Measuring the Effectiveness of Special Operations

Appendixes

LINDA ROBINSON, DANIEL EGEL, RYAN ANDREW BROWN

Prepared for the United States Army
Approved for public release; distribution unlimited

For more information on this publication, visit www.rand.org/t/RR2504

Published by the RAND Corporation, Santa Monica, Calif.

Preface

This report documents research and analysis conducted as part of a project entitled Special Operations Forces (SOF) Measures of Effectiveness, sponsored by U.S. Army Special Operations Command. The purpose of the project was to develop a system for evidence-based assessments to evaluate the success (or failure) of various special operations missions and demonstrate how Army special operations forces contribute to theater and national security objectives. This volume provides the appendixes to Linda Robinson, Daniel Egel, and Ryan Andrew Brown, *Measuring the Effectiveness of Special Operations*, Santa Monica, Calif.: RAND Corporation, RR-2504-A, 2019.

This research was conducted within RAND Arroyo Center's Strategy, Doctrine, and Resources Program. RAND Arroyo Center, part of the RAND Corporation, is a federally funded research and development center (FFRDC) sponsored by the United States Army.

RAND operates under a "Federal-Wide Assurance" (FWA00003425) and complies with the Code of Federal Regulations for the Protection of Human Subjects Under United States Law (45 CFR 46), also known as "the Common Rule," as well as with the implementation guidance set forth in DoD Instruction 3216.02. As applicable, this compliance includes reviews and approvals by RAND's Institutional Review Board (the Human Subjects Protection Committee) and by the U.S. Army. The views of sources utilized in this study are solely their own and do not represent the official policy or position of DoD or the U.S. Government.

Contents

Figure and Tables

Figure

Tables

APPENDIX A

Assessment Toolbox

This appendix provides a detailed review of the types of data and associated analytical techniques that are available for the assessment team. This is designed to augment the step-by-step assessment process described in Chapter Four of the main report.[1]

The first section describes the different types of data that are (potentially) available for analysis. Although the general guidance is to use as many different classes of data as possible because each provides a different perspective, both the permissiveness of the environment and the resources available to the command will play a strong role in determining which types of data are accessible. As an important example, operational data are the cheapest and easiest to collect (because they are already being collected) but may not capture the full range of population perspectives (although population-focused methods may also fail in that regard, as discussed later). A general good practice is to try to identify data sources that can be used for several measures of effectiveness (MOEs) to limit the cost and effects of data collection on tactical units.

The second describes methods for analyzing these data. Both qualitative and quantitative methods offer results with causal interpretations, if the analysis of each is appropriately designed. The appropriate method will be determined by the type of available data and the background of the assessment team. *Triangulation* (also known as *mixed-methods* analysis) combines both qualitative and quantitative methods, which allows a team to use the findings from a quantitative approach to validate those of a qualitative approach (and vice versa). Given our goal of producing rigorous and defensible results, the discussion of analytical methods provides a variety of practical examples, based on existing research, of how these approaches might be deployed. However, when possible, we recommend that assessment teams use some version of triangulation, incorporating both quantitative and qualitative data sets for inference.

Data

Researchers often distinguish between qualitative and quantitative data. It is important to note that these two categories refer to a continuum of data types, from those that are less structured to those that are more structured. Much of qualitative analysis involves leveraging human judgment to detect patterns in relatively unstructured data, with the goal of creating more structure in the data with which to make reliable conclusions.

[1] Linda Robinson, Daniel Egel, and Ryan Andrew Brown, *Measuring the Effectiveness of Special Operations*, Santa Monica, Calif.: RAND Corporation, RR-2504-A, 2019.

With the rise of big data (see Chapter Six),[2] semistructured or unstructured data have become an increasingly large portion of all available data.[3] Luckily, technological approaches for analysis are keeping pace. Large numbers of documents, such as thousands of tweets, can be analyzed by using automated textual analysis to identify some initial lexical characteristics. Computational algorithms can group text into bins for more-detailed manual analysis.[4]

The following subsections discuss three different types of data—operational, ambient, and assessment specific—that can be analyzed for special operations forces (SOF) assessments. These are listed in order of increasing effort to properly collect, clean, and analyze. Table A.1 defines and provides examples of each type of data, as well as strengths and weaknesses.

Operational Data

Military units and supporting interagency elements collect a multitude of data during operations, typically including intelligence reporting, situation reports (SITREPs), planning documents describing specific activities being executed (e.g., concepts of operation [CONOPs]), and some systemized reporting on engagement with enemy and friendly forces (e.g., significant activities [SIGACTs], Blue force tracker, mission tracker, key leader engagements). Although rarely designed to specifically support assessments, operational data provide geographically and temporally precise information on both inputs and outputs of operations. These data can be useful for understanding how operations are expected to lead to effects, and are thus useful for constructing measures of performance (MOPs) and the causal model linking MOPs to MOEs, and can be useful sources of data on the effectiveness of operations in their own right. Some operational data are already heavily structured (e.g., significant activities) and therefore require minimal coding or other processing for quantitative analysis. Other operational data (e.g., SITREPs) are more loosely structured and require either significant hand-coding or

Table A.1
Data Sources for Assessment Toolbox

Type	Description	Example	Strengths and Weaknesses
Operational	Data collected as part of ongoing military operations and interagency efforts (e.g., U.S. Department of State)	SITREPs Intelligence SIGACTs CONOPs	Operationally relevant and easy to access Not much information on effects
Ambient	Nonmilitary data streams produced or collected for other purposes	Social media Commercial imagery	Covers broad areas and populations Can require significant processing
Assessment specific	Data collection designed specifically for assessment purposes	Polling Interviews with locals	Tailored to assessment needs Expensive and burdensome to collect

[2] See "Emerging Data Streams" in Robinson, Egel, and Brown, 2019, Chapter Six.

[3] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, and William Money, "Big Data: Issues and Challenges Moving Forward," paper presented at the 46th Hawaii International Conference on System Sciences, January 7–10, 2013.

[4] See "Emerging Data Streams" and "Emerging Technologies" in Robinson, Egel, and Brown, 2019, Chapter Six. See also Elizabeth Bodine-Baron, Todd C. Helmus, Madeline Magnuson, and Zev Winkelman, *Examining ISIS Support and Opposition Networks on Twitter*, Santa Monica, Calif.: RAND Corporation, RR-1328-RC, 2016.

machine processing before most types of analysis. The following are a few commonly collected (and used) forms of operational data, along with some of their strengths and weaknesses:

- **SIGACTs** track violent incidents involving coalition forces. While SIGACTs are often easily accessible and easily structured for analysis, one main weakness is that they are confounded with the presence of U.S. forces. That is, simply the presence of more U.S. forces (or more operations) increases the chances of enemy-initiated SIGACTs.[5] Thus, reductions or increases in SIGACTs in an area of operation (AO) should be interpreted with caution.[6] For analysis, SIGACTs are usually filtered (e.g., filtered down to enemy-initiated SIGACTs), and trajectories of SIGACTs over time are analyzed in specific geographic areas connected with ongoing operations.
- **SITREPs**, which all tactical commanders produce, typically contain information about both operations and their perceived effectiveness. These documents are produced frequently and contain a combination of structured (e.g., number of operations) and unstructured (e.g., atmospherics) reporting, typically consisting of loosely structured textual data. SITREPS typically provide a combination of information required by higher headquarters and information the unit considers important (and proper) to report up the chain of command. At a minimum, a SITREP will include information on the unit's position, enemy and friendly actions observed during the reporting period, and both recent and planned operations.[7] SITREPs can be useful sources for details about ongoing operations and subjective observations of operational success (or failure) but tend to require painstaking human coding to produce quantitative data (or enough qualitative observations to make reliable inferences).
- **Intelligence reporting** can come from a wide variety of sources and usually consists of unstructured or semistructured text. Like other operational data, coverage is limited because such reporting is available only when someone is actively listening or gathering it. This information is, however, often more inclusive than other data streams. Intelligence reporting can provide useful insights into more subtle aspects of enemy behavior—including (sometimes) detailed observations on morale and other more-subjective properties. Intelligence reports can be binned by category and counted, or more-detailed analysis of their content can produce thematic insights.
- **CONOPs** provide idealized conceptual depictions of various operational activities and how they are expected to be sequenced and feed into each other. They can be very useful during assessment planning because they spell out expected operational activities and effects. Inferences drawn from CONOPs should be checked against SITREPs, mission tracker data, or other forms of operational reporting to properly validate what actually occurred versus the planned operations. CONOPs can be useful for thinking through MOPs but are not sources of information for effects or the outcomes of operations.

