THE VOLCKER ALLIANCE
Working for Effective Government

FIVE SUGGESTIONS TO AVOID MEASUREMENT MISMANAGEMENT

Thoughts on Using Metrics for Cybersecurity Centers for Academic Excellence

Remarks by Shelley H. Metzenbaum
President, The Volcker Alliance
The Second Annual Cybersecurity Summit
Presented by AFFIRM and USCC
Washington D.C.
March 26, 2015

I am clearly the odd man out on this panel and at this meeting today. Everyone else is an IT expert, a cyber expert, or both. I am neither.

What I am is someone who knows something about measurement and, more specifically, measurement in government. What I appreciate (and, frankly, can sometimes be evangelical about) is the enormous potential of performance and other measurement. Used wisely, performance measurement and complementary data are power tools for government. They are an incredible resource that helps organizations and the people who work in them accomplish more mission for the money, reduce risk, enhance people's experience with government, improve fairness, and strengthen accountability to the public.

Used inappropriately, however, measurement creates non-trivial problems.

That is why, when I received an email a few months back from the National Academy of Public Administration (NAPA) about an effort to identify useful metrics to help the Department of Homeland Security and National Security Agency assess whether cybersecurity education programs met a threshold level of quality, I got worried.

I shared that worry with NAPA, suggesting that it take care not to answer the wrong question, "What are the right metrics?" I worried that it would, indeed, try to answer this wrong question, and so should approach this study with great caution. Specifically, I suggested, the study should try to answer another question: how do we find schools ready to work with government and others to use performance metrics to find ways to improve cybersecurity education and to apply those findings? I guess I am here today talking to you and am sitting on the NAPA panel to make recommendations because the folks with whom I spoke found my worries valid.

Let me elaborate on them now, and suggest better ways to move forward to increase the quantity of high-quality cybersecurity employees supplied by cybersecurity educators. One way to think about these worries is five "don't's" of performance measurement – five things **not** to do. Then,

I will follow up with five specific suggestions. Perhaps think of these as five performance measurement "do's."

Five Worries

1. <u>Lack of clarity about the intended use of measurement</u>. The NAPA Panel was asked to identify metrics to "assess cybersecurity education programs." Frankly, the word "assess" always worries me. Is "assessment" the real purpose of metrics? Or do we want to use metrics for other purposes, such as informing and motivating improvement? It is important to be clear about why we want metrics and how we will use them. Let me elaborate a bit on possible uses of measurement for cybersecurity education, and suggest an order of priority among them.

- **Qualify eligibility.** My guess is that the Department of Homeland Security and the National Security Council want to use performance measures to qualify schools to be Centers for Academic Excellence (CAE), establishing performance thresholds schools need to exceed to qualify for designation as a CAE, presumably benefiting in some ways from that designation.
- **Find schools that produce, comparatively, the best students to tackle *general* government cybersecurity needs.** Perhaps some government agencies want to use the metrics not just to qualify schools as meeting minimum thresholds but also to find the schools that graduate the strongest people. (The ability to compare across graduates is clearly more important when the supply of people qualified for cybserscurity jobs exceeds the number available for hire.)
- **Find schools that produce, comparatively, the best students to tackle *specific* government needs.** Some agencies, I suspect, need employees with specific types of cybersecurity skills and would prefer to hire graduates with those skills than a generalist. The capacity to screen for specific skills, either to confirm a minimum threshold has been met or to determine comparative position, requires greater specificity in measurement.
- **Find schools that best meet a student's needs and preferences.** Perhaps you want metrics to inform a student's decisions about the best school to meet his or her needs. Students undoubtedly want some information similar to what government wants to know, but most likely also want some information of less interest to government, such as information about the cost of the education relative to expected return (except for employers assuming the cost of education), geographic location, other college offerings, and teaching philosophy.
- **Inform improvement.** Do you want metrics that help Centers of Academic Excellence find ways to improve? It is my hope that for everyone, the answer to this question is yes. I would argue, in fact, that informing improvement should be the priority use of measurement. Measurement needs to be collected, analyzed, used, and shared in ways that engage and inform key decision-makers in government and in the schools, including central office administrators and especially front-line teachers and department heads. Using measurement this way does not happen on its own, however. It requires intentional management, assignment of responsibilities, and adequate resourcing.
- **Motivate improvement.** Measurement can motivate in a number of ways. Most people instinctively like to do well so a well-designed measurement system motivate by providing fast feedback about how well things are going in a reasonably accurate,

