

## Design of Valid Operational Tests

By nature, the conduct of a JT&E can be considered an operational test. This is especially true in those tests that include changes to tactics, techniques, or procedures. Testing involves the evaluation of the effect of one or more stimuli, or treatments, on a system or process. For the results of a well executed JT&E to be defensible, the test design process must be capable of providing valid data that are used to calculate measures which are used to evaluate the test issues and sub-issues. The article, an extract from the International Test and Evaluation Journal, June/July 1997 issue, provides the test analyst a consolidated list of things that can have detrimental effects on the planning and execution of a test event/scenario. The "threats" to statistical validity should be considered by the test analyst and test director throughout the JT&E. All too often, the best test plan can fail due to the implementation of poor data collection procedures or a lack of checks and balances that ensures that all relevant data has been collected with appropriate procedures and policies in place to resolved problems with missing or incomplete data. There will be situations that arise which will prohibit the test analyst from obtaining perfect test data. By considering the situations addressed in the article, the likelihood of serious difficulties are reduced.

Of particular note, many JT&E's do not instill a high level of data collection rigor in their test planning and execution. The collection of valid data is the reason for conducting a JT&E. This requires a diligent, and sometimes, pessimistic approach to the Data Management and Analysis Plan. In all JT&Es, the JT&E Director should explicitly appoint a Data Collection Officer/Manager who is responsible for ensuring that all data elements listed in the APA, PTP, DMAP are collected wherever and whenever an opportunity occurs, that all data collection instrumentation has been calibrated and in place, that missing data or incomplete data situations are immediately resolved, that data collection forms, tapes, and other media are expeditiously transported to the data reduction site for cataloging and safe storage. This is one of the most important positions in a JT&E organization.



# Design of Valid Operational Tests

By Richard H. Kass

Test and Experimentation Command, Fort Hood, Texas

*The test community sometimes has difficulty delineating what is meant by a "valid test." Some respond by saying something about sample size, prototype systems, and representative player units. All test agencies have accumulated lessons learned about good test practices. A framework is needed to organize and relate these good test practices to test design validity. This paper provides a definition of test validity, categorizes and relates lessons learned as nineteen threats to test validity, and discusses the implications of this validity concept for designing valid tests. The validity framework presented in this paper can be used as a checklist when designing tests, as well as in the development of valid test plans, comparison of alternate test designs, and training of data collectors and test player units.*

**"I**s this a valid test design?" "Are these results valid?" Test and evaluation (T&E) agencies regularly address such questions concerning test validity. What is meant by the term "valid," and how can tests be evaluated with respect to validity? This paper presents a conceptual framework for designing valid tests based on the validity parameters discussed by Cook and Campbell (1979).

## Test design language

### Test purpose

Causality is the key to understanding the purpose of a test. Tests are conducted to verify a causal proposition. A causal proposition is represented by the statement "A causes B." All tests involve an explicit or implicit causal proposition, as evident in the formulation of criteria for test issues. A criterion statement is an expectation concerning the causal proposition to be observed in the test. It is written as a requirement: "A must cause specific outcome B," for example, System X must detect 90% of the targets. The ultimate product of the test is to provide information to decision-makers, and this information is a report of the ability of the test to verify a causal proposition in the form "A causes B"—the system under test did cause a specific effect or it did not. Understanding tests as a verification of a causal proposition allows the structuring of test components and the meaning of test validity into an heuristic framework.

### Test components

All test designs have the following five components:

(1) **Treatment.** The A of the test, the possible cause, is called the treatment and is sometimes called the *independent variable*. The test treatment may be a new system, a new process, or a new organization.

(2) **Effect.** The B of the test, the possible effect of A, is the measure of the results of the trial, the measure of effectiveness (MOE) or performance (MOP) of interest, e.g., time to complete task, message received or not, and hit or miss. The effect is sometimes called the *dependent variable*.

(3) **Test unit.** The test unit is the smallest unit to which the treatment can be applied. The size of the test unit may be a single soldier, a crew, or even a division or corps, depending on the size of the unit required to implement the treatment.