---

[5] Daniel Egel, Charles P. Ries, Ben Connable, Todd C. Helmus, Eric Robinson, Isaac Baruffi, Melissa A. Bradley, Kurt Card, Kathleen Loa, Sean Mann, Fernando Sedano, Stephan B. Seabrook, and Robert Stewart, *Investing in the Fight: Assessing the Use of the Commander's Emergency Response Program in Afghanistan*, Santa Monica, Calif.: RAND Corporation, RR-1508-OSD, 2016.

[6] Ben Connable, *Embracing the Fog of War: Assessment and Metrics in Counterinsurgency*, Santa Monica, Calif.: RAND Corporation, MG-1086-DOD, 2012.

[7] Field Manual 6-99, *U.S. Army Report and Message Formats*, Washington, D.C.: Headquarters, Department of the Army, August 19, 2013.

Ambient Data

A wide variety of data—social media, satellite imagery, news, etc.—is typically produced and collected by other entities in environments where military operations are being conducted. These nonmilitary "sensors" compile data on the movements and actions of individuals and groups, both passively and actively. Passive data collection includes social media; individuals leave a digital record when posting on social media. Other data are collected actively; for example, the National Oceanic and Atmospheric Administration and the U.S. Geological Survey collect satellite imagery to understand economic or agriculture activity, and both commercial and nongovernmental organizations (NGOs) collect imagery and other environmental data. Such data are frequently amenable to the secondary purpose of assessing military operations. Finally, newspapers, magazines, newsletters, academic articles, and other print and online media have provided a rich source of intelligence and background contextual data for decades and continue to be useful for some assessment work.

Ambient data can require more time and effort to acquire than operational data. Depending on their format, ambient data can also require significant effort to filter, clean, and process to provide insights on the effectiveness of operations. Specifically, some ambient data are heavily prestructured (or can be easily structured) for analysis; for example, it is easy to determine how many tweets over a certain period used one or more hashtags. However, sorting out the polarity of these tweets (that is, whether they reference the hashtag in a positive, negative, or neutral manner) is more labor intensive and can require corroborating information outside the text of the tweet.[8] Other ambient data streams require even more-advanced processing to prepare for analysis; for example, image or video data must either be coded by a human or a sophisticated computational algorithm, and computational image recognition is still in its early stages.[9]

Social Media

From social networks (e.g., Facebook) to blogs (e.g., Tumblr) and miniblogs (e.g., Twitter) to communication platforms (e.g., WhatsApp), social media present a rich and attractive data set for exploitation and analysis. Indeed, social media have been useful in detecting support and opposition groups on Twitter in operational environments and in determining the types of content that networks are sharing with each other.[10]

However, social media also present multiple challenges. One is that different types of individuals—younger versus older, male versus female, introverted versus extroverted, etc.—participate in different types of social media platforms.[11] Another challenge is access to data: Many platforms have made third-party data access increasingly difficult over time; local gov-

---

[8] Darshan Barapatre, M. Janaki Meena, and S. P. Syed Ibrahim, "Twitter Data Classification Using Side Information," in V. Vijayakumar and V. Neelanarayanan, eds., *Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges*, New York: Springer International Publishing, 2016.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, Vol. 115, No. 3, 2015.

[10] Bodine-Baron et al., 2016.

[11] Teresa Correa, Amber Willard Hinsley, and Homero Gil de Zúñiga, "Who Interacts on the Web? The Intersection of Users' Personality and Social Media Use," *Computers in Human Behavior*, Vol. 26, No. 2, March 2010. Also see Maeve Duggan and Joanna Brenner, "The Demographics of Social Media Users—2012," Pew Research Center Internet & Technology website, February 14, 2013, pp. 1–14.

ernments may block access to in-country social media platforms; and individuals are moving to more-secure platforms (e.g., WhatsApp).

Enemy-focused analyses of these data face a particular challenge in that state and nonstate actors also post false or misleading information on social media platforms to deliberately confuse U.S. and coalition situational awareness and analysis.[12] Furthermore, such platforms as Twitter are full of automated advertising or other spam content that often must be filtered out before analysis.[13] This issue of the credibility of social media information also presents a significant analytic problem, although some computational approaches appear to be making headway, for example, in filtering for tweets that are more credible or less credible.[14] Nevertheless, social media have been useful, such as for detecting support and opposition groups on Twitter in operational environments, and for determining the types of content that networks are sharing with each other. Such inferences often require applying multiple automated tools, including lexical and social-network analysis.

Print and Online Media

Data pertinent to operational effects can also be gleaned from news sources and other publicly available documents and information sources. While the exploitation of such sources must be undertaken cautiously (because of the same sort of deliberate misinformation and filtering as for social media), the quantity and variety of online media are expanding rapidly.[15] As with social media data, the sheer volume of online and print media available creates challenges for filtering and binning information, often making automated approaches critical for detecting relevant information and rendering timely insights. To make sense of the often textually rich (and frequently foreign language) content requires either robust, high-powered automated lexical analysis or time-consuming and painstaking hand-coding by analysts with the appropriate language training.

Public-Use or Commercial Imagery

Satellite imagery from commercial and other government entities is becoming increasingly available and accurate. For example, satellite imagery can be used to estimate population density, population movement, and economic activity. For SOF lines of effort (LOEs) that involve protecting the population, encouraging migration back to war-torn areas, or encouraging economic recovery, these sorts of estimates can be very useful—especially in areas that are otherwise difficult to monitor or would require more military intelligence, surveillance, and reconnaissance assets than are available. For example, LandScan data use satellite imagery to

---

[12] Ulises A. Mejias and Nikolai E Vokuev, "Disinformation and the Media: The Case of Russia and Ukraine," *Media, Culture & Society*, Vol. 39, No. 7, 2017.

[13] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia, "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?" *IEEE Transactions on Dependable and Secure Computing*, Vol. 9, No. 6, 201.

[14] Sue E. Kase, Elizabeth K. Bowman, Tanvir Al Amin, and Tarek Abdelzaher, "Exploiting Social Media for Army Operations: Syrian Civil War Use Case," *Proceedings of SPIE*, Vol. 9122, July 2014, pp. 1–18.

[15] Stephen C. Mercado, "Sailing the Sea of OSINT in the Information Age," *Studies in Intelligence*, Vol. 48, No. 3, 2004.

estimate population density across the globe at a high level of geospatial resolution,[16] and nighttime satellite data have been used to estimate the humanitarian impact of the Syrian conflict.[17]

Data from Other Agencies

A wide variety of nonprofit organizations and NGOs maintain data sets on political violence and conflict (e.g., The Armed Conflict Location and Event Data Project, Peace Research Institute Oslo),[18] government corruption (e.g., Transparency International),[19] and related topics. A challenge with these data sets is that NGOs may not be operating in the AO due to security concerns, or they may be operating in limited areas of the AO. Furthermore, global data sets may lack the geographic specificity for accurate information on the effects of certain SOF operations.

Assessment-Specific Data

Since the Vietnam War, researchers have conducted purposeful assessments of SOF and other military operations through direct data collection in the AO in coordination with U.S. forces. A prominent—and later controversial—example of assessment-specific data from the Vietnam War was the RAND Hamlet Evaluation System, which derived scores for medical and educational services, drinking water availability, civic action, and other elements for each village and hamlet involved in Operation Sunrise to track progress and assess elements of stability operations.[20] Elements of this assessment approach, complemented by polling, interviews and focus groups with local police and other actors in the AO, surveys collected from Operational Detachment Alphas (ODAs), and other methods of direct data collection would later become part of the assessment of village stability operations in Afghanistan.[21]

There are multiple ways to collect assessment-specific data in an AO, and assessment teams often gravitate toward this form of data collection because it is possible to customize the format and type of data collected. However, assessment-specific data are generally the most burdensome to collect properly and rigorously, that is, so that the resulting data are useful. Tasking operators with completing surveys or other assessment materials adds to the burden on already very busy ODAs and other personnel. Direct interactions with locals require force protection

---

[16] Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Marie L. Urban, "LandScan USA: A High-Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics," *GeoJournal*, Vol. 69, Nos. 1–2, June 2007. Also see Edward A. Bright, Amy N. Rose, and Marie L Urban, *LandScan 2015 High-Resolution Global Population Data Set*, Oak Ridge, Tenn.: Oak Ridge National Laboratory, 2016.

