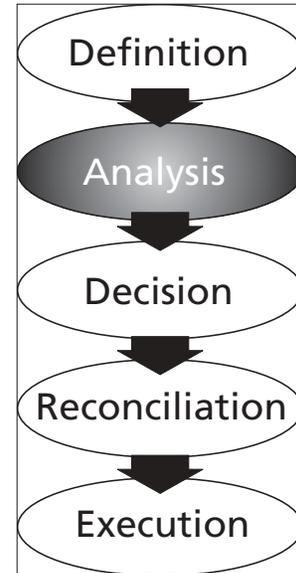


# ANALYSIS CONCEPTS: EFFECTIVENESS

*Facts are stubborn things; and whatever may be  
our wishes, our inclinations, or the dictates of our passions,  
they cannot alter the state of facts and evidence.*

John Adams



**A**FTER WE ARE SATISFIED WITH THE RESULTS OF THE DEFINITION phase, we begin analysis. Our primary goal in the Analysis Phase is to gain sufficient knowledge to meaningfully differentiate among alternatives. Most of the knowledge we seek concerns their effectiveness, their cost, and our uncertainty about the quality of the information we have about each choice. In defense resource allocation, analysis is the coin of the realm; other organizations are unlikely to take our proposals seriously unless we can back them with demonstrably robust analysis. Thus, we will address the standards we use to identify what we want to measure to compare alternatives and their likely consequences.

We may require research before we can begin analysis. In our framework, "research" is collecting original data and taking measurements whereas "analysis" is examining and interpreting data and measurements. We cannot conduct good analysis without sound data; therefore we may choose to be involved with the analyst's methods for data collection as well as with the tools he or she selects for evaluating the data.

Analysis almost invariably requires us to use models to organize our thoughts and evaluations. Models vary from the very simple, e.g., a ratio; to very complex theater warfare simulations, as we will see in the next series of chapters that cover the Analysis Phase. We will begin by addressing the most important constituents of models in the next few chapters, then we will discuss models themselves, and we conclude the phase by demonstrating how models are used in force-on-force and policy analysis.

## Action Officers, Decision Makers, and Analysts

As decision makers in DoD organizations, we seldom conduct our own formal analysis of complex problems.<sup>1</sup> Either our in-house analytical unit completes it or we contract with professional analysts. The decision maker is responsible for providing or approving his organization's

1. Major joint and service staffs have resident analysts, usually identified on their organization charts in the J-8 Directorates or as Analysis and Simulations staff assistants to the Commander-in-Chief or Service Chief. Sometimes we execute our own analysis; the Commanding Officer of a unit or base may not have the need or resources to execute a contract for expert analysts.

guidance to the analyst. He often delegates routine oversight to an action officer; indeed the action officer may be the instigator of the ideas that require analysis. Between them, they must provide the analyst with their military judgment, particularly in areas that are intuitive, operational, and experiential. Allowing analysts to proceed without the involvement of our action officers, or without the decision maker's approval of the analytic objectives, greatly increases the risk they will make serious mistakes. If we neglect to provide guidance to analysts, they will create their own, for better or worse.

Our relations with the analyst should be collegial, but we must take his or her background and different perspective into account as we proceed in this phase. Earlier, by our careful construction of analytic objectives in the Definition Phase, we notified the analysts that we are very interested and will be involved in their efforts. Now we seek to combine the powerful mathematical tools of the analyst with the operational experience, judgment, and intuition of military decision makers to sustain our rational approach.

## **Types of Analysis**

One of the first things the analyst and we must agree upon is the kind of analysis that will best achieve each analytic objective. There are three basic types of analysis: Exploratory, Cost-Risk-Effectiveness, and Causal. We decide upon the type of analysis now because it will influence the nature of our modeling later.

### **EXPLORATORY ANALYSIS**

Exploratory analysis examines alternatives that are in the early stages of development. During the mission needs and concept development stages of defense acquisition, we cast a wide net because we are looking for the best of all possible solutions. Because we are forecasting environments and encouraging creative, often non-traditional alternatives, we have a large amount of uncertainty and we do not expect very much detail from exploratory analysis. We must examine our assumptions from the definition phase very carefully, sometimes treating them as variables. The results of exploratory analysis are often controversial, so we must structure these studies clearly and exactly, particularly where we have made key assumptions. We should be able to comfortably explain the logic behind them upon demand.

### **COST-RISK-EFFECTIVENESS ANALYSIS**

Cost-risk-effectiveness analysis is the most common type of analysis in DoD; we use it almost universally to evaluate procurement options. Its purpose is to differentiate among problem-solving alternatives, e.g., to select a design for a major weapons system, to allocate funds among competing program alternatives, or to revise the roles and missions of active and reserve forces. When we execute cost-risk-effectiveness analysis, the problem is usually well defined and bounded, and often the alternatives already exist. Therefore, cost-risk-effectiveness analysis generally takes an engineering or mathematical approach. It, too, is hostage to the worthiness of its assumptions.

### **CAUSAL ANALYSIS**

We use causal analysis to determine why something happened in the past, how a previous action created the state we find in the present, or why actions we take now will create results we desire in the future. Causal analysis—establishing cause and effect—is central to making policy decisions, such as discovering why accident rates have increased, how best to conduct basic training,

or whether a pay raise or the provision of more recruiters would best increase the number of new enlistees. We want our analysts to rigorously separate facts from values and conduct causal analysis dispassionately. Our values may have entered the decision process in the Definition Phase, but we do not want the analyst to include his or her own subjective opinions unless we so specify.

## Selecting Alternatives

We may know the alternatives for solving the problem before we start decision making or we may develop them during the Analysis Phase. When we have the alternatives in advance, that knowledge may help us select criteria and build models that will best expose whatever important differences exist among them. Foreknowledge of the alternatives also indicates the likely range of values we can expect as we evaluate them, saving time and energy by limiting the scope of our analysis, i.e., if we know there are miniscule differences between certain aspects of the alternatives, we do not need to measure them. Nonetheless, our analysis must still be sufficiently general to accept a new, unforeseen alternative and compare it to the options we already have. Indeed, a pitfall of knowing the alternatives in advance is that we may design our model to emphasize differences between the alternatives although these differences may be trivial to the analytic objective. Worse, we may inadvertently favor one alternative for reasons outside the bounds of the analysis by seeking to emphasize differences. Because of these concerns, we always leave open the possibility of generating the alternatives later in the Analysis Phase.

Our set of alternatives should exhibit the following characteristics:

- Breadth
- Viability
- Neutrality

The alternatives must span the scope of possible solutions of the problem, including the extremes as well as the middle of the range of alternative solutions. Extreme solutions may include disruptive technologies that may have enormous spillover effects on our organization and others; they may require delicate handling and we should discuss them with the decision maker to determine whether they are within the boundaries of this problem's solution set.

When we have a continuum of alternatives, we select representative alternatives that permit study and enable clear choices; e.g., most studies of Overseas Troop Strength add increments of 25,000 soldiers as they build alternatives. They identify the capabilities of each force level so decision makers can see how much capability each increment adds. The actual alternative may not be a multiple of 25,000, but the decision maker will have a clear sense of capabilities after reading the analysis.