(4) **Trial.** A trial is one manipulation in the test to produce an observation of A and B. Each trial is the test unit using treatment A to produce one observation B, e.g., to determine if a single transmission from an experimental radio (treatment) is successfully completed (effect). The conditions associated with the execution of the trial are all of the controlled variables (scenario, time of day, training) and uncontrolled variables (weather, motivation, free play) existing during that trial.

(5) **Analysis.** The analysis phase of the test compares the observed results of the trial to other trials or to criterion threshold.

## Meaning of validity

The term "validity" implies the authenticity of something. Is this a valid signature? Is this a valid will? Something is accepted as valid if its authenticity is based on evidence or sound reasoning. The validity of a signature is certified by a notary public. The validity of a will is certified by marshaling evidence to the assent of the person signing the will. Similarly, questions concerning the validity of a test are questions concerning whether the conclusions of the test are justified based on the test conduct. Since the purpose of a test is to demonstrate cau-

validity, a valid test allows conclusions of causality (*A causes B*) to be based on evidence (observations) and sound reasoning. Sound reasoning is demonstrated by designing the test so that known threats to test validity are eliminated.

### Components of validity

Test validity can be addressed by assessing three logically sequenced components of validity: *statistical*, *design*, and *operational* (Table 1). Each component builds upon and extends the earlier component. The first two, *statistical and design validity*, address the internal validity of a test, i.e., the extent that a test allows conclusions indicating that A caused B. One first needs to establish that B, the effect, changed during the test. The ability to provide evidence and sound reasoning that B changed as A changed during the test is *statistical validity*.

Given that B changed, the next question is whether the result was due to the intended treatment or to some unintended cause. The ability to provide evidence and sound reasoning supporting the isolation of the cause of observed result is *design validity*.

*Operational validity*, the third component, addresses the external validity of tests—the ability to provide evidence that the test is related to the operational environment outside the test. Given that B changed and there is reason to believe that A caused the change, will the results observed in the test occur in actual military operations? The ability to provide evidence and sound reasoning that the cause and effect will occur in actual operations is operational validity.

TABLE 1. Test validity can be addressed by assessing statistical, design, and operational validity.

VALIDITY COMPONENT	WITH EVIDENCE	WITHOUT EVIDENCE
Internal Validity	A causes B	A does not cause B
Statistical Validity (ability to detect change)	B changed as A changed	B did not change as A changed
Design Validity (ability to isolate true cause)	A alone caused change in B	Something other than A caused change in B
External Validity	Test relevant to real operations	Test not relevant to real operations
Operational Validity (ability to apply results to actual operations)	Change in B due to A is expected in actual operations	Change in B is unique to test

### Threats to validity

Threats to each component of validity make it difficult to draw the appropriate conclusions. Table 2 depicts nineteen threats to validity. The remainder of this paper discusses these specific threats to validity and how to control or eliminate them when designing a valid test.

### Statistical validity

Statistical validity involves detecting a pattern of change. If data observations (such as target hits, message completions, or times to detect) fluctuated widely from trial to trial, no consistent pattern would be discernible. Statistical validity is the ability to draw quantifiable conclusions, i.e., the ability to detect covariation between treatment A and effect B. Covariation occurs when effect B differs systematically with different applications of treatment A, e.g., System X night trials and System X day trials, or System X trials and System Y trials.

### Threats to statistical validity – Type I errors

Threats to statistical validity are grouped according to whether they increase the risk of a *Type I* or *Type II* error. *Type I* errors occur when we mistakenly conclude that covariation exists between the treatment and the effect when, in reality, it does not. *Type II* errors occur when we mistakenly conclude that A and B do not covary when, in reality, they do.

Erroneously concluding that A and B covary leads to the incorrect conclusion that a test system is associated with a positive result. The easiest way to incorrectly conclude that a positive result exists is not to conduct statistical analysis of the data. After conducting an event three times (three trials) and observing a positive result two out of three times, we are tempted to conclude the test system is better. However, we know that flipping a coin three times can result in two heads even though heads and tails are equally likely. Computing statistical analysis of test data and getting “statistically significant results” indicates that the observed positive results did not occur by chance alone. All test results should be subjected to statistical analysis before drawing conclusions. When conducting statistical analysis, however, the following two threats need to be considered to ensure that the analysis technique itself does not produce the false positive conclusion that the statistical analysis is designed to guard against.