[17] Christina Corbane, Thomas Kemper, Sergio Freire, Christophe Louvrier, and Martino Pesaresi, *Monitoring the Syrian Humanitarian Crisis with the JRC's Global Human Settlement Layer and Night-Time Satellite Data*, Luxembourg: Publications Office of the European Union, 2016.

[18] Kristine Eck, *A Beginner's Guide to Conflict Data: Finding and Using the Right Dataset*, Sweden: Uppsala University, 2005. Also see Clionadh Raleigh, and Caitriona Dowd, "Armed Conflict Location and Event Data Project (ACLED) Codebook," Armed Conflict Location and Event Data Project website, 2015.

[19] Transparency International, *Corruption Perceptions Index 2016*, Berlin, January 2017.

[20] Anders Sweetland, *Item Analysis of the HES (Hamlet Evaluation System)*, Santa Monica, Calif.: RAND Corporation, D-17634-ARPA/AGILE, 1968. The Hamlet Evaluation System was later criticized for producing "fabricated or irrelevant data that had no real utility at the tactical, operational, or strategic levels of operation and decision-making" (Connable, 2012, p. 131)

[21] Headquarters, Combined Joint Special Operations Task Force–Afghanistan, *Village Stability Operations and Afghan Local Police: Bottom-Up Counterinsurgency*, April 1, 2011. Also see Egel et al., 2016, and Brian Petit, "The Fight for the Village: Southern Afghanistan, 2010," *Military Review*, May–June 2011.

and put the personnel collecting data—and sometimes the locals, by association with U.S. or coalition personnel—at risk. Surveys, polling, interviews, and focus groups require careful design and execution to avoid or control for multiple forms of bias that can otherwise lead to faulty or completely incorrect inferences. Nevertheless, for certain assessment requirements, direct data collection from the AO may be the best (or perhaps the only) feasible approach.

Polling

One-on-one structured interviews with host nationals have been a prominent tool used in recent years to assess the operational effects on perceptions of security, economics, and governance.[22] This axis of data has strong face validity because long surveys can cover many different topics and provide at least the impression of rigor due to large numbers of responses. However, host nations do not always have suitable infrastructure and training to conduct rigorous and accurate polls, which may lead to data falsification or other problems with data quality. Additionally, conducting surveys in foreign environments incurs difficulties having to do with loss of meaning across languages and, sometimes, the lack of equivalent cognitive frames or constructs in different cultural environments.[23] Survey data analysis has usually involved examining demographic or area differences in patterns of responses and changes in response patterns over time. Survey data can be collected or analyzed in combination with specific axes of operational data (for example, in Afghanistan, village stability operation surveys were collected in areas with Afghan Local Police units) to measure operational effectiveness.

Interviews or Focus Groups

Interviews and focus groups use semistructured guides to obtain qualitative data from individuals or groups. Rather than the constrained, predetermined, and closed nature of most survey or polling questions, interviews and focus groups are designed to let respondents guide the direction of the conversation. The advantage of this more-open format is that the interviewer or focus group facilitator may discover unexpected insights that would have been lost or never talked about with a predetermined survey or polling format. Focus groups have the advantage of collecting data from multiple individuals at once but can also be subject to bias because of group dynamics.[24] Interviews, meanwhile, allow in-depth exploration of individuals' experiences. Both are subject to interviewer bias and require some training to properly administer. The rich qualitative data that interviews and focus groups produce require team-based coding to produce robust insights for assessment purposes.[25]

---

[22] In Afghanistan, there were at least five major polling data collection efforts conducted by CJPOTF, ISAF, MISTF, SOJTF-A, USAID—that would each conduct as many as 40,000 individual interviews.

[23] Susan Ervin and Robert T. Bower, "Translation Problems in International Surveys," *Public Opinion Quarterly*, Vol. 16, No. 4, January 1952; Brandon A. Kohrt, Mark J. D. Jordans, Wietse A. Tol, Nagendra P. Luitel, Sujen M. Maharjan, and Nawaraj Upadhaya, "Validation of Cross-Cultural Child Mental Health and Psychosocial Research Instruments: Adapting The Depression Self-Rating Scale and Child PTSD Symptom Scale in Nepal," *BMC Psychiatry*, Vol. 11, No. 127, 2011.

[24] Margaret C. Harrell and Melissa A. Bradley, *Data Collection Methods: Semi-Structured Interviews and Focus Groups*, Santa Monica, Calif.: RAND Corporation, TR-718-USG, 2009.

[25] Gery W. Ryan and Russell H. Bernard, "Techniques to Identify Themes," *Field Methods*, Vol. 15, No. 1, February 1, 2003.

Partner Capacity–Focused Assessments

Assessment and research teams have developed multiple frameworks for assessing the capacity and performance of partner forces. These usually involve mapping available data onto a variety of prestructured domains, sometimes involve direct interaction with military personnel in train, advise, and assist roles, and sometimes even involve direct observation of partner-force training and activities.

Composite Assessments

When SOF are involved in large campaigns and embedded with conventional forces, SOF operations are also assessed with respect to their contribution to theaterwide LOEs. Such assessments often use quantitative rating scales (or sometimes stoplight charts) to assess progress on strategic LOEs such as security and economic development. However, composite scores that summarize across vast regions and multiple units have come under fire for obscuring too much underlying complexity and are most useful when backed up with a rigorous, structured qualitative process for deriving scores.[26]

Bias Considerations

At the top of the bias cascade depicted in Table A.2 are concerns about sampling and response bias. Does the data source sufficiently represent the population of concern? For the answer to be yes, the first step is to eliminate *sampling bias*—that is, to give a fully representative sample of the population equal chances of participating in the data-collection effort at hand. This itself is usually practically impossible; a survey may require some basic literacy, for example, which will exclude less-educated members of the population. In operational assessments, some areas (and sometimes the areas of greatest interest) may be inaccessible for data collection for security reasons.

But even if it were possible to eliminate sampling bias, some individuals would be more likely to respond to the survey (or show up in another form of data collection); for example, youth are more likely to use Snapchat than older generations are. This is known as *response bias*. While it is impossible to fully eliminate sampling and response bias, these must be taken into account both when considering what data are worth collecting and also when analyzing data and making conclusions. Thus, the analyst should consider whether it is easy to detect

Table A.2
Bias Considerations

Area of Concern	Key Considerations
Sampling bias	Was a representative sample included or recruited?
Response bias	What sort of individuals were more likely to provide data? Is this quantifiable?
Social desirability bias	How have individuals self-censored or otherwise distorted the data set?
Culture	Do the data "translate" cross culturally?
Validity	Do the data mean what the analyst thinks they mean?
Reliability	Do the data represent a strong and stable signal?

[26] Connable, 2012; William P. Upshur, Jonathon W. Roginski, and David J. Kilcullen, "Recognizing Systems in Afghanistan: Lessons Learned and New Approaches to Operational Assessments," *Prism*, Vol. 3, No. 3, June 2012.

and understand the sources of any systematic bias in the data. The next level of the bias cascade are various forms of bias that result from the data collection process itself. During face-to-face data collection, individuals can tailor their responses to seem more socially acceptable to the interviewer—leading to *social desirability bias*;[27] for example, a respondent might choose to lie about behaviors they perceive the interviewer would disapprove of, such as domestic abuse, substance use, or support for local insurgents. Similarly, individuals often seek to put their best foot forward when posting to social media, limiting the type of content that appears.[28] Interviewers may also influence respondents through subtle facial expressions or posture or even through gender or appearance; during qualitative data collection, interviewers and focus group facilitators can shape the direction of conversation.