contextualized manner. The motivational power of feedback can be magnified by linking to a few stretch targets; specific, ambitious goals that energize people in the delivery chain provided they are neither too numerous nor overly ambitious. Measurement also motivate some people through fair comparisons because some strive to lead the pack, some try to avoid the back of the pack, while others aim to lead their league and beat out those they consider peers. The possibility of reward or punishment linked to target attainment, relative position, or relative to past performance can also motivate, although, as I will discuss later, it can sometimes motivate dysfunctional rather than constructive responses.

- **All or some of the above and possibly other purposes.** Of course, it is possible that you want to use measurements for all of these, and possibly other, purposes I did not list. The question is: is it possible to use the metrics for all of these purposes simultaneously? Measurement's use for one purpose can drive out its value for other purposes unless these intended uses, and preferences among them, are carefully considered. I urge you, in thinking about measurement selection, to consider intended uses of the measurement and the order of preference for that intended use.

To simplify, let me suggest using measurement the following ways. In addition to qualifying participation, use measurement to illuminate, communicate, and motivate.

- **Illuminate.** Use measurement to illuminate what is working and worth promoting for adoption in other locations and adaptation to other situations; what is not working that needs fixing; root causes of problems; risks to manage and, most likely, reduce; and unwanted side effects. Use measurement, also, to illuminate the prevalence and characteristics of problems and opportunities needing attention and inform more precise treatment (e.g. teaching) design to address specific needs. When measurement fails to illuminate, I would argue it has little real value.
- **Communicate.** Use measurement to communicate what is important. This notion is captured in the well-worn phrase "what gets measured gets managed." Also, use measurement to communicate what has been and is being learned (e.g., which schools, which teachers, and which teaching methods produce more able cybersecurity employees.) In addition, use measurement to communicate what still needs to be learned, for example, spotlighting divergent trends and anomalies and inviting others to probe why they are happening or counting and characterizing problems to inform priority-setting and descriptive research.
- **Motivate.** Finally, use measurement to motivate. Measurement can motivate in a number of ways. Most people like to do better and measurement provides them feedback allowing them to do better both by showing them how well they are doing relative to their own expectations and, if peer measurement is available, by identifying those doing better whose practices might be worth adopting and adapting. Also, many people are motivated by healthy competition and not only want to learn from their better-performing peers, but do better than them. And, of course, the possibility of reward or the threat of punishment can also motivate, but, as I will discuss in a moment, not always as intended.

Starting with the motivational purpose of measurement (especially when linked to incentives) can easily impair measurement's capacity to illuminate, so focus first on finding measurements

that inform efforts to improve and then figure out how to use them to communicate and, finally, to motivate.

2. Inadequate communication of intended use of measurement. I have a second worry. I worry that organizations will fail to communicate, clearly and broadly, how they intend to use performance measurement, and, more specifically, fail to communicate that they intend to use measurement primarily to guide the path to improvement, and only after satisfying the "illumination" need, to motivate movement down the path to doing better. (I will suggest some ways to do that in a moment).

"Grading Teachers by the Test" is a very thoughtful piece that ran in the New York Times on March 25, 2015.[1] The author, Eduardo Porter, talks about Goodhart's Law, apparently referring to work by the British economist Charles Goodhart. Porter summarizes the law in a way I find appealing and constructively provocative: "A performance metric is only useful as a performance metric as long as it isn't used as a performance metric." That seems counter-intuitive, but, from my experience, he's got it right. People (and organizations) tend to run away from the insight-generating potential of measurement if they fear the same measurements could be used to punish them.