Additionally, each alternative we study must be a viable solution and meet our minimum requirements; we will not include throw-aways.<sup>2</sup> If an alternative is unacceptable, we should identify whether it can be improved to meet our standards, e.g., a city may be willing to upgrade, at its own cost, the hotel services at its piers or its mass transit to encourage the Navy to homeport ships there. We must be very careful whenever we dismiss an alternative for not meeting our standards; its proponents may ask us later to justify its exclusion.

---

2. There is an apocryphal story about Secretary of State Henry Kissinger and President Richard Nixon. During a crisis with the Soviet Union, Secretary Kissinger presented the President with three alternatives. "Mr. President," he said, "first, we may begin global nuclear warfare immediately; second, we may capitulate abjectly. I think we should explore the third option more fully."

We strive to shed bias from our alternatives, therefore we describe each in a similar manner with the same level of detail. We test each neutrally, according to the same standards and under similar conditions. One of the traits we value highly in analysis is its empiricism, the fairness we get by testing options and comparing results in a dispassionate manner. A fair competition among ideas is essential to discovering which is best for solving our problem. Besides, each alternative in defense resource allocation will have its proponents, many of whom we will encounter in the Reconciliation Phase and our analysis must be persuasive in that phase; it can only be persuasive if it is thorough and unbiased.

#### **WHEELS VICE TRACKS: THE ARMY'S MEDIUM-WEIGHT COMBAT VEHICLE ALTERNATIVES**

When the Chief of Staff of the Army, General Eric Shinseki, unveiled his vision for the Army's transformation to a medium-weight force on October 12, 1999, he was addressing concerns that the Army's heavy forces, although highly capable, were too heavy to move to the fight quickly enough. To reduce the size and weight of the equipment the U.S. Transportation Command would have to lift between and within theaters, he stated that his vision included a new family of wheeled armored vehicles that C-130 intratheater lift aircraft could haul and that would replace tracked vehicles.<sup>3</sup> Most observers understood his desire to lighten up the Army, but it was unclear to many why General Shinseki specified wheeled vehicles in his introductory comments. A variety of senior Army leaders has since said that a family of wheeled vehicles was one likely expression of the Chief's vision and that his comments should not be taken so literally as to exclude the possibility of a new tracked family of medium-weight armored vehicles; all options were on the table and because a wheeled option would break with tradition the Chief chose to emphasize it.

During the following winter, the Army held a vehicle competition for nine contractors with 35 different systems. The only U.S. manufacturer of the three that submitted tracked alternatives was United Defense LP; they introduced reengineered, modernized variants of their venerable M-113 armored personnel carrier. Following the demonstration, the Army revised its draft Request For Proposals with some lower performance standards to reflect what they had observed in the trials and to encourage as many contractors as possible to continue participating. United Defense accused the Army of relaxing its requirements because it realized wheeled vehicles could not meet the performance standards while tracked vehicles could.<sup>4</sup>

Army officials denied any bias, but skeptics could not help but note the Army had already leased 46 wheeled light armored vehicles from Canada for use by its two interim brigades as they test new operational concepts central to Army transformation. Senior Army leadership again denied that they had ruled out tracked vehicles, but many of their briefing materials gave exactly that impression in the Spring of 2000. (At an Association of the U.S. Army meeting during 16-19 February 2000, the U.S. Army Training and Doctrine Command's organizational graphics for the battalions of the interim brigades used the symbol for motorized infantry (wheeled vehicles) vice

3. For example, the current M-1A2 Abrams tank weighs 70 tons and can be carried one at a time only by strategic lift aircraft like the C-5 and C-17. It is too heavy for most bridges and maneuvers with difficulty in congested terrain and on narrow roads. By comparison, the maximum weight for the new vehicles is 19 tons.

4. Sean D. Taylor, "Wheels Vs. Tracks: Is Shinseki Moving Too Far, Too Fast?" *Army Times*, 28 Feb., 2000: 12.

mechanized infantry (tracked vehicles)<sup>5</sup> The Army has since created a new symbol that combines both.)

Inspired by concerns that the Army was moving too rapidly toward unproven capability—and at least in part by Congressmen from districts who manufactured track vehicles—Congress held hearings to explore how the Army was selecting these interim vehicles. As a result, while Congress funded the medium armored vehicle procurement program for fiscal year 2002, the Army, despite its protests, was made to hold side-by-side tests of the leased wheeled vehicles against M-113s before full production could begin.

The Army received 20 vehicle proposals on 6 June 2000. Over the next four months, they evaluated 17 of them and tested samples at Aberdeen Proving Grounds using the parameters from their Operational Requirements Document (largely parallel to the revised Request For Proposals that the Army issued the prior spring). After receiving final proposals on 6 October 2000, the Army awarded the contract for the new family of wheeled vehicles to a consortium of General Dynamics and General Motors of Canada on 8 November 2000 based on trade-offs in the following areas.

- Suitability to support operations with the new Interim Combat Teams
- Transportability requirements
- Quality of the training support package
- Technical requirements for the different variants, e.g., characteristics of armament
- Crew protection.

United Defense LP objected to the award decision soon after it was announced, based (they said) on the Army's failure to consider its own requirements, i.e., United Defense LP contends their vehicle is 50% less costly, can be delivered sooner, and that it meets all the Army's performance specifications unlike the wheeled vehicle selected. They also assert that the Army's communications with the contract-winning General Motors of Canada/General Dynamics consortium "substantially exceeded the nature and extent of information conveyed to United Defense LP" and that the competition was pro forma, evidenced by briefing slides prepared before November that incorporated substitute vehicles for the Mobile Gun System variant. (The substitution is significant, posits United Defense LP, because it means that the variant would not be available as required by the timeline specified in the Operational Requirements Document, i.e., the Army was pre-approving a deviation before the contract was awarded... and the delivery schedule was one of the criteria for selecting among manufacturers.)<sup>6</sup>

Conversely, the Army contends the wheeled vehicle family provides overall superior performance, according to its weighted criteria, than United Defense's reworked M-113s. Because of the protest to the General Accounting Office, the Army issued a stop work order on 5 December 2000 while the General Accounting Office reviewed the award.

The Army leadership stood by its decision. In December 2000, Secretary Caldera stated that he believed the selection process would stand up to the Government Accounting Office review

5. Ibid.

6. Kim Burger, "UDLP Offers Additional Evidence of Army Bias in Favor of LAV III," *Inside the Army*, 15 Jan. 2001:1. UDLP contends the wheeled vehicles failed to meet performance requirements for ammunition storage of ready rounds, separation of ammunition and crew, internal noise, braking, and, for the mobile gun system, battlefield sighting indexing requirements, amongst other shortfalls. They further claim many of the required improvements, such as the mortar variant and swim capability, are high risk and that the armor has to be removed to make it C-130-transportable and question how life cycle costs were arrived at without reference studies.

and that the new administration, like Congress, would find the medium armored vehicle program compelling. General Shinseki has called for armor traditionalists, concerned about the lesser firepower and protection of new vehicles, to stifle their dissent and, "If you chose not to get on board, then that's okay, but get out of the way."<sup>7</sup>

How much of this controversy could have been avoided if General Shinseki had not appeared to exclude tracked vehicles from his vision? Many, in addition to United Defense LP, still feel tracked vehicles are viable alternatives for transforming the Army. For major defense decisions with many stakeholders, the range of alternatives must cover the range of possible solutions without the perception of arbitrary exclusions or we may expect those stakeholders to react to protect their equities.