In foreign *cultural* contexts, certain constructs might not properly translate into the local social environment[29]—particularly Western academic or ethical constructs, such as *human rights*. If data collection does not take into account the local culture, responses on certain items might not represent what we expect them to.

Note also that collecting data takes time and resources. It is important to consider the ability to collect, clean, and prepare a high-quality data set (in which bias is limited and/or understood), given constraints in the operational environment.

Related to bias is the issue of *data validity*. It is important to consider whether a particular data stream presents a meaningful signal. Part of this equation is whether the signal itself is reliable—or stable to some degree. For example, (1) if survey respondents do not understand a survey item, (2) if their opinions on the item shift moment-to-moment, or (3) if they do not care enough or are not paying attention, they might provide one response one day and a different response the next. This type of data would have low test-retest reliability and, therefore, could be of minimal use in analysis.[30] A second and very important area of validity is content and construct validity; in plain terms, this means considering whether the data stream actually represents what the analyst thinks or claims it represents.[31] Consider a case in which an analyst is examining locals' complaints about local police forces. At face value, this might seem to be a valid signal of actual corruption and misbehavior. However, if the locals are mostly using claims of corruption to target social or political rivals, this would not be a strong or valid signal about local police behavior.

Analytical Methods

The analytical methods used in assessment must be designed to address the challenge of causality. Specifically, the methods need to convincingly demonstrate that a change in the appro-

---

[27] Maryon F. King and Gordon C. Bruner, "Social Desirability Bias: A Neglected Aspect of Validity Testing," *Psychology & Marketing*, Vol. 17, No. 2, February 2000.

[28] Even this varies by personality; for example, extraversion and narcissism increase the amount of self-referential posting, as illustrated in Eileen Y. L. Ong., Rebecca P. Ang, Jim C. M. Ho, Joylynn C. Y. Lim, Dion H. Goh, Chei Sian Lee, and Alton Y. K. Chua, "Narcissism, Extraversion and Adolescents' Self-Presentation on Facebook," *Personality and Individual Differences*, Vol. 50, No. 2, January 2011.

[29] Kohrt et al., 2011.

[30] Edward G. Carmines and Richard A. Zeller, *Quantitative Applications in the Social Sciences: Reliability and Validity Assessment*, Thousand Oaks, Calif.: SAGE Publications, 1979.

[31] Carmines and Zeller, 1979.

priate MOE—whether that change reflects an improvement or degradation in conditions—is attributable to the SOF activity being studied and not to other environmental factors. The dynamic settings in which SOF operations are conducted make this particularly important; a variety of other stakeholders (e.g., interagency actors, international community, host-nation elements, enemy forces) are also trying to influence these outcomes.[32] Indeed, there will be circumstances in which conditions in an area are degrading *in spite* of a successful SOF activity—e.g., a military information support operations effort successfully reduced enemy morale, but the arrival of a charismatic enemy leader caused a net improvement in morale—and vice versa.

Analytical results with causal interpretations can be produced using quantitative and qualitative methodologies. Methodologies that involve a blend of both are referred to as *mixed methods* or *triangulation*. No single analytical approach will be appropriate for all settings because each approach has strength and weaknesses. Table A.3 summarizes the conditions under which each type of methodology can produce causally defensible results and its strengths and weaknesses.

Quantitative

Quantitative analysis is an appropriate technique for assessing operations when two conditions are met. The first is that all relevant data—the activities SOF is conducting (inputs, or explanatory variables), indicators for the MOEs (outputs, or dependent variables), and environmental factors likely to influence the efficacy of SOF activities (control variables)—must be quantifiable. This is likely to encompass a large share of SOF assessment activity because most types of data are typically quantifiable (discussed earlier in this appendix, under "Data"). The second condition is that SOF activities are affecting a sufficient number of unique clusters of individuals (e.g., villages, military units).

Table A.3
Overview of Analytical Approaches Sources for Assessment Toolbox

Type	Requirements for Causal Inference	Strengths	Weaknesses
Quantitative	1. All data (inputs, outputs, controls) are quantifiable. 2. Large number of clusters (>50) impacted by the SOF activity 3. Longitudinal data available 4. Data for quasi-experimental methods available	• Statistical outputs • Replicability increases confidence	• Causality difficult to prove • Requires specialized training and tools • High risk of analytical bias
Qualitative	1. Longitudinal data are available 2 Detailed narrative data are available on MOEs and other correlates	• Explain why effects are being achieved • Support decisionmaking by understanding system	• Perceived as less objective • Time consuming
Triangulation	1. Either quantitative or qualitative analyses provides a causally defensible result	• Increases confidence by cross-validating findings • Adaptable to data constraints	• Time consuming • Resource intensive

[32] In a worst case, this may "devolve into simply gauging the environment, with little understanding as to how changes in that environment are the result of military activities" (Leo J. Blanken and Jason J. Lepore, "Principals, Agents, and Assessment," in Leo J. Blanken, Hy Rothstein, and Jason J. Lepore, eds., *Assessing War: The Challenge of Measuring Success and Failure*, Washington, D.C.: Georgetown University Press, 2015, p. 8).

In these cases, a *quasi-experimental, multivariate panel regression* is the preferred analytical technique. This approach has two defining characteristics.[33] The first is that it relies on longitudinal data, ideally with data available before SOF operations begin. This *ex ante* data serves as a baseline for the assessment, so that analysis focuses on how operations affect changes in the assessment's indicator variables. The second is that the approach includes data on areas with and without the SOF activity being assessed, which allows the analyst to use quasi-experimental methods. These quasi-experimental methods assess the effect of an activity by comparing changes over time in the area with the SOF activity against those in the areas without. We describe a step-by-step process for conducting a multivariate panel regression later.

Quantitative approaches offer several advantages over qualitative approaches. The first is that quantitative approaches produce statistically defensible results, which are typically perceived as being more objective those using qualitative methods. The second is that the analysis is replicable and employs standardized technical tools, which adds to the confidence in the results.

However, quantitative approaches have several important limitations. The first is that it is typically more difficult to establish causality using quantitative tools than most analysts appreciate, including many who rely heavily on quantitative methods. Indeed, some development professionals have argued that causal inference is not defensible if the treatment—in our case, the SOF activities—is not randomized, which is impossible in this context.[34] While others have argued that advanced statistical techniques can be used when randomization is not possible,[35] there is a substantial risk that analysis using observational data—i.e., assessment without randomization—can produce precise yet inaccurate statistical results.

The second major limitation is that analysis requires specialized technical training. The preferred tool of many analysts—Microsoft Excel—is unlikely to be sufficient for this kind of analysis.[36] Many military analysts now have training in relevant statistical toolkits (e.g., R, Stata) and can be taught these techniques relatively rapidly, although it may still prove easier to maintain a centralized capability for quantitative assessment at a larger headquarters element.[37]

---

[33] The following text provides a heuristic definition of what is referred to in the academic and development valuation literature as the *difference-in-difference approach*. For a more detailed, but still accessible, discussions, see Prashant Bharadwaj, *Quasi Experimental Methods: Difference in Differences*, San Diego, Calif.: University of California, 2010.

[34] For example, see Esther Duflo and Michael Kremer, "Use of Randomization in the Evaluation of Development Effectiveness" in George Pitman, Osvaldo Feinstein and Gregory Ingram, eds., *Evaluating Development Effectiveness*, New Brunwick: Transaction Publishers, 2005; Abhijit Banerjee, *Making Aid Work*, Cambridge, Mass.: MIT Press, 2007.

[35] For a detailed review of these advanced techniques, see Guido Imbens, "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, Vol. 86, No. 1, February 2004. For a discussion of their utilization in impact evaluations, see Martin Ravallion, "Evaluating Anti-Poverty Programs," in T. Paul Schultz and John Strauss, eds., *Handbook of Development Economics*, Vol. 4, Amsterdam: North-Holland, 2008; Angus Deaton, "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development," Cambridge, Mass.: National Bureau of Economic Research, Working Paper No. 14690, 2009.

[36] It is possible to conduct regression analysis in Microsoft Excel, but we found it extremely difficult to appropriately structure the data and then implement the regressions for the slightly more complicated multivariate *panel* regression (the panel element being what creates the complexity). Also, Excel is an inappropriate platform for combining many different types of data, which is often required in these settings.