3. Failure to choose measures genuinely reflective of the value schools add. Unless well-tested metrics exist that are broadly accepted as accurate and useful, premature adoption of performance metrics can lock in counter-productive decisions and actions by government and schools. If measurements are going to be used for school selection, whether to determine that the schools meet minimum thresholds or to find the schools that best match government or student needs and preferences, certain conditions must be met. Do we know how to measure schools in a way that fairly reflects ways schools might add value – whether through the quality of what and how they teach or through how well they screen their applicants during the admissions process? If we don't, which I suspect is the case and the reason for the requested NAPA study, then we risk going down a very deep rabbit hole, at best, and more likely hard-wiring in some very bad practices that will be hard to undo.

4. Failure to use the right suite of measures. When choosing measurement, a suite of measures is often needed to support the kinds of uses discussed above. Problems arise, for example, when performance measures don't fully capture key attributes that matter. A classic piece of research documenting this problem was done about fifty years ago and remains as relevant today as then. Peter Blau (1963)[2] studied a government employment office that helped workers find jobs (and employers find workers) and saw how overemphasis on an easily tallied measure, the number of interviews conducted, encouraged government employees to arrange interviews but diverted them from doing the full array of tasks needed to help a client land a job. Not surprising, but not good. Counting job placements did a better job motivating government employees to tackle the full suite of tasks needed to help the client.

---

[1] Eduardo Porter, "Grading Teachers by the Test," Economic Scene, New York Times, March 24, 2015 (online version), http://www.nytimes.com/2015/03/25/business/economy/grading-teachers-by-the-test.html?r=0

[2] Peter M. Blau, The Dynamics of Bureaucracy: A Study of Interpersonal Relationships in Two Government Agencies (University of Chicago Press, 1963), p. 38.

Measuring the wrong thing or only a part of the picture, in this case the count of an activity believed to influence the outcome of interest rather than the outcome itself, tends to make a big mess of things and encourage unwise allocation of resources. The result is less progress on a program's actual objectives than should be achieved.

Care needs to be taken, of course, to select a reasonable number of measures so that the time needed to produce the measures does not overwhelm efforts to analyze the data to illuminate ways to improve and communicate priorities nor to qualify entry, inform choice, or motivate needed change.

5. Ill-structured incentives and wrong accountability expectations. If you want to use metrics in "high stakes" ways to screen out ineligible schools or to award funds in some form, it is important to have the right indicators as well as adequate funding to monitor and validate the accuracy of the measurements. Organizations and individuals must be cautious using measurement for school selection to make sure they do not motivate measurement manipulation rather than motivating actions that identify and produce better cyber security employees.

Think of the annual ratings done by US News and World Reports. School officials tell me they induce all sorts of perverse behaviors and measurement tricks (such as schools ginning up applicants to lower acceptance rates to earn a higher ranking) without improving either school quality or the match between students and schools.[3] Or consider the scheduling shenanigans at the Veterans Administration hospitals and the testing turbulence in the Atlanta schools. Where did the motivation to manipulate measurements rather than design a better scheduling system (so people who could afford to wait were given longer wait times so people needing quick attention could see a doctor sooner) or argue for more budget come from? [4] What prompted the good people who chose to become educators to choose instead to conspire on state standardized testing, according to a recent court decision? [5] I seriously doubt that these people were motivated by huge bonuses of the sort that tempted the Enron employees. What I do understand is that to unleash the full potential of measurement, leaders need to emphasize and re-emphasize through their words and actions their intent to use data to help people in the delivery chain – whether they work on the front-line, in the central office, or are grantees – find and apply ways to improve.

