A Primer on Developing Measures of Effectiveness

Major Shon McCormick, U.S. Army

Major Shon McCormick, U.S. Army, is currently deployed to Afghanistan. Prior to his deployment, he served as an instructor for the Department of Joint, Interagency, and Multinational Operations at the U.S. Army Command and General Staff College, Fort Leavenworth, KS. He holds a B.S. from the United States Military Academy, and an M.A. from Kansas State University.

PHOTO: A Shura with tribal elders on 5 September 2005 in the village of Mainasheen, Afghanistan. (U.S. Army photo by SGT Andre' Reynolds)

TO UNDERSTAND THE operational level of war, students must appreciate the newest Joint doctrine. At the Command and General Staff College, Fort Leavenworth, Kansas, an important doctrinal concept is *elements of operational design*. As an instructor at the college, I have observed that "measures of effectiveness" are a difficult aspect of operational design for students to understand. Because my own knowledge of the concept was lacking, I conducted some research on the topic by scanning existing Joint doctrine and asking around the school. My only success came from individuals at the U.S. Army Training and Doctrine Command (TRADOC) Analysis Center, who provided a NATO manual on best practices for assessing command and control systems.[1] This manual is informative about measures of effectiveness, albeit not in the context of operational design, and it is too technical and specialized for most staff officers.

The dearth of knowledge surrounding measures of effectiveness extends beyond the college student population. This conclusion is not meant to disparage anyone or any institution but to highlight the general lack of understanding surrounding the concept of measures of effectiveness. This article reflects my efforts to describe a practical but rigorous method to develop measures of effectiveness that a nonspecialist can employ. The article will cover—

● Examining the utility of metrics in general.

● Analyzing how current Joint doctrine portrays measures of effectiveness as part of the elements of operational design.

● Exploring how fields outside the military deal with concepts like measures of effectiveness.

● Mining those other fields for insights to help bridge some of the gaps in current military doctrine.

● Providing observations on the implications of my findings on the emerging Army doctrine of design and related concepts.

Metrics

Why should we care about measures of effectiveness? The answer is that current Joint doctrine says so. However, this is a circular argument, and the question warrants a better answer. Pragmatic military leaders should care about measures of effectiveness if for no other reason than that the American people's representatives in Congress care about them. The requirements to brief Congress on progress in Iraq and Afghanistan are examples. An article by Patrick Cronin notes that congressional members from both parties have indicated that continued support for efforts in Iraq is contingent on "credible evidence of tangible military progress."[2]

In addition, a series of recent U.S. Joint Forces Command studies reinforce the utility of assessment tools such as measures of effectiveness. *Joint Operations: Insights and Best Practices* cites the use of assessment measures as "an important best practice whose need is reinforced time and time again in operational headquarters."[3] The study makes measures of effectiveness especially important in today's complex operating environment, which challenges planners' abilities to predict the outcome of their plans accurately.[4]

Current Doctrine

One logical place to begin is by surveying existing doctrine for some guidance on how to develop measures of effectiveness. The authoritative doctrinal references for measures of effectiveness are Joint Publication (JP) 5-0, *Joint Operation Planning,* and its companion manual, JP 3-0, *Joint Operations*.[5] These two manuals combine both measures of effectiveness and measures of performance under the general title of *Assessment Measures* and direct staffs to develop assessment measures during mission analysis. Other than that, Joint doctrine provides no insight on the actual mechanics of developing suitable measures of effectiveness.

In the absence of doctrinal guidance, research beyond military publications becomes necessary. In that regard, I will explore three fields:

● The basic tenets of social science research methodology. This field has long dealt with the very issues that the design element of measures of effectiveness tries to address.

● Policy and program evaluation, which covers the same ideas as measures of effectiveness in the arena of domestic public policy.

● Emerging Army doctrine on campaign design and how these emerging concepts deal with the challenge of assessment.