[37] Progress is being made toward standardizing and automating quantitative analysis, making it easier for those without advanced statistical or mathematical training to run and interpret quantitative models, but these approaches are still being developed. See Cheney J. G. Drew, Vincent Poile, Rob Trubey, Gareth Watson, Mark Kelson, Julia Townson, Anne Rosser, Kerenza Hood, Lori Quinn, and Monica Busse, "Integrating Technology into Complex Intervention Trial Processes: A

Finally, because of the precise quantitative results and their sensitivity to the details of how the quantitative approach is implemented, quantitative techniques are unusually prone to analytical bias (see the "Analytical Biases" subsection later for an extended discussion).

Despite the requirement for an appropriately trained analyst to prepare the data and conduct the analysis, the required steps for conducting a multivariate panel regression are relatively straightforward:[38]

1. *Identify clusters of individuals affected by SOF activity.* The MOEs determined during Step 3 of the assessment process should imply the appropriate clusters for the quantitative analysis.[39] A cluster is a group of individuals who are anticipated to be very similar to each other based on geography, affiliation, or some other similar characteristic.[40] For our purposes, the most common type of cluster will be either geographic (e.g., village, town, or neighborhood in a city) or associational (e.g., same military platoon, same terrorist cell). Analysts should identify roughly 50 clusters affected by the SOF activity or activities being studied and at least 50 more that are not being affected that can be used as controls.
2. *Select appropriate time variable.* The assessment team should be able to specify a reasonable time over which it is reasonable to see change. A standard time variable for most SOF operations will be months, although this can be shorter or longer depending on the specifics of the operational-level headquarters being assessed.
3. *Collect and organize data longitudinally.* For each cluster, quantifiable data should be collected on the input (explanatory), output (dependent), and control variables for each of the units of time selected in step 2. These data can be assembled in Microsoft Excel following the example in Figure A.1, which is focused on measuring how dropping 500 pamphlets on Village 1 in March and April affected enemy will, which is measured, in this case, using social media on a scale from 0 to 100. Additionally, because we believe that the religious makeup and income of the villages may influence the efficacy of the SOF activity, we include these as control variables.
4. *Conduct the basic multivariate panel regression analysis.* With the data quantified and organized longitudinally by cluster in Figure A.1, a trained analyst can immediately conduct a multivariate panel regression to produce a statistically defensible estimate of the impact of the pamphlet drop (or whatever other SOF activity is being studied).

Case Study," *Trials*, Vol. 17, November 17, 2016; Anmol Rajpurohit, "Automatic Statistician and the Profoundly Desired Automation for Data Science," KDnuggets website, 2015.

[38] There are other ways to structure and analyze quantitative data. For example, a person-by-variables matrix can be rotated and analyzed as a variables-by-person matrix to determine relationships among individuals rather than variables. This technique can be powerful when estimating concordance in belief systems, sometimes called *cultural consensus*. For more information, see Susan C. Weller, "Cultural Consensus Theory: Applications and Frequently Asked Questions," *Field Methods*, Vol. 19, No. 4, November 1, 2007. Meanwhile, social network analysis often relies on a person-by-person matrix that describes the presence and strength of relationships among multiple individuals and has its own class of quantitative techniques and metrics. For additional information on social network analysis, see Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca, "Network Analysis in the Social Sciences," *Science*, Vol. 323, No. 5916, 2009.

[39] See Robinson, Egel, and Brown, 2019, Chapter Four.

[40] This degree of internal similarity is referred to as *intraclass correlation*. For a more detailed discussion, see Shersten Killip, Ziyad Mahfoud, and Kevin Pearce, "What Is an Intracluster Correlation Coefficient? Crucial Concepts for Primary Care Researchers," *Annals of Family Medicine*, Vol. 2, No. 3, May–June 2004.

Figure A.1
Example Organization of Longitudinal Data for Analysis

	A	B	C	D	E	F
1	Cluster	Time	Input	100 s	Control	
2			(# Pamphlets)	(Enemy will -- 100 is highest)	% Protestant	Average Income
3	Village 1	JAN	0	75	25	$1,000
4	Village 1	FEB	0	70	25	$1,000
5	Village 1	MAR	500	65	25	$1,000
6	Village 1	APR	500	60	25	$1,000
7	Village 2	JAN	0	65	30	$1,110
8	Village 2	FEB	0	60	30	$1,110
9	Village 2	MAR	0	55	30	$1,110
10	Village 2	APR	0	50	30	$1,110
11						

5. *Validate results with quasi-experimental matching techniques.* To increase confidence in these approaches, the quantitative analysis should be augmented with now-standard quasi-experimental matching techniques (e.g., propensity score matching). These methods can be easily implemented in any of the statistical computer programs that can conduct multivariate panel regressions.[41]

Qualitative

Qualitative analysis is an appropriate technique for assessing SOF when *either* of two conditions are met. The first is that at least one of the critical variables for analysis—input, output, or control—is inherently qualitative and not easy to convert into a quantity. The "Data" section earlier has an extended discussion of this issue and the cases in which it is likely to arise. The second is that SOF activities affect a relatively limited number of groups of individuals, referred to as *cases* in qualitative research.

In these cases, the preferred analytical technique is *process tracing*,[42] which uses both quantitative and qualitative longitudinal data to establish a causal narrative. This approach has three defining characteristics.[43] First, the analysis must be longitudinal, allowing an analysis of the sequencing of events, which is similar to the quantitative methods described earlier. Second, in addition to the outcome variables that are the focus of quantitative analysis, the

[41] We recommend the use of *propensity score weighting*, which is comparable to the other approaches in terms of its interpretation, which can be implemented by a trained analyst in about five minutes. A step-by-step discussion of how to implement this in R, a free and commonly used software package available in both unclassified and classified domains, is provided by Antonio Olmos and Priyalatha Govindasamy, "A Practical Guide for Using Propensity Score Weighting in R," *Practical Assessment, Research, & Evaluation*, Vol. 20, No. 13, June 2015.

[42] Process tracing is sometimes called *historical analysis* or *detailed case studies*. See, for example, Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton, N.J.: Princeton University Press, 1994.

[43] The description of process tracing is based on David Collier, "Understanding Process Tracing," *Political Science & Politics*, Vol. 44, No. 4, October 2011.

analysis must include intermediary evidence that is consistent with the proposed causal mechanism; intermediary evidence typically includes a blend of both quantitative and qualitative data. Third, the process tracing must include a detailed description of how conditions vary over time.

Qualitative approaches offer several important advantages over quantitative approaches. The first is that the process tracing approaches detailed below allow a nuanced understanding of what worked and what did not, which can provide the commander a critical tool for decisionmaking. The second is that the process tracing process itself—which generates additional indicators in addition to those that are the focus of the analysis—can provide a valuable tool for improving the assessment process throughout an operation.

Because qualitative approaches are typically perceived as subjective (they rely on expert judgment) and thus not reliable, they are used less frequently than quantitative methods.[44] A related challenge is that these methods are often labor intensive; conducting the detailed longitudinal analysis required for process tracing across multiple LOEs is time consuming for even an experienced analyst.

Unlike quantitative analysis, there is no standard way to conduct process tracing.[45] However, after conversations with practitioners of these approaches, we recommend the following:

1. *Characterize the historical evolution of the key MOE or indicator.* Drawing on both quantitative and qualitative data, this step focuses on developing a longitudinal history of the MOE or indicator, with a particular focus on when conditions changed—this could be either a change in the level (e.g., "economic activity increased by 20 percent") or a change in the rate (e.g., "economic growth increased from 0 percent to 3 percent").
2. *Describe the history of the relevant SOF activities.* This involves building an analysis longitudinal data set of the activities that are being evaluated.
3. *Assess temporal relationships among the data.* This step focuses on descriptively answering two questions:
    a. Are significant changes in the outcome variable associated with SOF activities?
    b. Are SOF activities consistently associated with changes in the outcome variable?
4. *Validate the outcome.* A final step examines the variety of other factors that might have influenced the observed change in the outcome measure, using the longitudinal nature of the data to rule out other possible explanations.