### THREAT 1: VIOLATING ASSUMPTIONS OF STATISTICAL TESTS

Statistical analysis of data requires that certain assumptions be met to assess hypotheses at a specified risk level. Not all assumptions are equally important. Analysis of variance (ANOVA) is fairly insensitive to departures from assumptions of normality or equal within-cell variances.

TABLE 2. A framework consisting of nineteen threats to test validity can be used to support the test design process.

TEST COMPONENT	STATISTICAL VALIDITY	DESIGN VALIDITY		OPERATIONAL VALIDITY
		Single Group	Multiple Group	
TREATMENT	4-System Variability. Do test systems in like trials have the same hardware and software?	8-Systems Changes Over Time. Are there system hardware or software changes during the test?		15-Nonrepresentative System. Is the test system production representative?
TEST UNIT	5-Player Unit Variability. Do individual soldiers/units in like trials have similar characteristics?	9-Player Unit Changes Over Time. Will the player unit change during the test?	12-Player Unit Differences. Are there differences between groups unrelated to the treatment?	16-Nonrepresentative Unit. Is the player unit similar to the intended operational unit?
EFFECT	6-Data Collection Variability. Is there a large error variability in the data collection process?	10-Data Collection Changes Over Time. Are there changes in the instrumentation or manual data collection during the test?	13-Data Collection Differences. Are there potential data collection differences between treatment groups?	17-Nonrepresentative Measures. Do the performance measures reflect the desired operational outcome and have adequate, corroborating data sources for key measures?
TRIAL	7-Trial Condition Variability. Are there uncontrolled changes in trial characteristics for like trials?	11-Trial Condition Changes Over Time. Are there changes in the trial conditions (e.g., weather, light, start conditions, and threat) during the test?	14-Trial Condition Differences. Are the trial conditions similar for each treatment group?	18-Nonrepresentative Scenario. Are the doctrine, tactics, techniques, and procedures employed by the player unit and threat realistic? 19-Nonrepresentative Site. Are the test site conditions similar to the intended area of operation?
ANALYSES	1-Statistical Assumptions. Are assumptions for statistical techniques justified? 2-Error Rate. Are many statistical tests planned? 3-Low Statistical Power. Is the statistical analysis efficient?	<p><i>The purpose of a test is to verify whether A causes B. Valid tests allow the conclusion "A causes B" to be based on evidence and sound reasoning by eliminating or reducing the nineteen threats to validity.</i></p>		

Analysis of covariance (ANCOVA), on the other hand, is quite sensitive to its requirement for homogeneous within-group regression slopes. Nonparametric techniques require fewer assumptions than parametric statistics concerning the level of measurement and underlying distribution. During the design stage, evaluating whether field data will meet the assumptions of the planned statistical analysis is based on analysts' experience with similar data. After data collection, most assumptions for use of a particular statistical technique can be assessed empirically.

#### THREAT 2: ERROR RATE PROBLEM

The likelihood of committing a Type I error increases as the number of statistical comparisons increases. This is relevant when collecting data on many different MOPs in

one test, e.g., detection times, detection ranges, and detection rates. Binomial probabilities can be used to estimate test-wide error. If data for four different MOPs ( $k=4$ ) are collected, and each is independent and analyzed in a statistical hypothesis at the 80% confidence level ( $\alpha=.20$ ), then there is only a 41% confidence  $[(1-\alpha)^k = (1-.20)^4 = .41]$ , rather than an 80% confidence, that all four hypotheses will be true. In other words, there is a 59% probability that at least one of the four individual comparisons will erroneously be accepted as positive (incorrectly concluding A and B covary). A 59% probability of an erroneous conclusion when making multiple comparisons is much higher than the advertised 20% probability of an error for a single comparison. One way to decrease the multiple comparison error rate is to increase the confidence level for the individual comparison.

sons. A higher individual comparison confidence level, for instance, 95% instead of 80%, would increase the overall confidence level from 41% to 82%  $[(1-.05)^4 = .82]$ .