Social Science

First, an explanation of social science as opposed to physical science is in order. In simple terms, social science involves the behavior of people. One attribute of the study of social science is the inability to conduct research in controlled experimental environments; we cannot conduct social research in an environment where we can control all influences. Both practical and ethical considerations prevent us from experimenting on human groups the same way as on lab rats. As a result, when we do social science, we accept that a certain amount of error, both random and systematic, is inevitable.[6]

Although current doctrine often conflates causation and correlation, social science treats the concepts very differently. Correlation means that two events tend to occur together with some frequency. The classic example is that of a rooster crowing at dawn. One can observe that almost every time the sun comes up, a rooster crows. The two events display a high degree of correlation. However, correlation does not equal causation. Falsely attributing causation is the *post hoc fallacy*. Based on our simply observing the sun and the rooster, we cannot determine whether the rooster's crowing causes the sun to rise, the sun's rising causes the rooster to crow, or even if the two events have any causal relationship. How

> *Pragmatic military leaders should care about measures of effectiveness if for no other reason than that the American people's representatives in Congress care about them.*

to determine the degree of causation between two correlated activities is the essence of science. Since both physical and social science have been wrestling with causation for a long time, an accepted body of knowledge has emerged. While the body of knowledge is vast, a few key points are relevant here:

- Correlation does not equal causation.
- One can only determine causation by employing a hypothesis.
- One can never absolutely determine causation; one can only reduce uncertainty.

Hypothesis

A hypothesis is simply a proposed causal relationship between two activities that lends itself to testing. For example, the concept of the "surge" in Iraq was essentially a two-step hypothesis that tested whether increasing the number of coalition troops in Baghdad would reduce insurgent violence, and whether this reduction in violence would lead to reconciliation between Iraq's Sunni and Shiite political factions.[7] We can never be certain that an increase in troop strength truly led to a reduction in violence. Instead, we can only reduce our uncertainty by applying a number of techniques to determine if other causal factors are at play. The actual procedures to do this are beyond the scope of this article.

A hypothesis is necessary to test for causation, so the next challenge is to develop the hypothesis. Our doctrine is vague in this regard, but science offers three alternatives: employ a hypothesis developed by someone else for similar circumstances, develop your own, or employ a combination of the two previous approaches. The simplest way to find existing hypotheses is to consult the existing body of knowledge on the topic of interest. To determine the link between security force presence and insurgent violence, a good place to start would be studying research done by others on that same topic. However, no two situations are exactly

A U.S. Marine Corps captain talks with village elders during a shura in the Nawa district of Helmand province, Afghanistan, 10 August 2009.

alike. Even the most similar circumstances can have important omitted factors. Should we arm the Afghan tribesman with the same weapons we gave the Awakenings Council in Iraq? Will what worked in Iraq work in Afghanistan, given the two nations' very different histories and levels of development?

If there is no suitable existing hypothesis, then one has to create a hypothesis from original research. In simple terms, creating a hypothesis requires one to speculate on a causal relationship between two activities or variables. The source of this casual relationship can simply be a hunch or some other form of insight. After an initial hypothesis, the researcher must then test it against suitable cases from history. This is difficult because no two cases from history are completely alike or have the same casual factors. The goal of the researcher is

We can never be certain that an increase in troop strength truly led to a reduction in violence.

to determine which factors across several cases are general and which are specific to one case. Regardless of the source of the hypothesis, the next step is to employ the hypothesis to predict future events. In simple terms, if an idea seemed to work in the past, it may work in relatively similar circumstances in the future. In light of the already established caveat that the past is not a perfect predictor for the future, our hypothesis at best provides an educated guess on some unknown outcome. Since we must accept that our hypothesis has some degree of error, our task is to determine when our hypothesis has failed, or is false. Unfortunately, we may already have our plan in execution before we can reach any conclusion on our hypothesis. This is where measures of effectiveness become important.

To better explore the role of measures of effectiveness in testing hypotheses, we turn away from social science and enter the field of program evaluation. A quick internet search of the term "program evaluation" reveals a broad discipline with a large body of research. Nonmilitary agencies have been dealing with ways to assess the effectiveness of various programs in a formal way since the mid-1940s. Recent programs like "No Child Left Behind" or even President Obama's stimulus package are simply efforts to influence some system in a desirable way.

Programs

Before exploring the field of program evaluation, a few definitions are in order:

● A *program* is a "set of resources directed toward one or more common goals,"[8] or a hypothesis that "if followed, then the expected results will follow."[9]

● An *input* is simply that which goes into the program.