Triangulation

Triangulation refers to using more than one mode of data or analysis to view a problem or issue from multiple perspectives simultaneously.[46] In SOF assessments, the combination of multiple methods is particularly important because it builds confidence in the analysis. Specifically, it requires results from different data sets—each analyzed with different methods—to point in

---

[44] This is the core critique offered by King et al., 1994.

[45] In fact, practitioners of these approaches rarely use the specific approaches that methodologists have developed because these approaches involve confusing jargon that detracts from the analysis (authors' conversation with a practitioner). This is another reason these methods are less used.

[46] R. Burke Johnson, Anthony J. Onwuegbuzie, and Lisa A. Turner, "Toward a Definition of Mixed Methods Research," *Journal of Mixed Methods Research*, Vol. 1, No. 2, April 2007. Also see Michael Quinn Patton, "Enhancing the Quality and Credibility of Qualitative Analysis," *Health Services Research*, Vol. 34, No. 5, Pt. 2, December 1999.

a similar direction before making an assertion or conclusion. Adding data sets and analytic methods thus raises the bar for evidence and can also paint a fuller picture, revealing both patterns in data and potential causal relationships underlying the patterns.[47] Triangulation gives causally defensible results only if either the qualitative or quantitative method gives such results, but including both types of data collection and analytical approaches maximizes the possibility of a valid causal analysis and allows the user to cross-validate the findings from each analytical approach.

For example, an assessment team might first use structured polling to understand how local villages conceive of priorities for their community and for their families. However, polling data may indicate that certain regions show dramatically different patterns of community priorities from one polling period to the next—for example, how communities define *security*. The assessment team might decide to conduct qualitative follow-up work in regions with particular patterns. A simple but important finding from this work might be that certain regions do not conceive of security in the same coherent way military forces and more-stable, developed countries do. The assessment team could exclude polling responses related to security for these regions from current analyses and could then develop an alternative construct that is more grounded in local cultural understanding (for example, one more focused on protecting the family) for the next wave of polling.

One important aspect of triangulation is that the analysis team should deliberately try to challenge conclusions reached from one data set with those from a second (or third) data set.[48] For example, assume that the assessment team has been receiving very positive reports of the training level for the Atropia Special Forces.[49] To verify these reports, the assessment team decides to do a site visit to the training facility in Baku, observe Atropia Special Forces ODA training exercises, and interview ODA team leaders. The training exercises seem to go smoothly, but, in interviews, the ODA team leaders indicate that they are very frustrated with team command and control during night raids and think that many team members are unskilled with night vision equipment. The assessment team can now present the overall positive news with these important caveats and limitations.

The SOF assessment team should proceed through the following steps to conduct triangulation:

1. Conduct a rigorous quantitative analysis of the available indicators (within time, data, and resource constraints).
2. Look for counterintuitive and unexpected patterns in the quantitative data and develop a set of questions generated by this pattern analysis.
3. Brainstorm ways of collecting additional qualitative data that could challenge existing assumptions and answer the questions generated by the quantitative analysis.
4. Collect qualitative data (through interviews, focus groups, ethnographic observation, etc.)

---

[47] In the academic literature, this is called either *triangulation* or *mixed methods*.

[48] See Patton, 1999, pp. 654–656.

[49] See Robinson, Egel, and Brown, 2019, Chapter Five, for the relevant scenario.

5. Structure and analyze qualitative data using team-based thematic analysis or other methods to sort and group unstructured qualitative data in ways that allow the team to draw inferences.
6. Deliberately look for ways the quantitative and qualitative analyses conflict; for example, conduct a concerted search for interview transcripts that suggest a pattern opposite the one in the quantitative data.
7. Examine qualitative data for evidence that questions prevailing assumptions about patterns of cause and effect in the AO (i.e., evidence that challenges the CONOP), helps add context or caveats to prevailing assumptions, or even confirms existing assumptions.
8. Produce a modified assessment product that takes the lessons learned during triangulation into account.

Analytical Biases

Assessment teams should consider how vulnerable certain methods are to subtle manipulation by the analyst (*analytic sensitivity*). Regressions with small numbers of observations are highly sensitive to small changes—adding or subtracting variables, filtering out small numbers of observations, and the like.[50] In contrast, regressions with large numbers of observations often produce results that are statistically meaningful but reflect such weak relationships that their real-world meaning may be limited. Assessment teams should be careful to avoid advanced statistical techniques that could overfit a small or idiosyncratic data set.[51]

Qualitative analysis can be particularly vulnerable to analytic bias, especially when qualitative analysis occurs via less structured processes (e.g., "I will read these interview transcripts and write down the important and common themes"). Analysts will, of course, look for and notice or document themes with which they are already familiar, guided either by prior training or direct experience with the population or region being analyzed. Assessment teams can ameliorate this tendency for interpretive drift and bias when analyzing qualitative data by implementing structured analytic processes with teams.[52]

Two additional overarching issues should be considered when choosing analytic approaches. First, it is important to consider whether the analytic method chosen includes an exploratory phase intended to discover unexpected patterns in the data (*exploration and discovery*). This is particularly important when assessment teams expect that there are unknown unknowns—that is, causal dynamics or contextual factors in the operational environment that no one has heard of or anticipated. Many SOF operational assessments can encounter a considerable amount of discovery during operations, so it is important to include at least one analytic method that is capable of novel discovery. A qualitative analytic approach that defines all the codes at the outset, for example, might not allow for this type of discovery. Similarly, a statistical approach that narrowly defines the variables it includes based on expected causal factors could also miss this type of discovery. But some quantitative approaches are capable

---

[50] Samprit Chatterjee and Ali S. Hadi, *Sensitivity Analysis in Linear Regression*, Hoboken, N.J.: John Wiley & Sons, 2009.

[51] Michael A. Babyak, "What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models," *Psychosomatic Medicine*, Vol. 66, No. 3, 2004.

[52] H. Russell Bernard and Gery W. Ryan, *Analyzing Qualitative Data: Systematic Approaches*, Thousand Oaks, Calif.: SAGE Publications, 2009.

of discovery, including automated machine-learning approaches,[53] and exploratory statistical techniques, such as regression trees.[54]

Second, the assessment team should consider whether any of the analytic techniques selected have the ability to eliminate alternative hypotheses and/or confirm hypotheses: Can these tools (along with the available data) rule out alternative explanations and narrow in on *causal certainty*? Part of this, of course, depends on the range and quality of data collected. But part relies also on the analytic techniques applied. Analysts can employ quantitative techniques exploiting longitudinal data and statistically adjusting for multiple sources of contextual influence (if data are adequate), but causal inference is not limited to quantitative approaches. In particular, attention to causal-process observation in qualitative data can be used along with logical inference to build causal stories and eliminate or confirm hypotheses using process tracing.[55] Also, triangulation among multiple qualitative and quantitative analytic approaches can help build a more holistic causal picture of complex phenomena.[56]

---

[53] Indranil Bose and Radha K. Mahapatra, "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests," *Psychological Methods*, Vol. 14, No. 4, December 2009.

[54] Carolin Strobl, James Malley, and Gerhard Tutz, "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests," *Psychological Methods*, Vol. 14, No. 4, December 2009, p. 323.

[55] For a brief introduction and tutorial, see Collier, 2011.

[56] For an example, see Ryan A. Brown, David P. Kennedy, Joan S. Tucker, Daniela Golinelli, and Suzanne L. Wenzel, "Monogamy on the Street: A Mixed Methods Study of Homeless Men," *Journal of Mixed Methods Research*, Vol. 7, No. 4, October 2013.

APPENDIX B

Criteria for Measure of Effectiveness Selection

Below is a list of criteria for selecting MOEs derived from joint doctrine. This provides a fuller explication of the MOE selection criteria described in Chapter Three of the main report.[1]

Relevance

This is a fundamental qualifying criterion for MOEs. MOEs must have a logical—and ideally *causal*—connection to the objectives and desired end states of a planned operation.[2] For example, the "population support for democracy" MOE is not directly relevant to the "reduce enemy's capabilities for lethal effects" objective. While increased security *might* allow democracy to flourish, this would be a second- or third-order effect, only distantly related to the objective at hand.

