SAND2017-6337C

SAND2017-XXXX

Criteria for Comparative Evaluations of Treaty Verification Monitoring Systems

Jay K. Brotz[1], Angela Waterworth[2], Jacob Benz[2], Danielle Hauck[3], Dan Krementz[4], Gary Cockrell[5], George Weeks[6], Matthew McDougall[2]

[1]Sandia National Laboratories, Albuquerque, NM
[2]Pacific Northwest National Laboratory, Richland, WA
[3]Los Alamos National Laboratory, Los Alamos, NM
[4]Savannah River National Laboratory, Aiken, SC
[5]Pantex Plant, Amarillo, TX
[6]Crimson Sky Consulting, Kennewick, WA

Abstract

Analyzing options for a monitoring system, including options for technologies, data, and procedures, to meet a specific treaty verification scenario can be complicated and difficult. Competing interests lead to a multitude of tradeoffs, especially considering the tension between the needs of a host (the party attempting to prove compliance) and a monitor (the party conducting the inspection to determine compliance). Evaluation criteria may offer an agreed and established standard against which to evaluate competing options. Effective criteria will identify and prioritize factors of importance to both parties and should be unambiguous, independent from each other, collectively exhaustive, and measurable (whether this is quantitative or by a systematic subjective evaluation). When used for evaluation of options, criteria help to create an unbiased and consistent framework that can assist in selecting the optimal solution among a set of competing alternatives.

In 2015, a team of subject matter experts from across the U.S. Department of Energy (DOE) national laboratory complex completed a study to identify and assess potential chain of custody (CoC) capabilities for monitored warhead dismantlement. As part of that effort, the team developed an initial set of evaluation criteria to enable the comparative assessment of three hypothetical monitoring regimes that were based on different levels of technological capability. The team provided lessons learned from their effort, which included recommendations to refine the evaluation criteria and the way in which they are used to assess a monitoring system. More recently, a team of new and returning members from across the DOE national laboratory complex has reconsidered these criteria. The major objective of this effort is to consider the recommendations and lessons learned from the 2015 effort and insight from other relevant efforts to improve the previously developed evaluation criteria and strengthen their usefulness for evaluating monitoring regime options. In this paper, we discuss the criteria identified as part of this effort and explain how they can be used to evaluate monitoring system options. Additionally, we will discuss the method we followed to develop the criteria and considerations and observations from the effort.

---

[1] Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2017-XXXX

Introduction

Policymakers working toward next-generation nuclear arms control need the ability to compare options for monitoring systems against the goal of a particular verification regime. These goals, and therefore these systems, are often more complex than their treaty verification precedents, contributing to challenges in delivering viable options to negotiators. In this work, we formulate evaluation criteria to provide a framework to assess and compare monitoring systems – made up of data, technologies, and procedures – implemented to support treaty verification. We will first define two sets of criteria – one from the host perspective and one from the monitor perspective – that represent the values and objectives that all relevant stakeholders have with respect to the verification problem. We then identify, based on stakeholder evaluations, the relative importance, or weighting, for each criterion. Finally, we apply the weighted criteria to each monitoring system option and roll up each set to arrive at a normalized score for each monitoring system. These scores are a tool that system designers and policymakers can use to understand the benefits and risks of potential monitoring systems, and this method functions best when all of the criteria, weights, and calculations are presented to decision makers in a transparent way so that the primary contributors to the final score are apparent.

This effort builds on an attempt to identify criteria for comparing monitoring systems for a monitored dismantlement regime conducted in 2015. The new evaluation criteria are significantly different from the previous; the most obvious difference is the development of one set of criteria to be used to evaluate the monitoring regime from the perspective of the monitor (the party that is attempting to confirm compliance with treaty obligations) and a second set of criteria from the perspective of the host (the party that is trying to demonstrate compliance and prove to the monitor that treaty obligations have been met). One other major difference is structuring the criteria according to the benefit in achieving their interests and the cost associated with the implementation of the regime, for both the host and monitor. In this effort, we assume that the host is complying with the treaty when considering the host perspective, but that the monitor is not necessarily acting honestly. Likewise, when considering the monitor perspective, we assume that the monitor is not attempting to tamper with anything or recover sensitive information, but that the host in not necessarily acting honestly.