● An *output* is the "products, goods, and services" that come out of the program and are then provided to the intended recipients.

● Finally, the *outcome* is a "change or benefit resulting from the outputs."[10]

The definition of the elements of program evaluation is similar to the military doctrinal terms of measures of performance and measures of effectiveness. A measure of output is analogous to a measure of performance and a measure of outcome is analogous to a measure of effectiveness. In light of this similarity, measures of outcomes from program theory should prove useful in helping explain measures of effectiveness from military doctrine.

"Logic models," or "modeling" are concepts that are central to the field of program evaluation. They clarify the relationship between a program's inputs, outputs, and outcomes. Implicit in a logic model is its program theory, the causal hypothesis that links the model's elements. Program theories predict outcomes in the development of the program and determine causal relationships between inputs and outcomes after program implementation.[11]

Comparing outcomes requires some measurement of those outcomes. *Operationalization* is the process of creating metrics for inputs, outputs, and outcomes. Some outcomes lend themselves to measurement more easily than others. Examples of easily quantifiable outcomes are financial costs or casualties.

Measuring Problematic Variables

However, not all outcomes are so easily measurable. Examples of more problematic variables related to the military are outcomes such as *security* or *democracy*. In the case of these more abstract concepts, the researcher must employ indicators or proxies. While seemingly straightforward, the selection of indicators is complex. For example, how does one measure democracy? The difficulty in developing valid measures for more abstract outcomes often requires reviewing the existing research literature and consulting experts and practitioners within the field of interest.[12]

A return to social science methodology is useful at this point. Gary King, Robert O. Keohane, and Sidney Verba advise researchers to determine as many "observable implications" of their hypothesis as possible in order to create more cases for testing the hypothesis. Their thought is that more testing of more implications will more likely reveal any problems with the proposed indicators.[13]

Inputs, outputs, and outcomes. Once the program is in execution, a comparison of inputs, outputs, and outcomes informs the program manager of the validity of the underlying program logic. If this program logic is flawed, then the manager must reexamine and perhaps refine the model. At this point, the military staff seeking to employ program logic theory would need someone trained

in statistical analysis to determine which of the elements is flawed. Two common methods are *randomized experiments* and *quasi-experimentation*.[14] The basic difference between the two methods involves the degree of control the analyst maintains over the environment. The actual mechanics of conducting randomized experiments or quasi-experimentation are beyond the scope of this article.

A military example of measures of effectiveness development is in order. Given ongoing operations in Iraq and Afghanistan and the military's emerging focus on stability operations, this example will focus on the challenge of establishing security in a post-conflict environment.

The setting for this example is a planning staff responsible for conducting a stability operation in a post-conflict environment. This operational environment has an unacceptably high degree of violence, which threatens the ability of the fragile host nation government to establish authority. The problem facing the planning staff is to strengthen the capacity of the host nation government to effectively control its own territory. (Notice that the problem is

not to simply reduce violence. To define the problem as reducing violence is to assume there is a causal relationship between a reduction in violence and increased host nation governance capacity.) For this example, assume that the staff has reducing violence as the objective, i.e., the *outcome* sought. The next challenge for the staff is to determine what resources are available and how to employ them to achieve the desired outcome. This step requires adopting a program theory that proposes a causal relationship between inputs, outputs, and outcomes. The staff elected not to conduct its own independent research due to time constraints and therefore had to rely on existing research. One obvious source is doctrine. However, as Christopher Paparone has noted, one of the problems with doctrine is that it never cites its sources. For example, FM 3-07, *Stability Operations,* offers good general guidance on how to conduct the stability task of "Establish Civil Security," but readers are unaware of which specific historical cases actually influenced this generalization.[15] Since doctrine is insufficient, the staff must broaden its research.

The commanding general of the Afghan National Army 215th Corps and his English-speaking linguist address local leaders and U.S. Marines during a regional security meeting at Camp Dwyer, Afghanistan, 13 May 2010.