Measurability

A brainstorming session can produce a wide variety of possible MOEs, but not all might be measurable with available data streams or current methods of data collection. For example, while it would be useful to know the differences in will to fight among top insurgent leaders, this may be very difficult (or nearly impossible) to assess. A key inclusion criterion for MOEs is that the assessment team can derive measurable indicators for that MOE.[3]

Malleability

To track whether SOF are doing the right things, it is also important for the assessment team to be able to derive indicators for each MOE that can be expected to change within the period during which SOF operations are carried out.[4] For example, while a long-term goal of U.S.

---

[1] See Robinson, Egel, and Brown, 2019.

[2] Joint Doctrine Note (JDN) 1-15, *Operations Assessment*, Washington, D.C.: Joint Staff, January 15, 2015, p. A-15; Joint Staff, *Commander's Handbook for Assessment Planning and Execution*, Vers. 1.0, Suffolk, Va.: Joint Staff J-7, Joint and Coalition Warfighting, September 9, 2011, p. II-1.

[3] See Joint Staff, 2011, p. III-6.

[4] JDN 1-15, 2015, p. A-16; Joint Staff, pp. II-1 and II-4.

intervention in a developing country might be a stable, prosperous economy, it is unlikely that significant progress toward this national goal could be made within the scope of limited SOF intervention. However, an assessment team might be able to assess economic stability and growth in select locales by assessing traffic in local markets.

Distinctness

When finalizing the MOEs to be kept, it is necessary for them to be as mutually exclusive as possible.[5] MOEs for a particular objective should not be constructed in ways that blur the boundaries between MOEs. For example, "decreased population support for the enemy" and "decreased buy-in to enemy narrative" are too conceptually related—and likely even on the same causal path—and thus should not be included as separate MOEs for the same objective.

[5] JDN 1-15, 2015, p. A-15.

Abbreviations

AO	area of operation
CONOP	concept of operation
JDN	Joint Doctrine Note
LOE	line of effort
MOE	measure of effectiveness
MOP	measure of performance
NGO	nongovernmental organization
ODA	Operational Detachment Alpha
SIGACT	significant activity
SITREP	situation report
SOF	special operations forces

References

Babyak, Michael A., "What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models," *Psychosomatic Medicine*, Vol. 66, No. 3, 2004, pp. 411–421. As of September 21, 2017:
http://journals.lww.com/psychosomaticmedicine/Fulltext/2004/05000/What_You_See_May_Not_Be_What_You_Get__A_Brief.21.aspx

Banerjee, Abhijit, *Making Aid Work*, Cambridge, Mass.: MIT Press, 2007.

Barapatre, Darshan, M. Janaki Meena, and S. P. Syed Ibrahim, "Twitter Data Classification Using Side Information," in V. Vijayakumar and V. Neelanarayanan, eds., *Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges*, New York: Springer International Publishing, 2016, pp. 363–368. As of September 21, 2017:
http://dx.doi.org/10.1007/978-3-319-30348-2_31

Bernard, H. Russell, and Gery W. Ryan, *Analyzing Qualitative Data: Systematic Approaches*, Thousand Oaks, Calif.: SAGE Publications, 2009.

Bhaduri, Budhendra, Edward Bright, Phillip Coleman, and Marie L. Urban, "LandScan USA: A High-Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics," *GeoJournal*, Vol. 69, Nos. 1–2, June 2007, pp. 103–117. As of September 21, 2017:
http://dx.doi.org/10.1007/s10708-007-9105-9

Bharadwaj, Prashant, *Quasi Experimental Methods: Difference in Differences*, San Diego, Calif.: University of California, 2010. As of September 21, 2017:
http://cega.berkeley.edu/assets/cega_events/36/Quasi-Experimental_Methods.pdf

Blanken, Leo J., and Jason J. Lepore, "Principals, Agents, and Assessment," in Leo J. Blanken, Hy Rothstein, and Jason J. Lepore, eds., *Assessing War: The Challenge of Measuring Success and Failure*, Washington, D.C.: Georgetown University Press, 2015.

Bodine-Baron, Elizabeth, Todd C. Helmus, Madeline Magnuson, and Zev Winkelman, *Examining ISIS Support and Opposition Networks on Twitter*, Santa Monica, Calif.: RAND Corporation, RR-1328-RC, 2016. As of September 22, 2017:
https://www.rand.org/pubs/research_reports/RR1328.html

Borgatti, Stephen P., Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca, "Network Analysis in the Social Sciences," *Science*, Vol. 323, No. 5916, 2009, pp. 892–895.

Bose, Indranil, Radha K. Mahapatra, "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests," *Psychological Methods*, Vol. 14, No. 4, December 2009.

Bright, Edward A., Amy N. Rose, and Marie L. Urban, *LandScan 2015 High-Resolution Global Population Data Set*, Oak Ridge, Tenn.: Oak Ridge National Laboratory, 2016.

Brown, Ryan A., David P. Kennedy, Joan S. Tucker, Daniela Golinelli, and Suzanne L. Wenzel, "Monogamy on the Street: A Mixed Methods Study of Homeless Men," *Journal of Mixed Methods Research*, Vol. 7, No. 4, October 2013. As of July 26, 2018:
http://journals.sagepub.com/doi/pdf/10.1177/1558689813480000

Carmines, Edward G., and Richard A. Zeller, *Quantitative Applications in the Social Sciences: Reliability and Validity Assessment*, Thousand Oaks, Calif.: SAGE Publications, 1979.

Chatterjee, Samprit, and Ali S. Hadi, *Sensitivity Analysis in Linear Regression*, Hoboken, N.J.: John Wiley & Sons, 2009.

Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia, "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?" *IEEE Transactions on Dependable and Secure Computing*, Vol. 9, No. 6, 2012, pp. 811–824.

Collier, David, "Understanding Process Tracing," *Political Science & Politics*, Vol. 44, No. 4, October 2011, pp. 823–830. As of July 24, 2018:
https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/understanding-process-tracing/183A057AD6A36783E678CB37440346D1

Connable, Ben, *Embracing the Fog of War: Assessment and Metrics in Counterinsurgency*, Santa Monica, Calif.: RAND Corporation, MG-1086-DOD, 2012. As of September 22, 2017:
https://www.rand.org/pubs/monographs/MG1086.html

Corbane, Christina, Thomas Kemper, Sergio Freire, Christophe Louvrier, and Martino Pesaresi, *Monitoring the Syrian Humanitarian Crisis with the JRC's Global Human Settlement Layer and Night-Time Satellite Data*, Luxembourg: Publications Office of the European Union, 2016.

Correa, Teresa, Amber Willard Hinsley, and Homero Gil de Zúñiga, "Who Interacts on the Web? The Intersection of Users' Personality and Social Media Use," *Computers in Human Behavior*, Vol. 26, No. 2, March 2010, pp. 247–253. As of September 21, 2017:
http://www.sciencedirect.com/science/article/pii/S0747563209001472

Deaton, Angus, "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development," Cambridge, Mass.: National Bureau of Economic Research, Working Paper No. 14690, 2009.

Drew, Cheney J. G., Vincent Poile, Rob Trubey, Gareth Watson, Mark Kelson, Julia Townson, Anne Rosser, Kerenza Hood, Lori Quinn, and Monica Busse, "Integrating Technology into Complex Intervention Trial Processes: A Case Study," *Trials*, Vol. 17, November 17, 2016, p. 551.

Duflo, Esther and Michael Kremer, "Use of Randomization in the Evaluation of Development Effectiveness" in George Pitman, Osvaldo Feinstein and Gregory Ingram, eds., *Evaluating Development Effectiveness*, New Brunwick: Transaction Publishers, 2005, pp. 205–231.