### Threats to statistical validity – Type II errors

Some threats to validity increase the risk of concluding incorrectly that an effective test system is *not* associated with positive results. The ability of a field test to produce discernible results is referred to as *statistical power*. The following five Type II threats to statistical validity are sources of low statistical power:

#### THREAT 3: LOW POWER STATISTICAL ANALYSIS

There are three ways to use statistical analysis that may produce an incorrect, "due to chance" conclusion:

- (1) **Inadequate sample size.** Techniques are available for estimating sample size requirements. In general, the larger the sample size, the greater the statistical power. While sample size is most often the main consideration for determining statistical validity, it is not the only contributor.
- (2) **Setting Type I risk too low.** There is a direct correlation between Type I and Type II risk. Setting the Type I risk higher (accepting more risk by using a 20% rather than 10% risk) correspondingly reduces the Type II risk. If the analyst focuses solely on preventing a Type I error to avoid incorrectly seeing a positive result that is solely due to chance, the analyst runs the risk of creating too stringent conditions to allow small positive results to show up as statistically significant.
- (3) **Inefficient statistical techniques.** Statistical techniques differ with respect to statistical power. Parametric tests are generally more powerful than nonparametric techniques. Analysts need to select the most efficient analytic tool the data will permit.

#### THREAT 4: SYSTEM VARIABILITY

Treatments should be constant throughout a test. This may not always be the case, especially for prototype systems. Sometimes prototype systems that are using a "test-fix-test" design approach undergo hardware, software, or training modifications during testing. To the extent these modifications affect performance randomly, the variance of effect B will increase, making it difficult to detect a true change and decreasing statistical validity. A planned "test-fix-test" design is best accommodated by discrete phasing of the test and analyzing each phase as a "statistical block."

#### THREAT 5: PLAYER UNIT VARIABILITY

Nonstandardization among different test player units increases error variance. Test unit variability is a concern when multiple similar players (e.g., ten sharpshooters) are examined within a particular trial or when a single test

unit (e.g., one tank platoon) is examined across multiple like trials. Nonstandardization occurs when each shooter has a different level of training, different experience, or even a different version of the new system. Standardization among multiple test players can be improved by selecting similar players and by bringing all players to the same level of training. Standardization of a single test player across trials is improved by training that unit to a consistent level of performance prior to the test. The analyst can assess the extent of standardization after the test is completed by comparing scores across players in a single trial or across like trials for a single test unit. Outlier cases can be identified and analysis performed with and without outliers to determine their effect.

#### THREAT 6: DATA COLLECTION VARIABILITY

Many different data collection techniques are used in tests to measure Effect B. Data collection devices include elaborate instrumentation for real-time measurement and not so elaborate procedures, such as stopwatches, data collectors, questionnaires, and observations from technically proficient observers known as subject matter experts (SMEs). Inaccuracy in these devices buries true change within measurement variance. Test agencies have experience in calibrating and honing the measurement precision of instrumentation. There are also techniques for calibrating and increasing the precision of manual data collection procedures (Kass 1984). When the precision of individual measurement devices cannot be adjusted further, measurement precision can still be increased by averaging the responses of two or more data collection systems, e.g., two side-by-side electronic data collection systems, two expert observers, or additional questionnaire items. Precision can also be increased by averaging across multiple observations from a single data collection system.

#### THREAT 7: TRIAL CONDITION VARIABILITY

The prevalence of uncontrolled variables in the test setting yields nonstandardized trials that increase error variance in comparisons across trials. Any increase in variance will obscure statistical differences. A test unit operating under different levels of temperature, weather, light conditions, and terrain across supposedly like trials will fluctuate in performance. The best approach is to control the test to ensure that similar conditions are experienced during like trials; this reduces the number of uncontrolled variables. Unfortunately, there is a trade-off between test control and operational realism. If the threat force attacks the same way each trial, the test unit will know what to expect and operational validity (discussed later) will suffer. When standardized trials are not achievable and the sources of the variability can be identified, some reduction in the variance can be accomplished by

using statistical techniques such as paired comparisons, blocking designs, and ANCOVA.