Evaluation Criteria

We organize the evaluation criteria according to the following hierarchy:

> **Values** – At the highest level, values represent what is important to the host and the monitor. These are organized according to benefits that support the host or monitor in achieving their interests and costs associated with implementing the regime.
>
> **Objectives** – Objectives capture more concretely what the host and monitor want to achieve. While values capture what is important to a stakeholder, objectives capture what the stakeholder wants to do with a monitoring system.

Criteria – Criteria are observable features or characteristics; the efficacy of a monitoring system in supporting the objectives and values can be evaluated by examining the features and characteristics of the monitoring system.

Metrics – Metrics are measureable factors related to the observables that enable the objective assessment of the criteria. Ultimately, all metrics will be quantifiable; while some metrics can be quantified, such as costs, further research is needed to develop approaches to evaluate other metrics, such as those related to the ability to detect or the likelihood of detecting tampering. Most metrics are expected to be evaluated subjectively, though this can be done systematically, as will be discussed in the next section. Metrics are traceable through the criteria and objectives to the host or monitor's values.

Building on the team's effort to identify criteria for a monitoring system used in a monitored dismantlement scenario in 2015, we present the following criteria hierarchy separately for the monitor and host perspectives. Figure 1 shows the monitor values and objectives, with Figure 2 and Figure 3 showing the monitor criteria and metrics. Figure 4 shows the host values and objectives, with Figure 5, Figure 6, and Figure 7 showing the host criteria and metrics. In each diagram, the values are denoted [MX] or [HX], the objectives are denoted [MX.X] or [HX.X], the criteria are denoted [MX.X.X] or [HX.X.X], and the metrics are denoted [MX.X.X.X] or [HX.X.X.X]. The numbers shown in each box identify criteria weightings, which will be discussed below. Although the criteria hierarchy was developed with a monitored dismantlement scenario in mind, it is generally applicable to many treaty verification systems that involve inspections.

