures;
- Coping with dynamic interdependencies;
- Treating uncertainty and sensitivity.

7.1 *Meaningful metrics*

Where assessment hierarchies involve a subjective quantification, the demands of rigour require that the manipulation of subjectively derived numbers be done as objectively as possible. A critical element of that objective manipulation is the need for commonly understood scales of measure for the quantities manipulated. The criteria within an assessment hierarchy, unlike those in an MCDA, must be associated with clear metrics. The author defines a metric as "a scale of meaningful extent". This definition emphasises the need for metrics to have definable scales and for values on those scales to carry meanings, which can be explicitly shared and understood between SME and decision-makers.

In MCDA, on the other hand, the numbers created by decision-makers typically represent their preferences rather than any measure of independently definable quantities. Care needs to be taken in MCDA to label aggregations of these preferences in a way that does not imply unjustified meaning to the scales. Care is also required, when combining an assessment hierarchy with an MCDA, to ensure that the top level metrics in the assessment hierarchy (which form the lowest level criteria in the MCDA) are understood in the same way by both SME and decision-makers.

Where criteria can be defined in terms of simple, real-world metrics this should be done. Often, criteria are too complex or abstract to permit simple metric definition. In these cases scales can be made meaningful by using text descriptions which indicate how scale extremes (and intermediate points) relate to a reality which the SME can interpret and understand.

7.2 *Suitable structures*

The structure of assessment hierarchies is an often neglected, but important contributor towards reliability. The standard triangular form is based on reducing a large number of criteria into a small number or even a single measure. In this way, the multiple criteria are summarised, usually through linearly weighted combinations.

Two major risks arise from such combination. Firstly, the compression, if taken too far, will produce quantities with little or no meaning, being only arbitrary aggregations of logically uncombinable metrics. In such cases, the assessment hierarchy process should be terminated and the uncombinable metrics should be re-classed as decision criteria and presented directly to decision-makers. MCDA can then be used to support decision-making directly.

The second risk of linearly weighted combination is its failure to capture non-linear relationships. In MCDA, where criteria are decision-maker preferences this may not be too much of an issue, but where meaningful metrics have been used in an assessment hierarchy, the linear weighting schema may often be invalid. The use of non-linear weights can overcome some of the problem, although their added complication could easily obscure the analysis and make it less easy to gain insights.

A better approach is to re-configure the hierarchy and the categories used to define it to minimise the extent of non-linearity. Careful selection of categories is also important in the case of dynamic interdependencies, which are discussed later.

Where the assessment is required to quantify the consequences of quite low level attributes or performance metrics on wider system performance or effectiveness, it is often useful build a structure which is progressive rather than aggregating. Figure 6 shows the architecture of such structure, which has been used successfully on a number of studies.

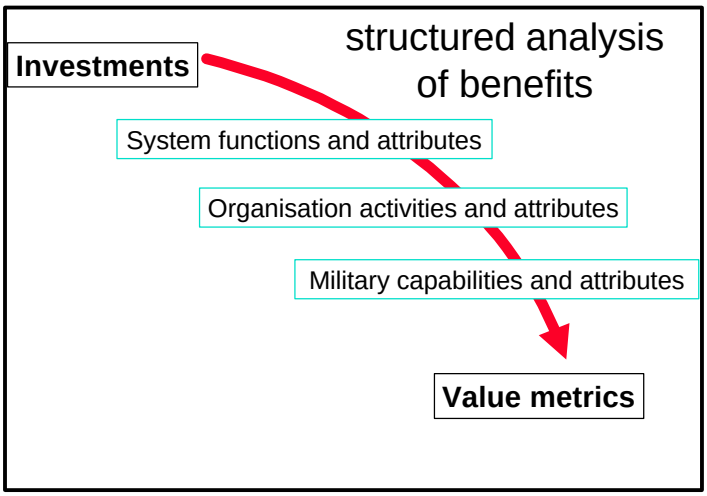

Figure 6: A progressive structure for multi-criteria analysis, based upon capturing the benefits of investments through a series of transitions at system, organisation and military capability levels.

The progressive structure seeks to transform measures of investment variables (e.g. option parameters) to system level attributes or measures of performance. The system level measures are then transformed into organisation level measures and, hence, to measures of military capability. These capability measures are finally transformed into value metrics, which can be used as decision criteria in an MCDA.

The progressive structure is usually implemented using a matrix formulation [see Mathieson et al, 1998]. This puts the focus on the progression from one stage to the next rather than on aggregation or summarising of measures. Causal mapping has proved a useful pre-cursor to aid the design of a progressive benefits structure.

7.3 *Validation of judges*

The need to ensure SME are valid judges has been discussed at length above. Given the often political nature of decision processes, and the fact that SME are typically also problem stakeholders, the concept of validation can be fraught. Practical steps to achieving confidence in the validity of SME include selection, accreditation and scrutiny.