### **Design validity**

The assessment of design validity is a logical assessment. This contrasts with statistical validity, which can be evaluated statistically. Assessment of design validity requires a knowledge of what factors other than the treatment might affect test results.

After the analyst is reasonably certain that the test is designed to detect change if it occurs, the next question is whether any observed statistical results (B) are caused by the treatment variable (A) or by some other influence. For example, suppose the test unit with the new system was more experienced than the baseline unit with the current system at the start of the exercise. The analyst could not conclude that any increase in performance for the test unit over the baseline unit is a result of the new system. The increase may have been a result of the test unit starting with more experience. Threats to design validity yield biased results and are often referred to as *problems of confounding*. Confounded results are results that may be attributed to an alternative, plausible explanation. A test high in design validity has eliminated or reduced the potential for alternative explanations to observed changes so that the only remaining explanation is the treatment.

### **Threats to design validity**

Threats to design validity depend on the design of the test. Almost all operational tests can be categorized as either a single- or multiple-group design. In *single-group designs*, a single test unit (individual, section, platoon) is trained with the new system and conducts operations with it during the test under multiple conditions—day and night, attack and defend. In *multiple-group designs*, different test units are assigned to different treatment conditions. Multiple-group designs are employed when a second player unit operates an alternative system in a side-by-side comparison test. If this alternative system is the baseline system, then the second player unit is the control group.

### **Single-group design threats**

The Achilles heel of single-group designs is the problem of order effects. Problems arise when one attempts to compare early trials to later trials. Trial order distorts comparisons between trial conditions. For example, if all day trials were conducted first and all night trials conducted last, any comparisons between day and night trials would be confounded with the order effect. Suppose the unit performed better in the day trials. The analyst could not conclude the test system will perform better during day than night, because if the night trials had been conducted first, the system might have performed better during the early night trials and not as well during the

subsequent day trials. This could occur if the system loses alignment as it is operated. Consequently, any observed increase in performance during the original early day trials could be attributed to the intended treatment (day versus night) or could be attributed to the order effect (early versus late).

The best techniques for negating potential confounding due to order effects is randomizing or counterbalancing trial presentation. This is not always possible because trials are often sequenced to accommodate resource availability rather than test design considerations. For example, battlefield smoke trials are usually conducted close together (early or late in the trial sequence) to coincide with the availability of smoke generators and appropriate wind conditions. The following four threats to design validity arise when a single test unit undergoes test trials in some sequence or order.

### **THREAT 8: TREATMENT CHANGES OVER TIME**

Treatments, whether a new system, organization, or procedure, should be constant throughout the test to determine if a specific treatment performs differently under different trial conditions. If both the treatment and trial conditions change, it will be difficult to isolate the true cause of any performance differences. Often systems, especially prototypes, undergo major modifications during testing. These may be hardware, software, or training modifications. A test-fix-test design encourages these modifications. If treatment changes fluctuate randomly, statistical validity is threatened (Threat 4). If the treatment change is increasing (or decreasing) performance nonrandomly over time, the change is a threat to single-group design validity. Treatment changes over time bias any intended comparisons between early and late trials. The threat to test validity is reduced by monitoring for changes in the treatment, counterbalancing trial sequence whenever possible, and checking for any changes in performance over time.

### **THREAT 9: PLAYER UNIT CHANGES OVER TIME**

Players participating in tests will change during the exercise. If the change is one of maturation, players become more experienced and proficient. This is called a learning effect. If the change is one of degradation, players become fatigued, bored, or less motivated. Player changes over time will produce an increase or decrease in performance in later trials, which is unrelated to the change in designed treatment conditions. This makes deciphering the real causality of change difficult. To reduce this threat, counterbalanced techniques should be used when possible. When not possible, test units should be trained to operate at a steady state. After the test is completed, checks for increasing or decreasing performance trends over the temporal sequence of trials should be made.

validity is best achieved when the first unit equipped is the player unit for the test, and this unit is adequately trained.