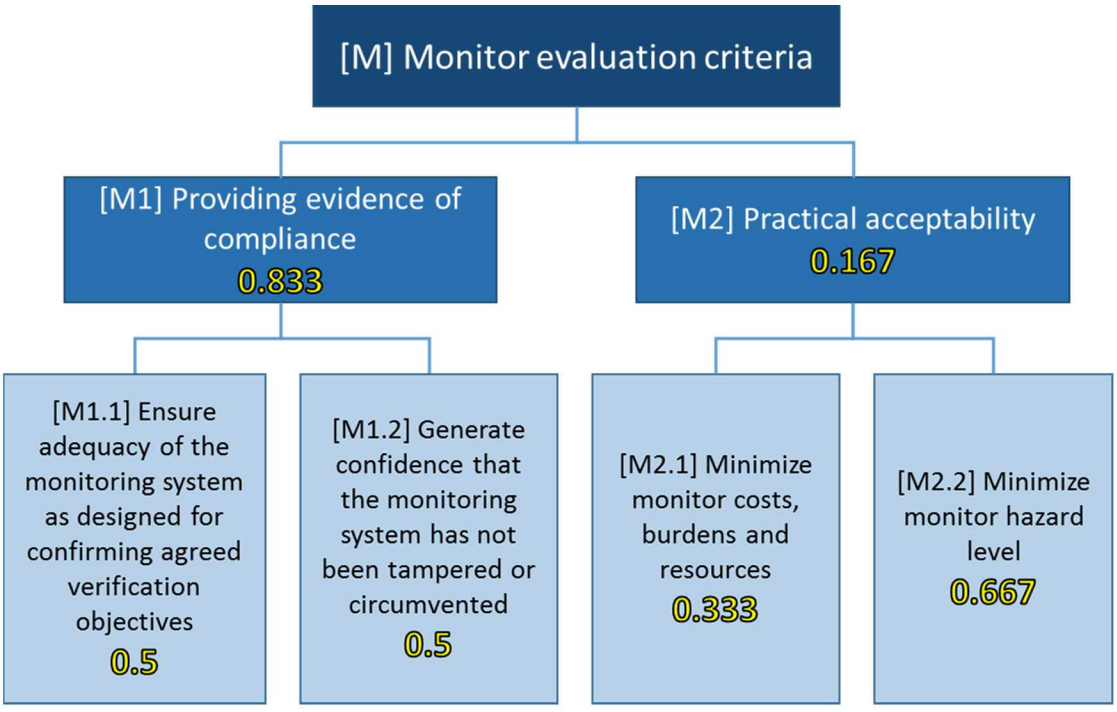

Figure 1. Monitor Values and Objectives

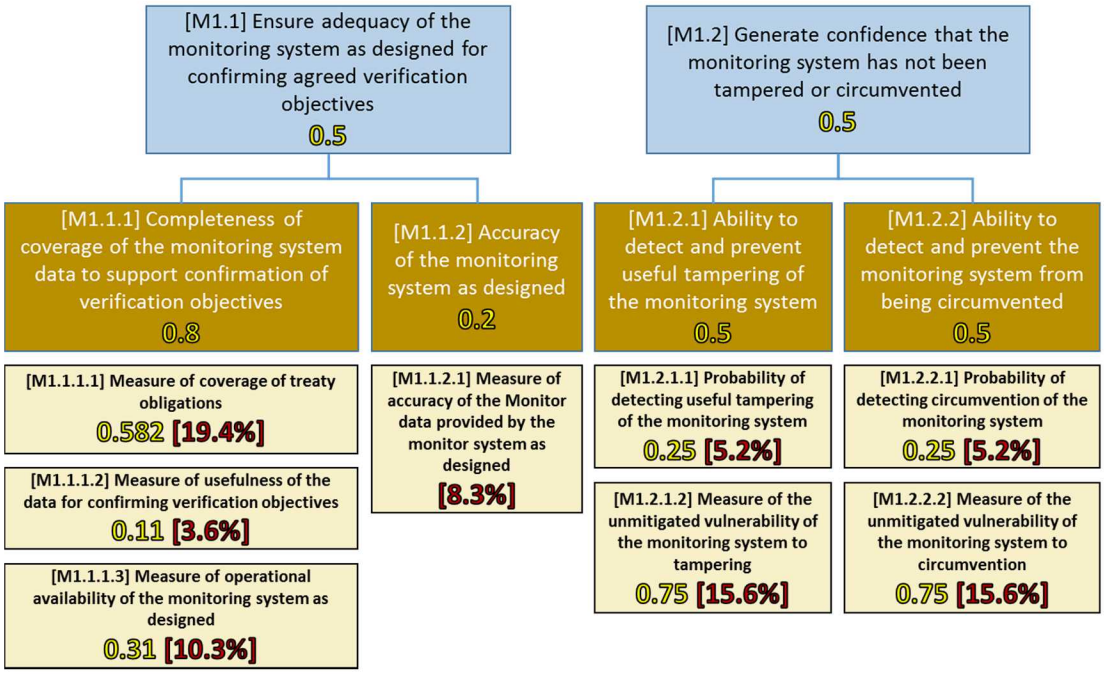

Figure 2. Monitor Criteria and Metrics under Providing Evidence of Compliance

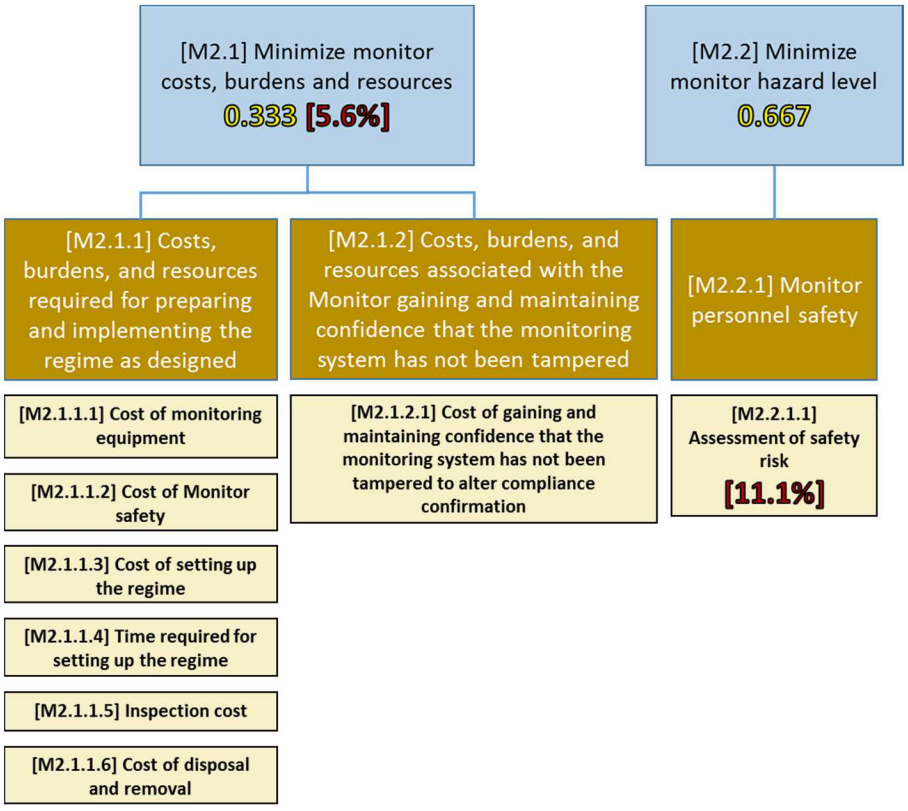

Figure 3. Monitor Criteria and Metrics under Practical Acceptability

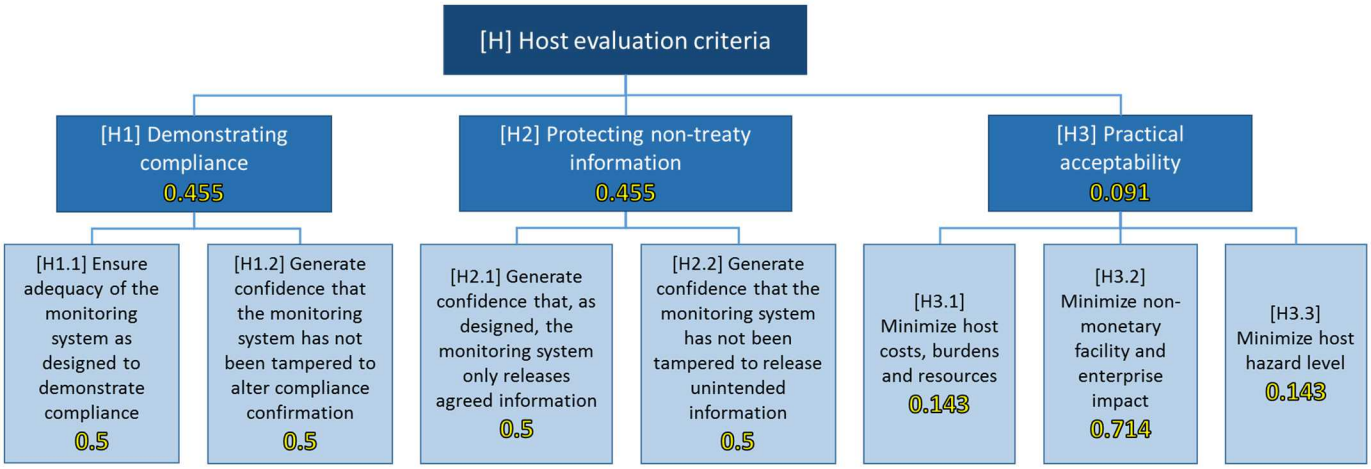

Figure 4. Host Values and Objectives

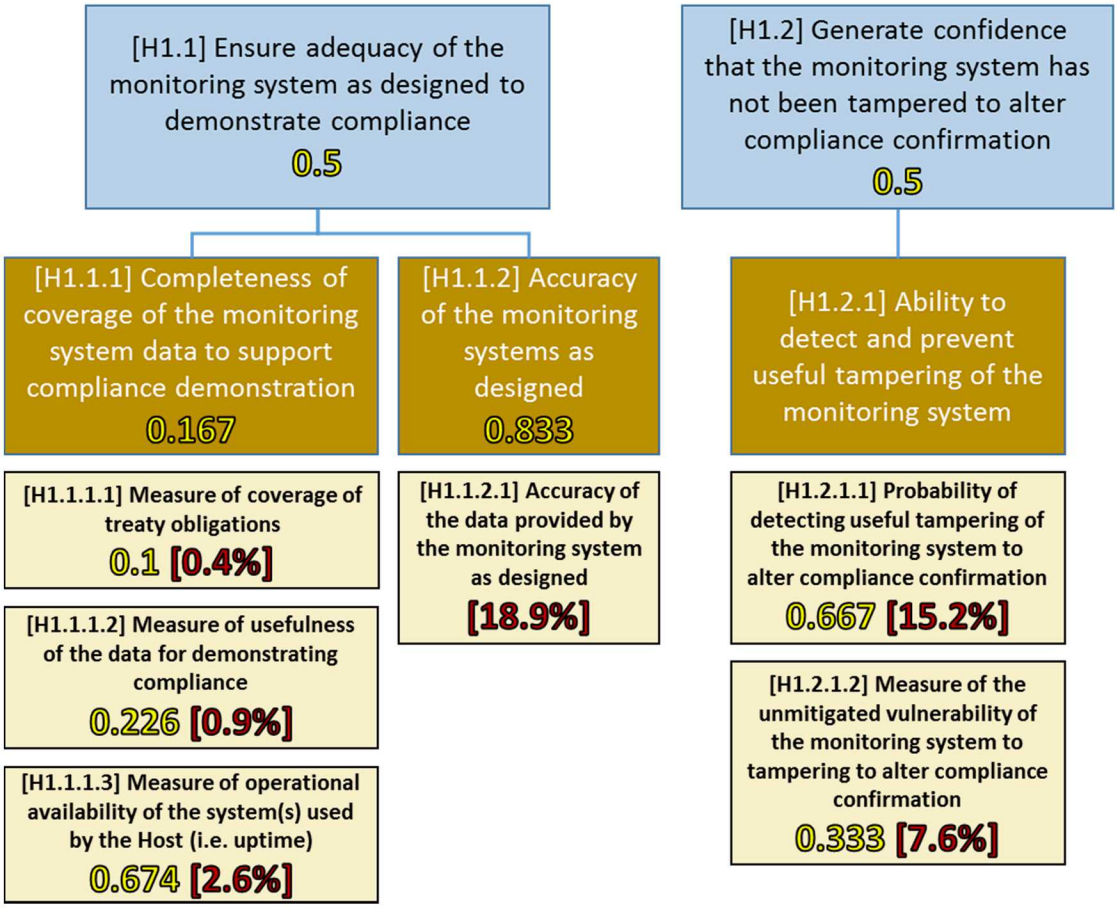

Figure 5. Host Criteria and Metrics under Demonstrating Compliance

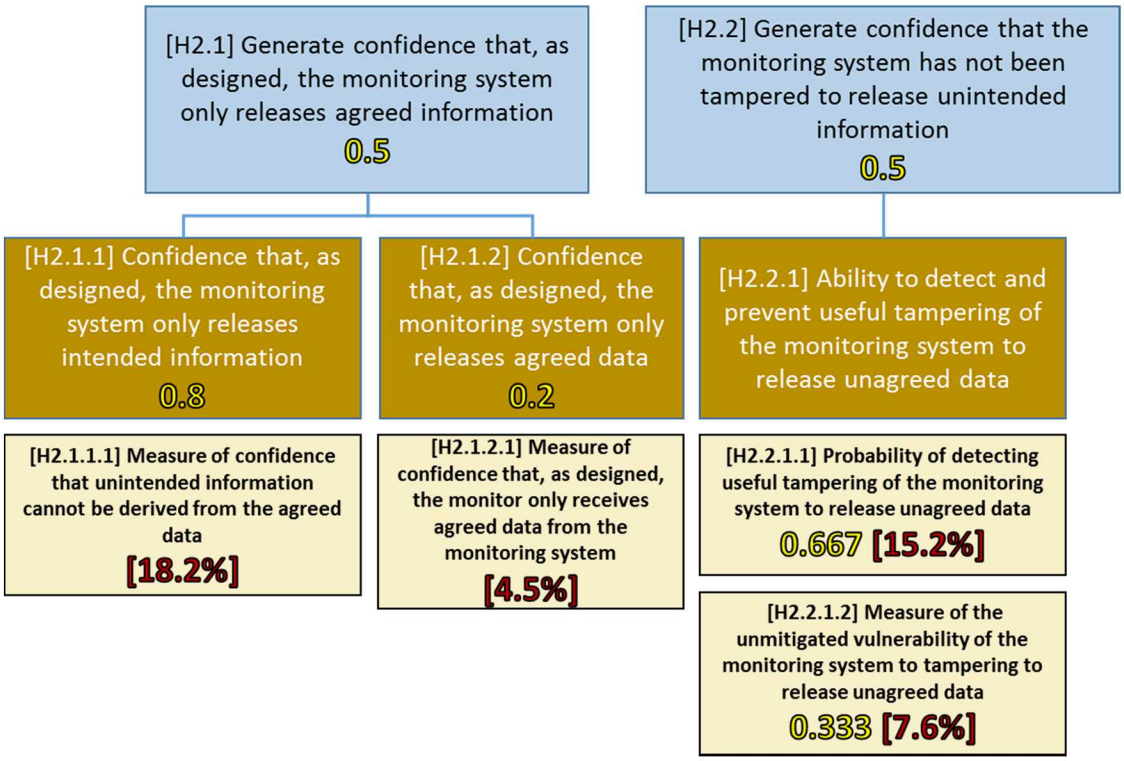

Figure 6. Host Criteria and Metrics under Protecting Non-Treaty Information

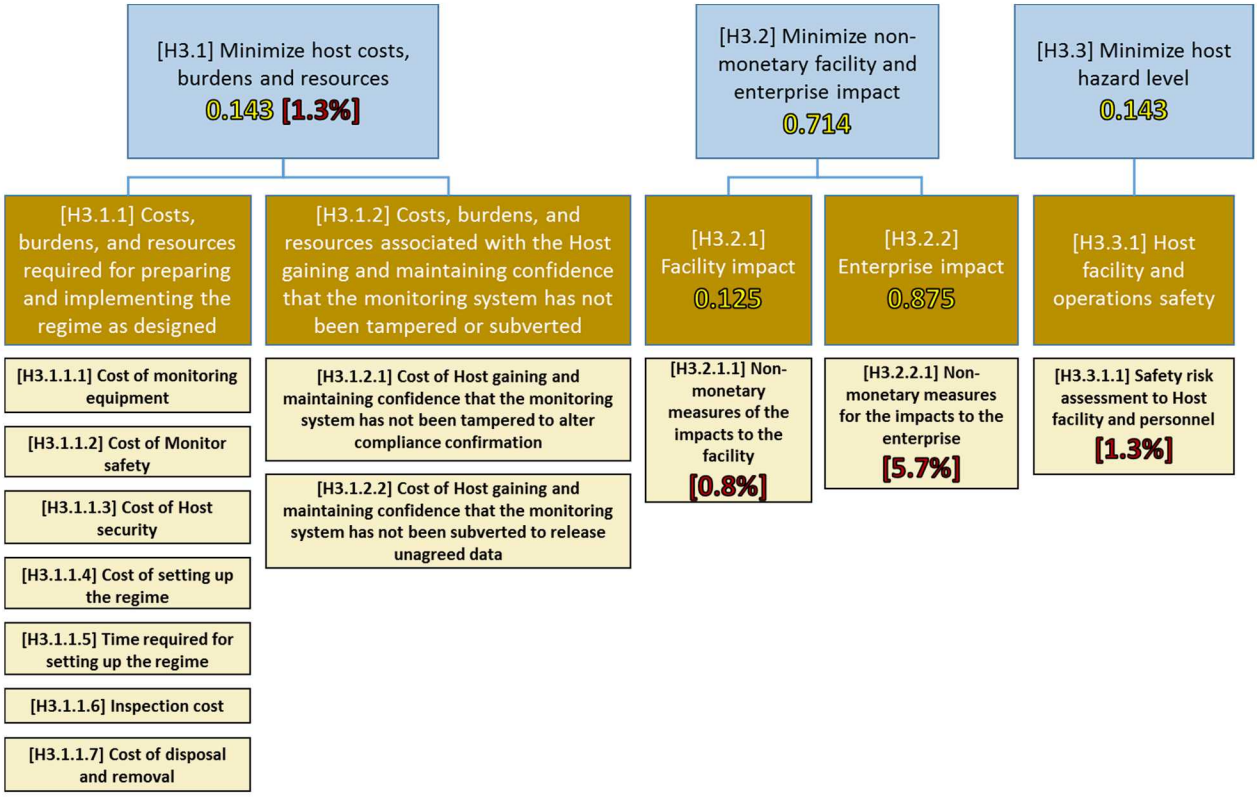

Figure 7. Host Criteria and Metrics under Practical Acceptability