Additionally, the alternatives must be viable in terms of the problem definition and it is the executive decision makers in DoD and their staffs who will approve the standards against which they are measured. How well were the analysts who designed the field trials in the winter of 2000 listening to their decision makers if they established overly demanding performance standards? The perception they created was that after the trials, when the wheeled vehicles did not fare well, they changed the requirements to make them viable. In reality, the new standards may well be the right ones, but now the issue is clouded.

Finally, there is the issue of neutrality or fairness. General Shinseki let his preference for wheeled vehicles and against the M-113 in particular be known early and clearly to his subordinates. How or whether that affected their decisions we cannot know, but United Defense LP perceived enough bias to raise objections that must be taken seriously—they resulted in the stop procurement order. Seldom can anything good come from promoting a particular alternative without robust analysis to explain rationally why this choice is favored.

The Army leadership may very well have made the right decision to purchase wheeled vehicles. The Light Armored Vehicle III family may best achieve Army transformation goals and therefore be best for the long-term health of the Army. But senior leaders' actions before the formal decision process engaged—especially the Analysis Phase—invited emotional responses. Were tracked vehicles effectively excluded from the beginning? Were the standards changed to ensure a wheeled vehicle choice? With better preparation before the decision and robust analysis, these questions can be answered to the satisfaction (if not the desires) of all the stakeholders during reconciliation; without them, expect controversy and disharmony.

As we consider or build alternatives, we know that the program or policy that is executed after a decision may not be, in the literal sense, any one of the alternatives we constructed in the Analysis Phase. Alternatives may be modified or even combined after analysis to incorporate the strengths of one to compensate for the weakness of another; the executive decision maker is usually in the best position to make these adjustments and usually sees their impact more clearly than the analyst. For example, an auxiliary airfield might be time-shared with civil aviation to defray costs and provide military access to a longer runway than is otherwise available or affordable. Extensive alterations to the alternatives may require that we conduct another analysis.

---

7. Thomas E. Ricks and Roberto Suro, "The Wheels Turn in Army Strategy: Transformation to Cut Tanks' Role" *Washington Post*, 16 Nov., 2000:1.

Where policy decisions are concerned, we recognize that the more steps or phases an alternative has, the less likely it is to be executed as the originator intended. Each succeeding phase of implementation is actually an opportunity to modify and shape the alternative further. Alternatives that are not phased have an all-or-nothing character to them—often the case in procurement decisions—and they present greater risk of failure for an organization committing to them.

## Attributes, Criteria, and Measures

Now we will examine the characteristics of alternatives and decide which to evaluate. Consider we are facing a decision, we have three alternatives that may or may not be effective, and our task is to select the one that best accomplishes our goal. We will do this by predicting the consequences of adopting each alternative and comparing those consequences to one another based on a set of standards we choose. Which standards we select are crucial in the Analysis Phase.

We begin our selection by noting that every alternative we encounter, however simple or complex, is or will be composed of attributes, that is, its entire family of qualities, characteristics, and distinctive features. Size, cargo capacity, weight, speed, and availability rate are typical attributes of vehicle alternatives. Equity, happiness, morale, and quality of life are typical attributes concerning policy choices. When we must select among alternatives, some attributes are more important than others because they are more relevant to the way we defined the problem and the way we expressed the decision objective. In our framework we call these more important attributes criteria; they are the standards upon which we will base our judgments and preferences among the alternatives. The most important subsets of criteria (and the ones we will assess) we call *Measures*, many of which we group in two categories, *Effectiveness* and *Cost*. Effectiveness is the ability of an option to achieve an outcome we desire. Cost is the amount and rate at which alternatives consume resources.

Figure 3-1 illustrates that Measures of Cost (MOCs) and Measures of Effectiveness (MOEs) do not constitute the entire universe of criteria. Schedule, risk, equity, and availability of resources are examples of criteria we measure for a procurement or policy selection that are not strict descriptors of cost or effectiveness. Our collection of criteria may include any number of measures, but cost and effectiveness are almost always relevant to defense decisions.

When we can evaluate a criterion numerically, we have a quantitative or objective measure. One alternative wheeled armored vehicle has a maximum speed of 65 mph, another 50 mph. A number that we can measure to a very fine degree differentiates them. But we can use numbers and combine them differently to measure alternatives. We do not purchase eggs in the same way that we do armored vehicles. All the eggs in a carton of a dozen do not weigh the same, nor are they the same size. We could measure both egg criteria (size and weight) in every carton we buy, and then buy the carton we find preferable, but there is a simpler way. We buy them based on a qualitative measure; instead of a number we assign eggs to a category that we understand, "large" or "medium," and then choose the carton that will satisfy us. In this

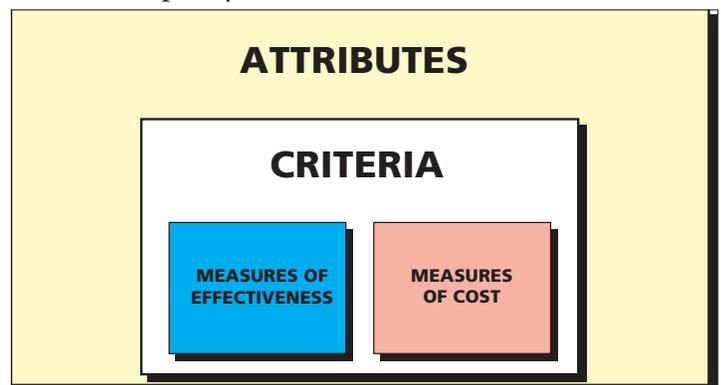


Figure 3-1. Characteristics of Alternatives

case, the qualitative measure (grade) is assigned based on underlying quantitative data; as we will see later, this is not always the case.

Most analysts are more comfortable measuring characteristics that are quantifiable. Left to our own preferences, many of us, too, will gravitate toward a numerical measure; we have less risk of being "wrong" about something we can count. Despite these tendencies, there is nothing inherently superior about quantitative measures, and nothing to suggest that qualitative measures are inferior or less rigorous. The character of the problem must drive the measures we choose.

## Selecting Criteria

Our next important task in the Analysis Phase is to identify the attributes we need to measure to support the needs of the decision maker and designate them as criteria. (Later, with the analyst, we will determine how to best measure these criteria.) We select attributes to be criteria based on their ability to indicate whether important differences exist among the alternatives, and, if that is the case, the degree of those differences. Knowingly or not, we generated indications of important criteria during the definition phase. Brainstorming—listing every possibility on a wallboard—is a good beginning for turning those indications into well-defined, candidate criteria. In general, we seek criteria with the following characteristics:

- A direct connection to the analytic objective
- Inclusiveness
- Precision
- Measurability
- Uniqueness
- Discrimination

There are many attributes that distinguish between procurement and policy options that are not germane to the decision and therefore they are not good criteria. Good criteria evaluate the performance of alternatives in the real world in a manner linked to the analytic objective. That is, they help us evaluate the alternatives in a way that matters. For many acquisition programs, the criteria for concept studies are derived largely from the operator's Mission Need Statement that first identified the requirement or deficiency.