If the test unit is under- or overtrained, the true capabilities of soldiers in a typical unit will be misrepresented. Undertraining results from compressed schedules to start the test and inadequate training development for new systems. Overtraining arises when player units undergo unique training not planned for units that will receive the fielded systems. Overtraining, like undertraining, is difficult to avoid. Everyone wants to ensure the test unit is well qualified to operate the new system so that the system is given a fair evaluation. The temptation is to overtrain the test unit to ensure success.

Does the test unit represent the appropriate organizational slice for the employment of the new system? If the new system will be employed as a platoon of six weapons, the test unit should be a platoon rather than a single system or a team of two systems. Employing the appropriate organizational-size test unit is a recognition that there are synergistic effects and command and control implications that affect operational capability.

#### THREAT 17: NONREPRESENTATIVE MEASURES

Threat 17 arises when the effect is a complex variable such as unit effectiveness, mission accomplishment, tempo, or command and control. Complex operational concepts are difficult to measure in a test. The best an analysts can do is develop good, approximate measures using several techniques.

Unit effectiveness may be definable in terms of concrete, measurable variables such as loss-exchange ratio, rate of movement, and time to complete a mission. The problem is that component measures may not covary in a similar fashion. In some instances, a slow rate of movement may be associated with a low loss ratio. In other instances, it could be associated with a high loss ratio. While the individual component variable scores can be reported, these scores by themselves do not address overall unit effectiveness, which is the measure of interest. One approach is to select a single component measure that represents the highest level of interest in the complex variable.

When measuring multiple components, analysts also need to ensure individual components are measured independently. If all components are measured in the same manner, any covariation among the component indices cannot be disassociated from the influence of its method of measurement. This is problematic whether the sole data source is an SME, a questionnaire, or electronic instrumentation. For example, if a single rater provides estimates for a unit's ability to maneuver, collect intelligence, and engage the enemy, and these three estimates are combined into a unit effectiveness score, the covariation of these component measures may be artificially high due to a "halo effect." Any inaccuracies in the

single data source induces the same error in each component score and results in an inflated component covariation. To avoid the halo effect, it is best to collect component data using independent sources.

Measuring a complex effect by means of an SME overall subjective rating alleviates the problem of defining, measuring, and combining data from component measures. The quantitative component scores are still valuable in providing a validity check on the composite rating provided by the SME. Credibility in these overall ratings can be increased by showing positive correlations with component measures and by having several SMEs rate the same critical events to demonstrate inter-rater agreement. Ultimately, however, the operational validity of SME composite scores rests on the experience, veracity, and credibility of the SME.

#### THREAT 18: NONREPRESENTATIVE SCENARIO

How realistic is the trial scenario for the test unit? Three factors should be considered:

(1) **Realistic doctrine, tactics, techniques, and procedures.** Many circumstances may make it difficult to use realistic tactics during a test. Modifying current tactics to incorporate a new system often follows rather than precedes new system development. Even when new tactics have been developed, adequate training is difficult due to prototype shortages. Additionally, terrain, instrumentation, or safety constraints during test execution may preclude appropriate tactical maneuvering. Similarly, representation of threat tactics and equipment in the test is difficult. Captured threat equipment is not always available and training operational units to emulate threat tactics is a low priority, except at centralized training centers. Sufficient time needs to be allocated for training the test unit and threat unit in appropriate tactics. Tactical units can assist the tester in developing realistic operational plans that provide for appropriate force ratios, missions, and maneuver space and time.

(2) **Battlefield intensity.** It is impossible to create conditions during a field test that approximate the noise, confusion, fear, and uncertainty of combat. Lack of player apprehension during test trials is a threat to operational validity and can be offset by increasing the realism of player participation. Use of lasers to simulate engagements increases the realism of direct fire engagements. Other methods include allowing the exercise to continue for many hours or days to generate fatigue-associated stress.

(3) **Player uncertainty.** Over time players can anticipate and prepare for scenario events. Directing a unit to an assembly area during continuous operations to calibrate instrumentation is a signal to the unit that a battle will soon occur. Surprise has evaporated. Additionally, player units that undergo the same scenario over a sequence of trials know what to expect. Anticipation of scenario events pro-

motes lack of apprehension and promotes nonrepresentativeness of unit reactions. Allowing for maximum free play and adding new scenario events promote player uncertainty.