Duggan, Maeve, and Joanna Brenner, "The Demographics of Social Media Users—2012," Pew Research Center Internet & Technology website, February 14, 2013. As of July 24, 2018:
http://www.pewinternet.org/2013/02/14/the-demographics-of-social-media-users-2012/

Eck, Kristine, *A Beginner's Guide to Conflict Data: Finding and Using the Right Dataset*, Sweden: Uppsala University, 2005. As of September 21, 2017
http://www.pcr.uu.se/digitalAssets/15/a_15928-f_UCDP_paper1.pdf

Egel, Daniel, Charles P. Ries, Ben Connable, Todd C. Helmus, Eric Robinson, Isaac Baruffi, Melissa A. Bradley, Kurt Card, Kathleen Loa, Sean Mann, Fernando Sedano, Stephan B. Seabrook, and Robert Stewart, *Investing in the Fight: Assessing the Use of the Commander's Emergency Response Program in Afghanistan*, Santa Monica, Calif.: RAND Corporation, RR-1508-OSD, 2016. As of September 22, 2017:
https://www.rand.org/pubs/research_reports/RR1508.html

Ervin, Susan, and Robert T. Bower, "Translation Problems in International Surveys," *Public Opinion Quarterly*, Vol. 16, No. 4, January 1952, pp. 595–604. As of September 21, 2017:
http://dx.doi.org/10.1086/266421

Field Manual 6-99, *U.S. Army Report and Message Formats*, Washington, D.C.: Headquarters, Department of the Army, August 19, 2013. As of September 21, 2017:
http://www.apd.army.mil/epubs/DR_pubs/DR_a/pdf/web/fm6_99.pdf

Harrell, Margaret C., and Melissa A. Bradley, *Data Collection Methods: Semi-Structured Interviews and Focus Groups*, Santa Monica, Calif.: RAND Corporation, TR-718-USG, 2009. As of September 22, 2017:
https://www.rand.org/pubs/technical_reports/TR718.html

Headquarters, Combined Joint Special Operations Task Force–Afghanistan, *Village Stability Operations and Afghan Local Police: Bottom-Up Counterinsurgency*, April 1, 2011.

Imbens, Guido W., "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, Vol. 86, No. 1, February 2004.

JDN—*See* Joint Doctrine Note.

Johnson, R. Burke, Anthony J. Onwuegbuzie, and Lisa A. Turner, "Toward a Definition of Mixed Methods Research," *Journal of Mixed Methods Research*, Vol. 1, No. 2, April 2007, pp. 112–133.

Joint Doctrine Note 1-15, *Operation Assessment*, Washington, D.C.: Joint Staff, January 15, 2015.

Joint Staff, *Commander's Handbook for Assessment Planning and Execution*, Vers. 1.0, Suffolk, Va.: Joint Staff J-7, Joint and Coalition Warfighting, September 9, 2011. As of July 25, 2018: http://www.jcs.mil/Portals/36/Documents/Doctrine/pams_hands/assessment_hbk.pdf

Kaisler, Stephen, Frank Armour, J. Alberto Espinosa, and William Money, "Big Data: Issues and Challenges Moving Forward," paper presented at the 46th Hawaii International Conference on System Sciences, January 7–10, 2013, pp. 995–1004.

Kase, Sue E., Elizabeth K. Bowman, Tanvir Al Amin, and Tarek Abdelzaher, "Exploiting Social Media for Army Operations: Syrian Civil War Use Case," *Proceedings of SPIE*, Vol. 9122, July 2014, pp. 1–18.

Killip, Shersten, Ziyad Mahfoud, and Kevin Pearce, "What Is an Intracluster Correlation Coefficient? Crucial Concepts for Primary Care Researchers," *Annals of Family Medicine*, Vol. 2, No. 3, May–June 2004, pp. 204–208.

King, Gary, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton, N.J.: Princeton University Press, 1994.

King, Maryon F., and Gordon C. Bruner, "Social Desirability Bias: A Neglected Aspect of Validity Testing," *Psychology & Marketing*, Vol. 17, No. 2, February 2000, pp. 79–103.

Kohrt, Brandon A., Mark J. D. Jordans, Wietse A. Tol, Nagendra P. Luitel, Sujen M. Maharjan, and Nawaraj Upadhaya, "Validation of Cross-Cultural Child Mental Health and Psychosocial Research Instruments: Adapting the Depression Self-Rating Scale and Child PTSD Symptom Scale in Nepal," *BMC Psychiatry*, Vol. 11, No. 127, 2011, pp. 1–17. As of September 21, 2017: http://dx.doi.org/10.1186/1471-244X-11-127

Mejias, Ulises A., and Nikolai E. Vokuev, "Disinformation and the Media: The Case of Russia and Ukraine," *Media, Culture & Society*, Vol. 39, No. 7, 2017, pp. 1027–1042. As of September 21, 2017: http://journals.sagepub.com/doi/abs/10.1177/0163443716686672

Mercado, Stephen C., "Sailing the Sea of OSINT in the Information Age," *Studies in Intelligence*, Vol. 48, No. 3, 2004.

Olmos, Antonio, and Priyalatha Govindasamy, "A Practical Guide for Using Propensity Score Weighting in R," *Practical Assessment, Research, & Evaluation*, Vol. 20, No. 13, June 2015.

Ong, Eileen Y. L., Rebecca P. Ang, Jim C. M. Ho, Joylynn C. Y. Lim, Dion H. Goh, Chei Sian Lee, and Alton Y. K. Chua, "Narcissism, Extraversion and Adolescents' Self-Presentation on Facebook," *Personality and Individual Differences*, Vol. 50, No. 2, January 2011, pp. 180–185. As of September 21, 2017: http://www.sciencedirect.com/science/article/pii/S0191886910004654

Patton, Michael Quinn, "Enhancing the Quality and Credibility of Qualitative Analysis," *Health Services Research*, Vol. 34, No. 5, Pt. 2, December 1999, pp. 1189–1208. As of September 21, 2017: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1089059/

Petit, Brian, "The Fight for the Village: Southern Afghanistan, 2010," *Military Review*, May–June 2011.

Rajpurohit, Anmol, "Automatic Statistician and the Profoundly Desired Automation for Data Science," KDnuggets website, 2015. As of July 26. 2018: https://www.kdnuggets.com/2015/02/automated-statistician-data-science.html

Raleigh, Clionadh, and Caitriona Dowd, "Armed Conflict Location and Event Data Project (ACLED) Codebook," Armed Conflict Location and Event Data Project website, 2015. As of September 21, 2017: http://www.acleddata.com/wp-content/uploads/2015/01/ACLED_Codebook_2015.pdf

Ravallion, Martin, "Evaluating Anti-Poverty Programs," in T. Paul Schultz and John Strauss, eds., *Handbook of Development Economics*, Vol. 4, Amsterdam: North-Holland, 2008.

Robinson, Linda, Daniel Egel, and Ryan Andrew Brown, *Measuring the Effectiveness of Special Operations*, Santa Monica, Calif.: RAND Corporation, RR-2504-A, 2019. As of October 2019: https://www.rand.org/pubs/research_reports/RR2504.html

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, Vol. 115, No. 3, 2015, pp. 211–252. As of September 21, 2017: http://dx.doi.org/10.1007/s11263-015-0816-y

Ryan, Gery W., and Russell H. Bernard, "Techniques to Identify Themes," *Field Methods*, Vol. 15, No. 1, February 1, 2003, pp. 85–109. As of September 21, 2017: http://fmx.sagepub.com/content/15/1/85.abstract

Strobl, Carolin, James Malley, and Gerhard Tutz, "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests," *Psychological Methods*, Vol. 14, No. 4, December 2009, pp. 323–348.

Sweetland, Anders, *Item Analysis of the HES (Hamlet Evaluation System)*, Santa Monica, Calif.: RAND Corporation, D-17634-ARPA/AGILE, 1968. As of September 22, 2017: https://www.rand.org/pubs/documents/D17634.html

Transparency International, *Corruption Perceptions Index 2016*, Berlin, January 2017. As of September 21, 2017: http://www.transparency.org/whatwedo/publication/corruption_perceptions_index_2016

Upshur, William P., Jonathon W. Roginski, and David J. Kilcullen, "Recognizing Systems in Afghanistan: Lessons Learned and New Approaches to Operational Assessments," *Prism*, Vol. 3, No. 3, June 2012, pp. 87–104.

Weller, Susan C., "Cultural Consensus Theory: Applications and Frequently Asked Questions," *Field Methods*, Vol. 19, No. 4, November 1, 2007, pp. 339–368. As of September 21, 2017: http://fmx.sagepub.com/cgi/content/abstract/19/4/339