U.S. Marine Corps photo by SSGT Ezekiel R. Kitandwe

One potential source of causal hypotheses is *Winning the Peace: An American Strategy for Post-Conflict Reconstruction.* This book serves as the basis for the State Department's "Post Conflict Reconstruction Essential Tasks Matrix," which influenced FM 3-07. According to Scott Feil's chapter on enhancing security capabilities, establishing security is a prerequisite for any development or reconstruction activity.[16] Successful security efforts consist of a combination of defensive and offensive protection activities that "remove the capacity for groups and individuals to engage in illegitimate violence." With respect to defensive measures, the general populace is one element requiring protection.[17] With Feil's research in mind, we hypothesize that protecting the general populace leads to a reduction in illegitimate violence. However, this hypothesis does not tell us how to secure the general populace, so we have to continue our research. In *The Quest for Viable Peace,* Ben Lovelock reports that increasing foot patrols in populated areas was a successful technique to secure the general populace in Kosovo in the 1990s.[18]

Combining Feil's and Lovelick's hypotheses produces the following logic narrative: If an organization increases foot patrols (inputs), then the general populace will be more secure (output). If the general populace is more secure, then illegitimate violence should decrease (outcome).

Elements of the program model. Having determined program logic theory and created a logic model, the next step is to determine measures of effectiveness for the various elements of the program model. Measurement of foot patrols is relatively easy. In this case, a measure of performance derived from Army doctrine would be appropriate. Metrics could include number and duration of patrols as well as the area covered. Measurement of the outputs of population security is more challenging because "security" is more abstract. As a result, we rely on proxies or indicators. Neither Feil nor Lovelick, our sources for our program theory, provide indicators, so further research is necessary. One work that does address indicators for security is the recent RAND *Guidebook for Supporting Economic Development in Stability Operations,* which lists a number of indicators for population security, such as the number of people fleeing their homes.[19]

Measurement of the outcome of reduced illegitimate violence could involve measuring reported crimes and violent death. In this case, the analyst would be relying more on intuition than existing theory to choose the metric. However, the RAND guidebook offers some guidance on employing crime data as a metric. The guidebook offers the caveat that the most likely source of this information is police data, which only reflects reported violence. In addition, the guidebook cautions that successful reconstruction projects often serve as lucrative targets to insurgents and may actually lead to an increase in violence. The analyst wishing to measure violence accurately would have to accept diminished accuracy due to unreported acts of violence and increased violence near reconstruction projects.

Once the analyst has created an accepted program logic model and employed it, the staff will need to determine the security program's effectiveness. The first step is to actually conduct measurements. The United States Institute for Peace offers four primary methods for collecting measurement data. These methods are—

- Content analysis of local media products.
- Consultation of a panel of experts.
- Statistical analysis.
- The use of polls and surveys.[20]

The next step is to establish the relationship between the measurements. According to the program theory, an increase in foot patrols should ultimately lead to a decrease in reported violence, which indicates an increase in overall stability. If this chain of hypotheses does not hold true, then the analyst must reexamine the logic model. We relied upon models of past events, and there is no guarantee that our program logic is entirely valid in the current environment. Perhaps increased foot patrols served only to further alienate the populace and increase the perception of the coalition forces as occupiers. If so, the situation calls for a more appropriate program theory. Perhaps the program theory and even the logic model itself are valid, but the indicators of security are not. In this case, the analysts must develop better indicators.

For a good real-world example of the application of the principles addressed in this section, see the Organization for Economic Cooperation and Development's working draft of *Guidance on Evaluation Conflict and Peacebuilding Activities*. This document

is a highly readable guide for both government and nongovernment practitioners who employ program logic theory during stability and reconstruction activities. Annex 6, "Understanding and Evaluating Theories of Change," has a tabulated summary of major theories suitable for use as program logic.[21]

The Future

The complexity of today's operational environment has led to a number of initiatives to improve military planning through the concept of *design*. The Army has incorporated the tenets of "systemic operational design" through the publication of FM 5-0, *The Operations Process*.[22]

The language of design expressed in FM 5-0 seems remarkably similar to that of program logic theory, which has been extant since the 1940s. Both constructs accept that initial solutions may not be valid. Both focus on explicit hypotheses linking inputs, outputs, and outcomes. Methods to create measures of effectiveness under the emerging framework of design are similar to those found in this article and in the Functional Area 49 Operations Research Systems Analyst community. The Army's systems analysts have long employed metrics in complex environments and can provide useful input into emerging planning processes. As the Army continues to discard mechanistic and deterministic planning methods associated with the defunct "effects-based approach" and incorporates tenets of design into doctrine, it should not neglect these existing bodies of knowledge.