#### **THREAT 19: NONREPRESENTATIVE SITE**

Test site constraints on the operational employment of the test system may restrict the operational representativeness of the effects. Hills and vegetation limit direct-fire engagements to close ranges, while flat, open areas permit long-range engagements. Terrain, instrumentation, or other site restrictions (such as environmental, frequency spectrum, and air corridors) may lower the operational validity of the test. Analysts must be aware of which aspects of the intended operational environment are available at the test site. The test report may conclude A affects B, when the test only demonstrated that some levels of A affect B.

#### **Judgment critical to operational validity**

Tests are never perfect representations of actual combat operations. Operational validity, however, depends on approximating the operational conditions to which the conclusions of the test are pertinent. To formally assess operational validity, the analyst would need to examine data from a series of similar tests involving different units and different environments. Field tests for the sake of replicating findings are seldom funded. Consequently, the assessment of operational validity rests on judgments as to the representativeness of the system, the measures, the player unit, the scenario, and the site conditions under which the test was conducted.

#### **Validity framework**

Tests are a balance, primarily between internal and external validity. Precision and control increase internal validity (statistical and design validity) but decrease external validity (operational validity). Tests with high internal validity emphasize strict control of trial conditions and multiple repetition of similar events. On the other hand, tests high in operational validity emphasize free-play exercises and uncertainty in unique scenarios. Consequently, 100% valid tests are not achievable. Different validity components can be emphasized in a test. Trade-off is inevitable; analysts need to be cognizant of validity trade-offs and explicit about priorities when designing for validity in order to minimize the loss of one type of validity because of the priority of another.

Test design priorities can differ. In tests where one expects a small effect and it is important to determine the precise relationship between the treatment and its effect, the priority should be internal validity. On the other hand, if one expects a large effect, and if it is important to determine if the effect will occur in the operational environment with typical units, and if there is less need to address questions of why the specific result occurred,

then external validity is the priority.

In most tests of new materiel, a case can be made for prioritizing design validity above operational validity. A very realistic test may be conducted; but in the end, if analysts cannot, with some degree of assurance, make a case for or against the new system, the test will have been an expensive training exercise. To ensure an adequate level of design validity, some operational realism may need to be sacrificed. A scenario calling for continuous tactical operations may have to be interrupted periodically to realign data collection instrumentation. Emphasizing design validity does not imply operational validity is not critical. It is critical, and every effort should be made to minimize the impact of the test on operational realism.

#### **Framework applications**

This description of nineteen threats to validity should contain no surprises to experienced test officers. These threats summarize accumulated test agency lessons learned in a coherent, heuristic framework.

Test plan writers and reviewers can use the 19-point checklist to determine if all components of validity have been addressed in the test design plan. Known test limitations in the plan can be related to specific threats to show the impact of limitations on the different components of validity. Test officers can evaluate alternate test designs with respect to the ability of each design to eliminate or control the 19 threats. Data collector training can enhance test validity by focusing on eliminating data collection threats to statistical, design, and operational validity. Similarly, test unit training can be monitored to adjust and reduce player unit threats.

Test validity is multifaceted. Most analysts, when asked about validity, dwell on the concept of sample size and conclude that a field test "is" or "is not" valid based on this single dimension. Sample size is only one of nineteen threats to validity, and it affects only one of the three validity components. A test is never totally valid or totally invalid. Maximizing some aspects of validity necessarily minimizes others. A good test design maximizes those aspects of greatest importance to the purpose of the test. □

#### **References**

- Cook, T.D. and Campbell, D.T. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.
- Kass, R. October 1984. "Calibrating Questionnaires and Evaluators," *The ITEA Journal of Test and Evaluation*, Vol V, No. 3, 26-32.
- RICHARD A. KASS, a senior test manager at the Army Test and Experimentation Command (TEXCOM) is currently refining data collection strategies for Army advanced warfighting experiments. He holds a Ph.D. from Southern Illinois University.