We naturally prefer a single inclusive criterion that covers a large portion of the desired analysis to several discrete ones so that we can simplify our data collection and display. We carefully and precisely describe each criterion to eliminate room for interpretation by the analysts or the participants in the decision. We prefer direct, quantifiable measurement to reduce error, even as we understand that such perfection is not always possible.

Each criterion should measure something unique and different from the others. "Double counting"—directly or indirectly measuring the same attribute twice—is usually undesirable, but in exceptional cases may be appropriate. Finally, the criteria should reflect value added for exceeding the minimum requirement to help us discriminate between alternatives. If an option must meet a specific minimum requirement to be eligible for consideration, but there is no value for exceeding that minimum, then that attribute is not a good criterion. It may be an important attribute, a benchmark that each alternative must satisfy, but that importance is not synonymous with being a good criterion. Requirements and thresholds are Go/No-Go filters;

they disqualify an option from further consideration unless the alternative brings itself up to the required standard. Criteria help us compare value beyond minimum requirements.

The more criteria we choose to measure, the more expensive and lengthy the Analysis Phase will be. There is a point of diminishing returns beyond which our attempts to refine the alternatives further are not worth the effort. We may even proceed to the point of over-specification, in which we define so many criteria so tightly that we cannot create any alternative that satisfies them all. Over-specification reduces the effect of individual measures when we weight them in a model. The Definition Phase helped us identify the point of diminishing returns when we evaluated the importance and urgency of this decision to our organization.

After we identify a range of potentially useful criteria, we identify the relative value of each criterion to the decision and determine which criteria we actually want to measure. Ideally, we would like a set of criteria we can measure directly, in quantitative terms. Unfortunately, objective attributes (the quantitative ones) are often far less important than subjective attributes (the qualitative ones). We must guard against choosing criteria that are easy to measure but less relevant to our decision, and we should not shy away from attributes that are difficult to measure.

<b>CRITERIA</b>	<b>MEASURES</b>	<b>EXAMPLES OF HOW WE CAN MEASURE</b>
<b>COST</b>	<b>UNIT COST</b>	<b>CURRENT OR CONSTANT DOLLARS</b>
	<b>PERSONNEL</b>	<b>PAY, MAN-HOURS, MANNING LEVELS</b>
	<b>TOTAL OWNERSHIP COST</b>	<b>CONSTANT DOLLARS</b>
<b>SCHEDULE</b>	<b>FIRST UNIT DELIVERY</b>	<b>CALENDAR DATE</b>
	<b>INITIAL OPERATIONAL CAPABILITY (FIRST UNIT)</b>	<b>CALENDAR DATE OR DATE RELATED TO THREAT</b>
	<b>FULL OPERATIONAL CAPABILITY</b>	<b>CALENDAR DATE OR DATE RELATED TO THREAT</b>
<b>EFFECTIVENESS</b>	<b>MAXIMUM SPEED</b>	<b>MACH, KNOTS, MPH, FEET/SEC</b>
	<b>MAXIMUM RANGE</b>	<b>MILES, KM; EMPTY OR WITH WEAPONS</b>
	<b>WEAPONS LOAD</b>	<b>NUMBER AND VARIETY</b>
	<b>STEALTH</b>	<b>RADAR CROSS SECTION, HEAT SIGNATURE, NOISE LEVEL, SIZE</b>
	<b>SIZE</b>	<b>DIMENSIONS, FT2, FT3, DECK SPOTS, CONTAINER-EQUIVALENTS</b>
	<b>WEIGHT</b>	<b>POUNDS, TONS, DISPLACEMENT</b>
<b>RISK</b>	<b>MATERIAL</b>	<b>% OF COMPOSITES, FIRST APPLIED USE OF MATERIAL</b>
	<b>TECHNOLOGY</b>	<b>NEW OR PROVEN, NUMBER OF TESTS BEFORE PROTOTYPE</b>
	<b>PRODUCTION</b>	<b>NUMBER OF TESTS BEFORE PRODUCTION, % NEW OR UNIQUE COMPONENTS</b>
	<b>POLITICAL SUPPORT</b>	<b>OPERATOR REQUIREMENTS, COMPETING FUNDING REQUIREMENTS, JOB DISTRIBUTION</b>

Table 3-1. Examples of Measuring Criteria.

## Assessing Criteria

In the process of identifying and selecting a set of criteria, we must assess the degree to which all of these measurements help us evaluate alternatives that satisfy the decision objective. We examine them, individually and as a group, through three lenses: Validity, Reliability, and Practicality.

### VALIDITY

Validity is the degree to which our criteria adequately predict, measure, or illustrate to the decision maker the important differences among alternatives: *Are we measuring the right things to support making this decision?* Are we gathering enough information to make a rational decision? Does each criterion add to our understanding of the alternatives? The set of criteria must somewhere address every aspect of the analytic objectives; when applied to the alternatives, they must help us select. We use analysis to simplify reality; by assessing validity, we ensure that we do not over-simplify or become distracted from the analytic objective.

Put another way, validity is the degree to which we are able to identify what we want to measure. We accept that usually one criterion will not reflect every facet of the alternatives' behavior. There is no single, ultimate criterion we can use to measure the performance of a fighter aircraft. We settle for what we *can* measure: components of the idyllic measure of "fighterness." The most common way to improve validity is to measure more attributes, i.e., to add more criteria, thus, at least in theory, we can move closer to the perfect set of measures that encompasses everything.

A related way to improve validity is to use surrogates for things that are difficult to measure directly. For example, we may estimate aircraft survivability by determining the number of enemy radar types that our electronic counter-measures suite can counter.

On a more abstract level, suppose that we are tasked to evaluate several alternative compositions for U.S. nuclear forces and that the different alternatives' deterrent effect is one of our criteria. This is a tough task, because deterrence is something that happens in the minds of our adversaries (if it happens at all) and is not a directly measurable physical attribute of our nuclear forces. One way to cope with this problem is to use several more directly measurable attributes (e.g., the quantity and size of warheads, their accuracy, and their ability to launch after an enemy attack) as surrogates. If we have reason to believe that our adversaries consider such attributes in deciding whether they are deterred from certain actions, then we can reasonably use these attributes as surrogates for our "deterrent effect" criterion.

The degree of validity for each criterion varies with the problem definition. Consider two decisions about Navy surface combatants. Ships have a huge family of attributes and therefore an equally large set of potential criteria. If we are deciding between builders' proposals to select the design for the next generation destroyer, cost, warfighting effectiveness, technology risk, delivery schedule, habitability, and maintainability are all highly valid criteria. If we are deciding among ships to send on a contingency deployment, a different but overlapping set of criteria is more valid; we may list cost, warfighting effectiveness, level of training, materiel readiness, and command climate as our criteria. Both decisions deal with ships, but they require different criteria; we evaluate each attribute's validity for use as a criterion differently for each decision.