A driving principle in this reexamination is the desire to end up with a set of criteria that are mutually exclusive and collectively exhaustive (MECE)[2]. Mutually exclusive criteria have no overlap and thus we are not double counting any portion of those criteria. Collectively exhaustive criteria span the complete set of considerations that the host and monitor care about with respect to a monitoring system. While it may not be possible to ensure perfectly MECE criteria, we are attempting to significantly improve those attributes of the criteria in this effort. Some of the criteria developed in the 2015 effort, specifically *certifiability* and *authenticability*, were considerably comingled. In the 2015 effort, the team began by first identifying high-level criteria that they deemed necessary for evaluating a monitoring regime; they then identified sub-criteria to capture the key elements that comprise the high-level criteria and that might be easier to assess for a particular monitoring system option than the more abstract high-level criteria. In this project, we took the opposite approach: building on experience from the 2015 effort and other projects, team members created lists of "considerations" which are relevant to the host and monitor in evaluating a monitoring regime. We used these lists of considerations to develop a complete list of potential criteria, eliminating duplicates and taking care to ensure all of these elements were mutually exclusive and defined at the same functional level. We also eliminated "solutions" from our list of potential criteria. Solutions represent ways to implement a monitoring regime rather than criteria to evaluate one and, therefore, should not be used as criteria. For example, when considering if a monitoring regime can be certified and at what cost, one solution is to use equipment that has been certified before. "Precedence of certification" is an option to address these interests; it is not a criterion.

Criteria Weighting

The utility of the criteria in evaluating multiple monitoring system options relies on a system for comparative evaluations. In this effort, we use the Analytic Hierarchy Process (AHP) to identify the relative importance of each criteria and metric (i.e., criteria weighting). It could also be used to compare monitoring system options after a weighted criteria hierarchy is established. The great benefit of AHP is that subjective input (as is critical in criteria weighting and often necessary in comparing monitoring system options given the metrics we have available) is still usable, and in fact can be used alongside objective input when it is available. AHP uses pairwise comparisons of a number of alternatives with subject matter expert (SME) judgements. For exhaustive pairwise comparisons, each alternative is compared against each other. For two alternatives, there is only one comparison, whereas for three alternatives, there are three comparisons, and for four alternatives, there are six comparisons, etc. To support criteria weighting, the group of alternatives considered at any one point in the process is dependent on the tree structure. As seen in Figure 1, the two monitor values are compared with a single pairwise comparison. Each value has two objectives, so that two more comparisons will identify the weightings off all elements in the hierarchy shown in that figure. In Figure 2, there are three metrics under the criteria [M1.1.1] Completeness of coverage of the