Emerging doctrine suggests that measures of effectiveness and associated concepts of operational design are not going away. A basic understanding of measures of effectiveness and how to create them will remain a fundamental skill for commanders and their staffs as long as the Army employs the elements of design. The concept of measures of effectiveness should not intimidate us. All but a few facets of constructing measures of effectiveness are within the capabilities of a typical field grade officer. *MR*

NOTES

1. *NATO Code of Best Practice for Command and Control Assessment* (Washington, DC: DOD Command and Control Research Program, 2002).

2. Patrick M. Cronin, "Irregular Warfare: New Challenges for Civil-Military Relations," *Strategic Forum* 234 (October 2008): 4, available at <www.ndu.edu/inss>.

3. Joint Warfighting Center, United States Joint Forces Command, *Joint Operations: Insights and Best Practice* 2d ed. (July 2008), Gary Luck, Mike Findlay, and the JWFC Joint Training Division, 33.

4. Ibid., 35.

5. Joint Publication (JP) 3-0, *Joint Operations,* incorporating chg. 1 (Washington, DC: U.S. Government Printing Office [GPO], February 2008) and JP 5-0 *Joint Operation Planning* (Washington, DC: GPO, December 2006).

6. Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton: Princeton University Press, 1994), 8-9.

7. Frederick W. Kagan, *Choosing Victory: A Plan for Success in Iraq,* Phase I Report, American Enterprise Institute Iraq Study Group (January 2007), 1.

8. Kathryn E. Newcomer, Harry P. Hatry, and Joseph S. Wholey, "Meeting the Needs for Practical Evaluation: An Introduction," *Handbook of Practical Program Evaluation,* 2d ed. (San Francisco: Jossey-Bass, 2004), xxxiv.

9. John A. McLaughlin and Gretchen B. Jordan, "Using Logic Models," *Handbook of Practical Program Evaluation,* 2d ed. (San Francisco: Jossey-Bass, 2004), 7.

10. Ibid., 9.

11. Ibid., 10-11.

12. Arnold Love, "Implementation Evaluation," *Handbook of Practical Program Evaluation* 2d ed. (San Francisco: Jossey-Bass, 2004), 76-78.

13. King, Keohane, and Verba, 28.

14. Charles S. Reichhart and Melvin M. Mack, "Quasi-Experimentation," and Robert G. St. Pierre, "Using Randomized Experiments," from *Handbook of Practical Program Evaluation,* 2d ed. (San Francisco: Jossey-Bass, 2004).

15. This comment applies analysis provided by Stephen Biddle in his review of FM 3-24 in "The New U.S. Army/Marine Corps Counterinsurgency Field Manual as Political Science and Political Praxis," *Perspectives on Politics* 6, vol. 2 (June 2008): 347-50.

16. Scott Feil, "Laying the Foundation: Enhancing Security Capabilities," *Winning the Peace: An American Strategy for Post-Conflict Resolution,* ed. Robert C. Orr (Washington, DC: The CSIS Press, 2004), 40.

17. Ibid., 42.

18. Ben Lovelock, "Securing a Viable Peace: Defeating Militant Extremists—Fourth-Generation Peace Implementation," *The Quest for Viable Peace: International Intervention and Strategies for Conflict Transformation,* ed. Jock Covey, Michael Dziedzic, and Leonard R. Hawley (Washington DC: United States Institute of Peace, 2005), 134.

19. Keith Crane, et al. *Guidebook for Supporting Economic Development in Stability Operations* (RAND: Santa Monica, 2009).

20. United States Institute of Peace, *Measuring Progress In Conflict Environments (MPICE): Metrics Framework For Assessing Conflict Transformation and Stabilization* (February 2008).

21. Organization for Economic Cooperation and Development, *Guidance on Evaluation Conflict and Peacebuilding Activities.* Working draft for application period (2008), available at <www.oecd.org/document/35/0,3343 ,en_21571361_34047972_31779555_1_1_1,00.html> (24 February 2009).

22. FM 5-0, *The Operations Process* (Washington DC: GPO, March 2010)