Finally, we check our set of criteria again against our analytic objective to ensure we have not left out an important attribute or that we have not accidentally double-counted the same attribute. We examine each criterion individually to ensure it contributes meaningfully to our understanding of the alternatives.

**MEASURES OF EFFECTIVENESS: THE SYSTEMS APPROACH AND CONVOY PROTECTION<sup>8</sup>**

We often use a Systems Approach as a tool during analysis, consciously or otherwise. In our lexicon, a system is the full array of elements (people, equipment, and processes) that operate together to perform a mission, create a desired state, or achieve an objective. Each system has input and a process that produces an output. We measure a system's output while the system is in operation to modify or adjust the process controls—feedback—to keep us heading toward the goal. The process becomes a loop, not just a linear path from input to output. We use the systems approach in mechanical applications ranging from driving an automobile to guiding missiles. We use similar feedback mechanisms in policy analysis, from efforts as diverse as dieting and exercising, to reducing the national debt and improving student population test scores.

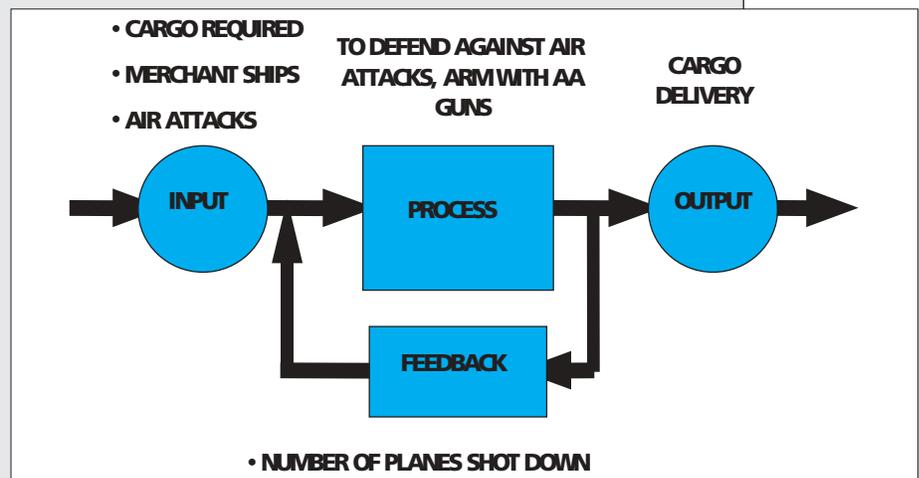
Although widely applied, the systems approach is not applicable for every analysis. If we can easily identify systemic elements in our problem and its potential solutions, such an approach is an easily grasped and appropriate tool for creating simple models of processes. Consider the usefulness of the systems approach as a convenient structure and display mechanism for the following problem, and for evaluating the validity of some analysts' choice of a measure of effectiveness.

During World War II, the British armed their merchant ships in the Mediterranean Sea with anti-aircraft guns to fight off enemy aircraft. These guns were in short supply, expensive, and badly needed elsewhere. After a few months of operation on the ships, the British government ordered an analysis to decide whether the guns should remain on the ships. Using a systems approach, the analysts' model looked like this:

After considering this information, the British government decided to remove the guns from the ships and redirect them to more gainful employment. Fortunately, before the decision was implemented, someone pointed out that the wrong measure of effectiveness was used to provide the feedback.

The objective was to protect the merchant ships, not to destroy enemy aircraft—that could be done more efficiently in other ways. The guns, however, forced the attacking aircraft to maneuver more, release their bombs at higher altitude, and otherwise impaired the bombers' accuracy. The

guns were serving their purpose because more cargo was arriving. When the MOE (feedback) was framed correctly against the decision objective, the analysts discovered only ten percent of the gun-protected ships were sunk during air attacks while twenty-five percent of the unprotected ships were lost. Based on this revised analysis, the British left the guns on the ships.



8. Adapted from *Methods of Operations Research* by Philip M. Morse and George E. Kimball (Cambridge, Mass.: MIT Press, 1951).

**RELIABILITY**

Next, we evaluate our set of criteria for reliability. In our lexicon, reliability is the accuracy and consistency of a measure. *How well can we measure?* We must specify to the analysts the resolution of the measurements we require, including the units of measure and the fidelity or degree of accuracy we desire for each measurement. We must tell them how much measurement error is tolerable. When we measure repeatedly, under identical circumstances, we should get the same, consistent results.

We select criteria with less engineering precision (resolution) to support decisions by the Secretary of Defense than we would for an acquisition program manager. Similarly, we are usually less specific during concept studies and become more granular as we approach production parameters. Do we need to know airspeed in terms of Mach, knots, or feet per second? Is greater precision of value to the decision maker or is it a distraction? The resolution we need to distinguish between alternatives in a meaningful way is the level of detail we should measure and display; this may be considerably less than the resolution we can possibly measure.

Ideally, we opt for criteria that we can measure directly, in isolation, and without disruption by the act of measuring in order to minimize error and improve repeatability. Measurement error is ever present; we can compensate for some measurement errors easily, such as that in a gauge that misreads by 10 psi across its entire range. Detecting or adjusting for other measurement errors is difficult, especially as our criteria become interrelated or more subjective. A missile's failure to intercept a target within lethal range (miss distance) is a typical test criterion. Test firing intercept failures may be due to hardware casualties in a sensor in the missile seeker head, problems in the missile's software, or its control system; but we cannot isolate the fault unless we measure at each control point. In the worst case, errors may cancel each other out and our miss distance may be small enough to score as an intercept even though the missile did not work properly. Miss distance does tell us something we want to know; it has high validity. Miss distance, if we measure it simply as distance from the target, has low reliability because we do not know how subsystem measurement errors interacted with one another or how they individually affected overall system performance.

Surveys present our most difficult reliability challenge, a circumstance wherein reliability is on a par with validity. When we commission surveys of personnel to research policy options, the quality of the questionnaire is central to the reliability of the results, so we test the questionnaire before we use it in a survey. By issuing the questionnaire, then interviewing the respondents and identifying why they answered the way they did, we gain confidence that responses from the general population mean what we think they do. If the questions are poorly worded, the respondents' answers will be skewed, compromising reliability. Reliability also suffers when we do not get a sufficiently large or random sample of the target population; we should not permit self-selection by respondents because the most vocal members of the population are seldom the most representative of the general population. Reliability suffers further when survey respondents do not answer truthfully, i.e., without necessarily meaning to be deceitful, some people answer questions based on how they think they should feel rather than how they actually feel. An old saw says voters speak from the heart but vote from the pocketbook; a similar process can happen with answers to surveys.

Sometimes, when we are assessing complex or intuitive behavior, there are limits to the amount of knowledge we can obtain about causal factors or future actions. When we compare two manufacturers' products based upon their anticipated mean time between failures, we can examine historical data from the companies, we can review their assumptions for calculating

projected failure rates in the past, but we cannot know if their estimated rates are correct (reliable) until the product is built and tested. Even then, we still have uncertainty. Will the mass-assembled products behave like a lab-built prototype, or even like each other? We will explore this further in Chapter 5, "Uncertainty and Risk."