The criteria for selection of judges for an assessment hierarchy should ideally be thought about in advance and recorded as part of the study concept of analysis. This will help to ensure a measure

of objectivity in the selection process and avoid the appointment of panel members based purely on their official position (or, more damaging, because they are stakeholders keen to ensure that the analysis process arrives at the "right" answer).

The credentials of judges should be explicitly recorded. This could involve no more than obtaining a curriculum vita (CV) or other record of experience. Evidence of credentials can be used to obtain stakeholder consensus on the constitution of a judgement panel, especially where one seeks to draw judges from a wider community than the study stakeholders.

Independent scrutiny of SME credentials is another important safeguard; particularly where study leaders may come under pressure to accept panel members for political reasons. Where the customer organisation already uses independent scientific scrutineers this facility should be exploited. Where this is not available, study leaders can co-opt independent reviewers and give them a scrutiny mandate.

7.4 *Legitimate questions*

Having constructed an assessment hierarchy which is suitable to address the problem at hand, and having identified SME who are valid judges for the scope of that hierarchy, it is critical to ask those judges legitimate questions. As mentioned above, the validity of SME as judges is based upon their operational or technical expertise. Often, in a broad problem, the expertise of any one judge is only relevant to part of the hierarchy. Asking SME to make judgements in parts of the hierarchy where their expertise is not strong is using them outside their scope of validity and is likely to produce a less reliable result, despite the potential benefit of a larger sample of judgements.

In one case study discussed in open session at a recent conference[1] it was reported that an assessment hierarchy had been used in parallel to a simulation-based analysis of the same problem. When the results (in terms of ranking investment options) disagreed the assessment hierarchy was re-visited. It was found that if only the judgements of those SME who were the lead experts in each area were counted, then the assessment hierarchy results changed to agree with the simulation-based ones. Although anecdotal, this experience is consistent with that in other studies, and supports the thesis that study reliability depends upon asking appropriate questions.

As well as ensuring that the scope of questions matches SME expertise, the content of the question must also be understood. This is intrinsically linked to the use of meaningful metrics as discussed above.

7.5 *Coping with dynamic interdependencies*

Accepting the need for meaningful metrics and appropriate, perhaps non-linear, assessment structures gives rise to limitations in the ability to select assessment hierarchy categories. This opens the way for the dynamic interdependencies between categories to become unavoidable.

[1] International Symposium on Military Operational Research, Royal Military College of Science, Shrivenham, UK, Sept. 1998

Rather than ignoring these interdependencies, it is possible to incorporate their effect using an interdependency matrix.

Mapping a set of categories onto itself using a matrix transform has been successfully used to make a first order correction for interdependency. Although not a perfect solution, this device can make the use of an assessment hierarchy more valid than would otherwise be the case.

7.6 *Treating uncertainty and sensitivity*

The demands of rigour include the explicit treatment of uncertainty. Whilst conventional sensitivity analysis, involving systematic variation of uncertain quantities, has proved effective for assessment hierarchies, a more complete test is possible using mathematical treatment of uncertainties. By eliciting estimates of the uncertainty in judgemental scores and weights, and by using these to calculate a mathematical uncertainty (typically a standard deviation) in intermediate and output metrics, it is possible to provide a robust statistical estimate of the reliability of the assessment. Figure 7 illustrates a standard linearly weighted sum equation and also the derived variance equation assuming input variances on both the scores and the weights.

$$OE^{Scen} = \frac{\sum_k w_k^{Scen} MoE_k^{Scen}}{\sum_k w_k^{Scen}}$$

$$s^2_{OE^{Scen}} = \frac{1}{\left(\sum_k w_k^{Scen}\right)^2} \sum_k \left[\left(w_k^{Scen}\right)^2 s^2_{MoE_k^{Scen}} \right] +$$

$$\frac{1}{\left(\sum_k w_k^{Scen}\right)^4} \sum_k \left[\left(MoE_k^{Scen} \left(\sum_k w_k^{Scen}\right) - \left(\sum_k w_k^{Scen} MoE_k^{Scen}\right) \right)^2 s^2_{w_k^{Scen}} \right]$$

Figure 7: Illustration of a typical weighted sum equation from an assessment hierarchy and the derived variance equation, assuming input variances in both scores and weights.

The mathematics can become somewhat complicated and this is often seen as a reason for scepticism. However, since variance equations are a straightforward derivation of the basic equations, there is no reason to treat them as any less valid than the basic ones. Indeed, if it is asserted that the variance equation is not valid then it must logically be concluded that the basic equation is equally invalid.