[2] Refer to the blog post "MECE Framework" by Tom Spencer at http://www.spencertom.com/2013/01/30/mece-framework/, or the book The McKinsey Way by Ethan Raisel, published in 1999.

monitoring system data to support confirmation of verification objectives, so that three pairwise comparisons are needed to fully weight those metrics.

Each pairwise comparison for criteria weighting uses a one through nine scale of relative importance, in which a value of one means that the criteria are equal in importance and the values indicate a larger different in importance as the number increases, as indicated in Figure 8.

The Fundamental Scale for Pairwise Comparisons		
Intensity of Importance	**Definition**	**Explanation**
1	Equal importance	Two elements contribute equally to the objective
3	Moderate importance	Experience and judgment moderately favor one element over another
5	Strong importance	Experience and judgment strongly favor one element over another
7	Very strong importance	One element is favored very strongly over another; its dominance is demonstrated in practice
9	Extreme importance	The evidence favoring one element over another is of the highest possible order of affirmation
Intensities of 2, 4, 6, and 8 can be used to express intermediate values. Intensities of 1.1, 1.2, 1.3, etc. can be used for elements that are very close in importance.		

Figure 8. Pairwise Comparison Scale

Criteria weighting should be done by policymakers or other decision-making stakeholders. To explore the concept of criteria weighting with AHP, the team (acting as a proxy for the monitor and host stakeholders) conducted the pairwise comparisons for the entire criteria hierarchy. The weightings of each grouping within the tree are indicated in yellow numbers in Figure 1 through Figure 7. That is, the yellow numbers indicate the portion of the next-higher element in the tree that that element is weighted. For each grouping, the weights should add to one. The leaves of the tree, typically metrics[3], also have a red number that corresponds to the global weighting of that element. A global weighting represents the importance of that element in the entire tree (monitor or host). All of the percentages in red sum to 100% for each tree.

When all of the pairwise comparisons have been conducted and there is a global weight for each metric, the entire hierarchy can be reduced to a list of metrics and global weightings. Then each monitoring system option can be measured by each metric, the options can be compared against

[3] The monetary costs, indicated by M2.1 Figure 3 in and H3.1 in Figure 7 have a single global weighting. This is due to the fact that all metrics under these objectives have units of dollars (or other currency), and as all dollars are worth the same as all other dollars, the total from each of these metrics can be summed before applying the criteria weighting. Therefore, the team did not find value in identifying the weighting farther down the tree for these objectives.

each other using pairwise comparisons, and the resulting percentages can be multiplied by the criteria weightings and summed in order to compare the options as a whole.

Conclusion

The approach presented here results in criteria that are appropriate, useful, mutually exclusive, and represent the distinct and, at times, conflicting interests of both the host and the monitor. The benefit and cost framework that serves as the foundation for the hierarchy of this framework supports the evaluation of both the effectiveness and efficiency of implementation options in meeting the host's and monitor's interests. Finally, the elimination of solutions from the criteria reduces bias and prejudice and ensures a fair evaluation of all implementation options.

Since many of the concerns of a monitoring system can only be measured subjectively at the present time, it is difficult to ensure mutual exclusivity of the criteria. The system performance is based not only on technical data but also on the perceptions of each party, which may differ from one set of individuals to another. The team therefore must assume the perspective of each party in the absence of an actual treaty negotiation with identified parties, and this will necessarily lead to some errors in judgement. In addition, there are likely to be disagreements about whether the criteria are mutually exclusive given their subjective nature. While we admit that this is not a perfect tool, we present it as useful for assisting policymakers in making decisions amongst monitoring system options.