We desire repeatability or consistency, the same results under the same circumstances, in our measurements of criteria. We may not be able to reproduce the same circumstances for each measurement, just as downhill skiers race on a slightly different course on each run. The more subjective our criterion is and the more dependent it is on the actions of others, the less repeatable it becomes. The mood of a respondent to a survey question may alter his choices on any given day. In a conflict simulation, the enemy response may vary depending on which analyst is playing Red, affecting Blue's optimal strategies and outcomes dramatically.

We can improve reliability in several ways. First, we can measure the same criterion in more than one way. If we decide unit manning is a criterion for selecting which of several like units to deploy, we can examine overall strength, manning levels for mid-grade Noncommissioned Officers and above (leaders), and projected rotations during the deployment. Together, they provide a better picture than any one measure alone, and they all concern manning. Should we make each a criterion by itself? We could; it depends on the situation and the level of detail the decision maker wants when we model this problem. More likely, we will measure these three items to justify our evaluation of manning and display only manning in our briefing; if asked, we are prepared to explain our evaluation—the proverbial back-up slides.

A related way to improve reliability involves taking advantage of surrogates that we chose in searching for valid criteria. To see this point, recall the example of using various physical attributes of nuclear forces as surrogates for those forces' deterrent effect. To the extent that we can measure those attributes objectively, we can improve reliability. (Of course, we can only increase reliability if the attributes we measure are also valid measures of what we care about.)

We can enhance reliability by improving our measurement methodology. Improved measuring equipment with more sensitive instruments, more complicated models, or a more isolated test environment will lead to more accurate measurements. If we are using computer simulations, we can run more iterations. If sampling is an important technique, then we increase the sample size.

## PRACTICALITY

We evaluate our criteria from a third perspective, practicality. *Does the knowledge we gain from measuring justify the resources that we consume?* Practicality in this application does not mean "easily used or applied," rather, are our criteria too costly to measure and use? Resources can be money, time, personnel, equipment, and the like—anything we consume to measure a criterion. Practicality involves a sense of the first two evaluations: Do we have enough validity and reliability? Can we afford more?

For example, there have been an enormous set of attributes that helped us to compare between the two prototypes of Joint Strike Fighters proposed by the two competing contractors. After we order them in terms of validity, practicality tells us how many are enough. We may be able to measure each to an extraordinary degree, and thus improve reliability, if we are willing to consume a large amount of resources to do it. Practicality considerations tell us whether we should. An example of a low level of practicality is a set of criteria that is both highly valid and re-

liable, but that requires more time to collect the data than is permissible to meet the deadline for this analysis.

Practicality may involve a tradeoff between validity and reliability. We can often improve both validity and reliability by consuming more resources. To conserve resources, we can choose more abstract, less costly, surrogate measures as long as they have enough validity and reliability to support our decision. Practicality constrains our analysis by tying it to resource limitations commensurate with the importance of the decision to our organization.

### VALIDITY, RELIABILITY, AND PRACTICALITY INTERACTIONS

Having discussed validity, reliability, and practicality at some length, we should reflect on how they interact and how they are distinct from one another for they are recurrent themes that permeate our decision-making framework. Logically, we view and evaluate them sequentially. Validity is often our first and most central concern. When we analyze a problem and its alternatives, we are analyzing an abstraction of the real world (a model) and validity is our evaluation of how well we have transferred reality to that model. Without valid criteria, there is little point in proceeding further; the most exquisite reliability cannot compensate for measuring the wrong things.

Reliability, then, is our next concern: poor reliability can lay to waste a perfectly valid model in several ways. If we measure poorly or inappropriately, our data is skewed and our analysis becomes tainted. Flaws in reliability may be more insidious than validity problems because they are not necessarily obvious when the results of analysis are documented and displayed. We must insist the analysts show us how they measured before we can have confidence in their results.

Practicality can be viewed as resource allocation between validity and reliability. Often, we would like to measure more criteria and often we would like to measure an individual criterion with more precision. Practicality is the balance between the two: are we measuring so many things that our reliability suffers too greatly from spreading ourselves too thinly? Are we omitting an important criterion because we are measuring the others in more detail than we need? Are there insufficient resources to support this analysis and bring it up to the standards we need to achieve acceptable validity and reliability? Most of our practicality problems can be resolved with more personnel, time, or money. Our practicality evaluation tells us whether such expenditures are worthwhile in the context of the decision and the organization.

Finally, validity, reliability, and practicality are not absolute qualities that are either present or absent; *criteria do not pass or fail a "Validity-Reliability-Practicality Test."* Simple statements that declare, "A criterion has high validity because it reflects the real world" are not helpful; we must consider all of a criterion's characteristics before we are satisfied. There is an important, deliberately subjective quality to our assessment of these traits—we evaluate validity, reliability, and practicality from our decision maker's perspective. Therefore, we are not surprised when other organizations and other decision makers select or emphasize different criteria. Because of practicality constraints, the decision maker must approve the decisions we make about criteria, including which imperfections in validity and reliability are tolerable.

## Measures of Effectiveness

We know that effectiveness is the ability to produce a result we desire, but there is usually no single measurement that will encompass all of the attributes we desire to measure in a set of alter-

natives. Speed and tire pressure are both attributes of an aircraft. One is clearly more important to the success of a fighter aircraft than the other. Speed contributes toward success in combat; it therefore becomes a criterion as a measure of effectiveness. Tire pressure is an attribute of the tire and ultimately of the plane, and it is a requirement for the proper functioning of an aircraft with inflatable tires. It is not a criterion that helps us evaluate how well an alternative satisfies the analytic objective. We do not care what the tire pressure is as long as it is adequate. It is possible to imagine a case in which every MOE of several aircraft is exactly the same, thus tire pressure emerges as the tiebreaker, but such circumstances are rare. (They might be more common in shopping for less expensive items. Color might be the discriminator among several suitcases that all have the same capacity). If such a condition occurs, we might ameliorate it by more accurate measurement of more important criteria.

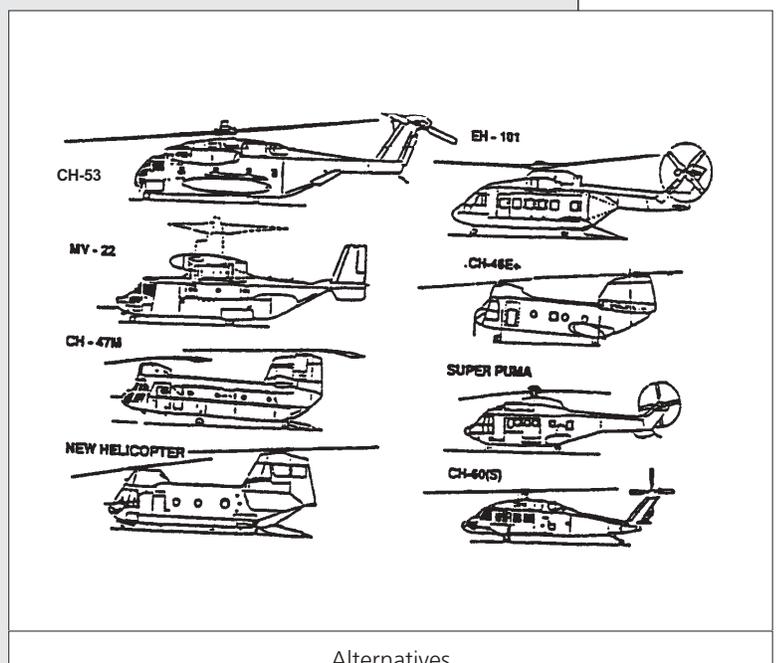
Criteria for procurement decisions thus tend to cluster around MOEs such as speed, range, capacities, weapons loads, combat power, lethality, and survivability. Note that we can measure some of these by direct means; others may require sub-measures to evaluate them meaningfully. We can measure the speed of a vehicle directly. The survivability of an armored personnel carrier may require a compilation of other measures like thickness of armor plate, profile, self-defense capability, redundancy of systems, etc. Note again that "self-defense capability" may require further specificity, such as the performance characteristics of an offensive capability such as a machine gun.