By using an explicit mathematical treatment of uncertainty, sensitivity analysis can be made more complete and reliable, and less prone to subjective bias

8. **Conclusions**

This paper has identified serious potential flaws in the application of assessment hierarchies, which can lead to unrigorous OA and the risk of unreliable advice to decision-makers. Despite the difficulty of applying the principles of rigour in the presence of subjective judgements, the paper has drawn on practical experience, to demonstrate that steps can be taken to improve the rigour of assessment hierarchies. The OA community is invited to develop and apply good practice in this increasingly important aspect of decision support.

9. **References**

[Goodwin & Wright, 1999] Paul Goodwin and George Wright. *Decision Analysis for Management Judgement*. 2nd Edition, John Wiley and Sons, 1999

[Abi-Zeid et al, 1998] Irene Abi-Zeid, Micheline Belanger, Adel Guitouni, Jean-Marc Martel, and Khaled Jabeur. *A Multicriteria Method for Evaluating Courses of Action in Canadian Airspace Violation Situations.* Presented at the Fourth International Symposium on Command and Control Research and Technology, Nasby Slott, Sweden, June 1998.

[Mathieson et al, 1998] Graham Mathieson, Anneliese Handley, Yvonne Rippeth. *Valuing Investments in a Digitized 'Systems Of Systems'*. Presented at the Fourth International Symposium on Command and Control Research and Technology, Nasby Slott, Sweden, June 1998.

[Mathieson, 1998] Graham Mathieson. *Rigorous Subjectivity - How will OA survive the "soft" revolution*. Presented at 15th International Symposium on Military Operational Research, held on September, 1998, at the Royal Military College of Science, Swindon, UK.

Appendix A

A1. Introduction

This appendix contains guidance on the application of the RIGOUR mnemonic to ensure a study will be repeatable, independent, grounded in reality, objective in process, dealing explicitly with uncertainty and robust against that uncertainty. For each element of the mnemonic, questions are presented, which have been found useful in identifying areas where rigour could be improved. Also presented are some archetypal problems associated with each area of rigour, and possible solutions that have proved useful in the past.

A.2 Is the study Repeatable?

Question: Given the "same" input and constraints will the analysis produce the "same" output?

- If yes, is the similarity real and substantial?
- If not, what prevents or inhibits replication?

Question: Do the resources impact on repeatability?
Question: What can be done to improve repeatability?

Problem: Uncontrolled inputs
Solutions:

- Control inputs
- Exploratory analysis

Problem: SME judgement
Solutions:

- Replace humans
- Use whole population
- Use multiple samples
- Use peer review

Problem: Analyst judgement
Solutions:

- Replace with logic
- Reduce analysis scope
- Use multiple samples
- Use peer review

A.3 Is the study Independent?

Question: Is the analysis independent of prejudice/bias?

- If yes, is the independence demonstrable?
- If no, what are the sources of prejudice/bias?

Question: Are these sources auditable?

Question: Could they be replaced with unbiased sources?
Question: How could the bias be detected/compensated?

Problem: Customer/Stakeholder prejudice
Solutions:
- Explicit elicitation of views
- Challenge from analyst
- Separation from analysis
- "Devil's Advocate"

Problem: Prejudice in inputs
Solutions:
- Alternate sources
- Critical scrutiny

Problem: Analyst prejudice
Solutions:
- Peer review
- "Red teaming"

A.4 **Is the study Grounded in "reality"?**

Question: Are the inputs and constraints "real"?
Question: How does the analysis link to "reality"?
Question: What "unrealities" exist in the analysis output?
Question: How can these be removed/compensated?

Problem: Incomplete data
Solutions:
- Limit analysis scope
- Extrapolation by modelling

Problem: Incomplete model
Solutions:
- Limit analysis scope
- Use sample situations

Problem: Unvalidated model
Solutions:
- Use anecdotal evidence

A.5 **Does the study have an Objective analysis process?**

Question: What inferences are made in the analysis?
Question: What judgements must the analyst make?
Question: How methodical is the analysis process?
Question: Can the judgements be replaced by logic?
- If not, can they be isolated as inputs/constraints?

Problem: Judgemental inferences
Solutions:
- Avoid judgemental steps
- Use multiple samples and consistency checking
- Isolate judgements and treat as inputs

A.6 **Does the study treat Uncertainty explicitly and is it Robust?**

Question: Are there express or implied uncertainties in the inputs, constraints or process?
Question: Does the analysis process explicitly treat these uncertainties?
- If yes, what treatment is used?
- If no, what prevents treatment?

Question: Do the outputs explicitly express uncertainty?
Question: Are the outputs significantly sensitive to uncertainties in the inputs, constraints or process?
- If yes, what could be done to avoid this sensitivity?

Problem: Defined uncertainties
Solutions:
- Reduce output sensitivity
- Present output uncertainty

Problem: Ill-defined uncertainties
Solutions:
- Avoid sources of uncertainty
- Reduce output sensitivity
- Present input uncertainty