#### CASE STUDY: THE ANALYSIS PHASE—ALTERNATIVES AND MEASURES OF EFFECTIVENESS USMC MEDIUM-LIFT REQUIREMENTS: THE V-22 OSPREY AND HELICOPTERS

Congress and the Department of Defense specified many of the aircraft options the Institute for Defense Analyses considered in their analysis of medium-lift alternatives, but they gave IDA license to explore other alternatives as well. Thus, Congress and DoD wanted IDA to consider the broadest range of options; they are shown above. As a result, IDA added the New Helicopter, a notional design based on a Boeing 360.

For each aircraft, IDA created two fleets for their assessment, one sized on the Marines' requirement to lift the assault force in two waves of aircraft (502 V-22s) and the other sized on the projected expenditure by DoD for replacement helicopters (356 V-22s). In each case, they calculated the cost of the V-22 fleet and used the

same funding level to buy the various helicopter fleets. All of the fleets were viable in the sense that they were plausible alternatives, however by fixing cost at these two levels, IDA did not evaluate whether a helicopter fleet less costly than the DoD proposal in Level II could achieve the mission, i.e., they used DoD's planned expenditure as a lower boundary. No reasonable options were



Alternatives

9. Simmons, L.D. et al, *Assessment of Alternatives for the V-22 Assault Aircraft Program*, Executive Overview, Institute for Defense Analysis, 1991, p. 12.

excluded from the study; in fact IDA underplayed some significant additional costs to keep the smaller helicopters in play.

No matter which aircraft is selected for procurement, the Marines' existing fleet of 76 CH-53E heavy-lift helicopters must augment the medium-lift fleet. Some of the smaller helicopter fleets would require *additional* CH-53Es. The smaller helicopters cannot lift certain "medium" weight cargos such as vehicles and artillery. As a reference point, at the time IDA did their study the Marines had 224 CH-46E medium-lift helicopters and 76 CH-53E's. Table 4 below reflects the size of the fleets at the two cost levels IDA considered:<sup>9</sup>

<b>Marine Corps Medium-Lift Assault Aircraft</b>	<b>Number at Cost Level I (\$33B FY88)</b>	<b>Number at Cost Level II (\$24B FY88)</b>
<b>V-22</b>	<b>502</b>	<b>356</b>
<b>New Helicopter</b>	<b>634</b>	<b>450</b>
<b>CH-47M</b>	<b>673</b>	<b>527</b>
<b>CH-60 (S)/CH-53E+</b>	<b>287/347</b>	<b>240/283</b>
<b>CH-46E+/CH-53E+</b>	<b>317/336</b>	<b>251/258</b>
<b>Puma/CH-53E+</b>	<b>330/322</b>	<b>260/246</b>
<b>EH-101/CH-53E+</b>	<b>252/335</b>	<b>200/256</b>

#### **MEASURES OF EFFECTIVENESS**

Congress and DoD together identified eight missions that they tasked IDA to evaluate. IDA evaluated the role of the aircraft in each mission area and explored the comparative performance of each aircraft fleet using the following MOE:

- *Amphibious Assault* (Move Troops and Equipment Ashore). IDA's MOE was the percentage of the assault force lost while building a 3:1 force superiority during a vertical assault. They used survivability of the different aircraft in the assault role as a proxy. Using aircraft speed, design, and size, IDA evaluated how likely enemy air defenses were to shoot down the aircraft under a variety of conditions, e.g., day, night, rolling and flat terrain, various air defense weapons. The defending force was a Soviet-style, Third World Motorized Rifle Division.

- *Sustained Operations for Logistics Support* (Move Troops and Equipment to Support Combat Forces Ashore). IDA compared the number of equivalent payload sorties flown in a 30-day period, based on aircraft reliability rates, payload, and speed for the different fleets of aircraft.

- *Hostage Rescue/Raid* (Insert and Extract Marine Rescue or Raiding Force and Hostages). For this mission, IDA evaluated the maximum distance from the objective a raid could be launched and, separately, how long it would take to reach an objective from a distance of 275NM, the V-22's most distant possible launch position. The helicopters had to have their ships close toward the objective before they could launch.

- *Overseas Aircraft Deployment* (Move to Overseas Theater and Transport Deployed Marine Force to Combat Positions). IDA assessed the number of C-5 sorties required and how long it would take to deliver a brigade's share of each fleet to an off-loading Maritime Pre-Positioning

10. Since 1990, the Marines have adopted Operational Maneuver From The Sea as their operational concept and it calls for Over-The-Horizon amphibious assault, incorporating the V-22 to land the vertical assault echelon from up to 50 NM off-shore.

Squadron or to the Marines' pre-positioned brigade equipment set in Norway. They also evaluated how long it would take the aircraft to deploy and tactically reposition combat troops.

- *Combat Search and Rescue* (Recover Downed Air Crews). IDA evaluated the percentage of rescues each type of aircraft could affect within two hours of a crash based on the distance of the survivors from the launch platform.

- *Special Operations* (Insert and Extract Special Operations Forces). Clandestine Special Operations often require aircraft to over-fly hostile territory at night, therefore IDA compared the fleets based on the number of missions that each could complete in darkness during nights of varying length.

- *Counter-Narcotics* (Trail Courier Aircraft and Boats, Deploy Law Enforcement Personnel). IDA evaluated the area to which each aircraft could respond in three hours and at maximum range without refueling.

- *Anti-Submarine Warfare* (Detect and Attack Enemy Submarines). IDA compared the V-22 fleet's capability using dipping sonar to detect submarines approaching the battle force to that of the Navy's S-3 patrol plane fleet (with other sensors).

See Appendix 3 for the results of IDA's analysis of each MOE.

**Validity.** IDA used a plethora of labels to measure the same thing in all eight missions: speed. This is a classic example of how seemingly different criteria can, in fact, be different representations of the same thing. Cycle time, area searched, time over an area, and the like are different measures of speed. This is why much of the IDA analysis seems repetitious.

Although we normally seek criteria that are unique, is the use of non-unique criteria justified in this case? Yes. Speed is a dominant criterion in each of the missions. The V-22 is more effective because it is faster; it is also more costly, as we shall see. Again, the crux of the decision is whether the additional effectiveness derived from the V-22's higher speed is worth its cost. With the IDA study, the validity question we should really ask is whether each scenario is truly representative of medium-lift aircraft employment: our standard question becomes, "Did we measure the right thing in the right context?"

The Marines validated the assault scenario, the most important medium-lift mission by far. It drives the overall size of the medium-lift fleet.<sup>10</sup> Survivability is an appropriate proxy for estimating how fast combat power will build up. Looking at how well each aircraft supports Over-The-Horizon assault was critical, and one could argue (despite the Congressional and DoD guidance), it is the only scenario that really merited evaluation. The sustainment scenario is based on how many sorties each aircraft can generate vice how many sorties and how much equipment the Marines require for support. This makes the measure of sortie rate questionable in terms of validity because the superior performance of the V-22 may not be necessary to achieve the mission, i.e., it may be over-capacity.

The Hostage/Raid scenario starts with the amphibious ships at the V-22 launch point and includes the steaming time for the ships to close launch points in the helicopter response times. To judge the validity of this MOE, one must examine the historical record for instances in which operations were delayed or canceled because of the additional ship transit time and then look at our current and projected needs. For example, the Marines have shown how the aborted Desert One raid and Non-Combatant Evacuations could have been executed more easily with V-22s. Our validity question is whether the 275 NM scenario, based on the operational range of the V-22 vice real world data bases and planning scenarios, will happen often enough in the future for it to be used as the test case in this study. If most operations will begin 1500 miles from the objective, the relative

response time difference is much smaller between different kinds of aircraft. If the predominant circumstance is that the ships are already nearby, then again the response time difference between aircraft types is quite small. We can tell that IDA's chosen scenario favors the V-22, but we cannot tell with the information available whether that kind of scenario is itself sufficiently valid.

The self-deployment scenario shows a clear advantage to the V-22. Less need for high demand supporting strategic airlift is important—and the earlier arrival of the V-22 to move troops is markedly better than the helicopter options... provided the 250 C-141 sorties of the Fly-In Echelon of the Marine Expeditionary Brigade arrive in time for the Marines to be transported by the V-22s.

For the non-USMC missions, speed is still the dominant criterion IDA used to compare aircraft options. Where range is concerned, the V-22 flies further because it flies faster every hour it is in the air, a significant advantage over helicopters. For Combat Search and Rescue, speed is indeed of the essence and its validity is strong for estimating success. For the long-range Special Operations missions, IDA assumed the assault force started at a great distance from the objective, and they assumed that more Special Operations are better. But the V-22 may again represent excess capacity: are more Special Operations required and are planners limited by the current inventory of helicopters?

For counter-narcotics operations, the response times from cueing to aircraft arrival in order to trail boats and aircraft or to move agents to a site is a highly valid criterion for an individual mission, much like for Combat Search and Rescue. We must ask, however, whether there are circumstances under which it would be more advantageous to have two less capable aircraft rather than one V-22.

Submarine detection and localizing (vice area searched based on speed of the aircraft jumping between dip points) is the most valid way to compare anti-submarine warfare systems because it is the most difficult chore in the detect-to-engage sequence. All the aircraft alternatives carry similar sensors and weapons.

**Reliability.** IDA measured their MOE well, using existing data for aircraft characteristics where available and they scrutinized projected aircraft characteristics from contractors carefully. IDA used military judgment from the Joint Staff and services to evaluate the subjective elements of the study such as the scenarios and missions, thereby improving the reliability of their analysts' estimates. The main reliability issues again revolve around the scenarios; did IDA measure aircraft performance accurately and consistently?

For the assault scenarios, IDA ran hundreds of iterations using the different fleets under varied simulated conditions to build a very large database. Field-testing the V-22 was not possible; however, the Marines had data based on helicopter-landed assault forces that IDA extrapolated to build the simulator runs. Scenario construction in terms of terrain, environmental conditions, and density of air defense along flight routes must all be realistic in order for the results of the simulation runs to be highly reliable; in this case they were as good as possible in 1990. The only way to improve reliability further in the all-important assault mission would have been for IDA to construct additional scenarios with a greater variety of opponents.

The outcome of the sustainment scenario depends upon the time between failures for the aircraft, i.e., how many round trips can each aircraft make with how much cargo before they go down for maintenance? The failure rates of the yet-to-be-built aircraft had to be estimated. IDA doubled the contractor's estimate, yet their calculations were still optimistic for a new (high risk) technology; in the IDA study, the V-22 was still more mechanically reliable than the advanced

technology helicopters. What happens if the V-22 fails at triple the projected rate? How did the contractor estimate the failure rate in the first place? Historical research may reveal a trend between aircraft manufacturers' predicted failure rates and their actual failure rates. IDA could use such a factor as a better multiplier than simply doubling the contractor's estimate.

For the Raid/Hostage Rescue and Overseas Deployment missions, the reliability of the study is very high: we can predict the mission transit times and the aircraft are equally affected by environmental factors. The reliability of the measures for the Combat Search and Rescue and long-range Special Operations scenario is also high because it was calculated on the basis of the speed difference between the options, a straightforward mathematical process. For the counter-narcotics mission, the calculations of area coverage are similarly very reliable. For the Anti-Submarine Warfare mission, IDA's figures for detecting submarines are questionable because we do not have an explanation of how they calculated them.

**Practicality.** IDA took a very pragmatic approach to this study because they had to complete it quickly. They maximized their use of existing force-on-force models and data from previous studies and researched when they found them lacking. For example, earlier studies did not consider survivability in the assault scenario. They balanced knowledge gained versus resources consumed extremely well, achieving very high levels of practicality.

Validity and reliability for the assault scenario are in balance; improvements to either would be costly and time consuming beyond their worth. It is appropriate that it consumed the majority of IDA's resources; improving validity and reliability for the other scenarios by consuming more resources is not very worthwhile unless the scenario or MOE is grievously flawed.

For the sustainment scenario, estimating a better factor for anticipating failure rates (described above) to improve reliability, or doing sensitivity analysis using a variety of failure rates was probably worth the investment. The most important improvement to the Raid/Hostage Rescue and Overseas Deployment scenarios would have been for IDA to determine whether the scenarios they used are truly representative of how we anticipate these missions unfolding in the future. If most missions, raids, and rescues do not fit the IDA scenario profile, then we need a larger or different family of scenarios. IDA should have reviewed the theater Commanders'-In-Chief Operational Plans that will tell them quickly whether the Fly-In Echelons are expected early enough to take advantage of the V-22s' earlier arrival.

## Summary

Executing the Analysis Phase forces us to answer some fundamental questions about how best to proceed: first about how much research we need to satisfy our analytic objectives and then what general approach we will take. For each analysis, we identify a likely range of viable alternatives that will reasonably satisfy our requirements. Sometimes we know them in advance and at others we decide upon them later in the Analysis phase, after we build our model.

We select criteria meaningful to the decision maker from the family of attributes that describe our options, beginning with Measures of Effectiveness. We evaluate each MOE for validity, reliability, and practicality individually and then the collection as a whole before moving on to address cost.