I am not saying that linking measurement to incentives is always a bad idea. It can, in fact, constructively redirect attention to problems needing attention, as incentives linked to the rate of hospital-acquired infections and patient return-to-hospital rates seem to be doing. When making these linkages, however, choosing measures that accurately and fairly capture performance is essential, as is a strong monitoring mechanism that assures accuracy and adequacy, especially

[3] John Tierney, "Your Annual Reminder to Ignore the *U.S. News & World Report* College Rankings," The Atlantic, September 10, 2013, http://www.theatlantic.com/education/archive/2013/09/your-annual-reminder-to-ignore-the-em-us-news-world-report-em-college-rankings/279103/

[4] "Statement of Richard J. Griffin, Acting Inspector General–Department of Veterans Affairs," US Senate, September 9, 2015, http://www.veterans.senate.gov/imo/media/doc/VA%20OIG%20Statement%20-%20SVAC%20-%20Sept%209%202014.pdf

[5] Associated Press, "Atlanta Educators Jailed in Test Cheating Scandal," New York Times, April 2, 2015, http://www.nytimes.com/aponline/2015/04/02/us/ap-us-atlanta-schools-cheating.html?_r=0

when self-reporting, or a robust third-party measurement system is involved. Surveying government cybersecurity supervisors about the quality of the students they hired, their areas of strength and weakness, and schools could be a good third-party measurement mechanism to capture key aspects of CAE performance.

Five Suggestions

I have shared with you my five worries. Now, let me offer five suggestions that address my worries.

1. Display and array measurement so they support intended information uses. Dashboards are all the rage these days, but lots of wasted effort goes into creating dashboards without much thought directed to who will use them, how they will be used, and whether they are likely to provoke the intended beneficial responses. Think carefully about intended uses of communicating information when displaying performance data and confirm that the shared information had the intended impact, and did not trigger unwanted, unhealthy reactions as so often happens with dashboards and other public reporting.

If using metrics as a tool for informing student choice or even government recruitment efforts, consider the approach used by *Consumer Reports* to display how products fare relative to different characteristics of the product. *Consumer Reports* shows which products fare best on each characteristics, which do best overall, and which do best for those looking for a "best buy" – value for the money, not just on performance. Government programs need a variety of different cyber skills likely to come from different programs and schools, so I am guessing it would be helpful to array performance information in a similar multi-dimensional way so both students and government know which schools do best on which types of skills and which do best overall. How you organize and share information will not only send a message about its intended use, but also affect whether that intent is achieved.

2. Join a Continuous Learning and Improvement Community. Require all CAEs to join a continuous learning and improvement community (CLIC). My mother always told me never to join a clique, but after looking at the decades-long work of National Highway Safety Administration to reduce traffic fatalities, more recent work with community colleges, as well as progress made in some areas of health care using iterative measured trials to test hypotheses and improve outcomes, I would argue that the federal government should create, support, and mandate participation in CLICs whenever awarding grants or other institutional benefits. Require CAEs to participate in a continuous learning and improvement community, agreeing to share data in real time, conduct appropriately rigorous measured trials to find better ways to teach, and continually brainstorm and analyze to find root causes of problems and discover ways to address them. Also, government (and others) should support skilled network coordinators to convene and support these CLICs, analyze the data, look for relevant research, formulate and test hypotheses, and promote adoption of better practices.

Some recent developments with community colleges may provide a possible model here. Anthony Bryk has a new book out, *Learning to Improve: How America's Schools Can Get Better at Getting Better*. He and his colleagues coined the term, "improvement science," I believe, as

well as “networked improvement communities.” I suspect these will be useful concepts to adopt and use with CAEs to improve cyber security outcomes. Here is a brief description of Bryk’s work from a blog he co-authored:

> This initiative addresses the problem of the extraordinarily high failure rates among the half-million community college students annually assigned to developmental (remedial) math instruction as a prerequisite to taking degree-level college courses. Traditionally, only about 20 percent of those enrolled ever make it through these courses -- a critical gatekeeper to opportunity. A network of faculty members, researchers, designers, students and content experts joined to create a new system built on the observation that "structured networks" accelerate improvement. They are a source of innovation, and of the social connections that facilitate testing and diffusion. They provide a safe environment for participants to analyze and compare results and to discover patterns in data. In addition, they involve the people on the ground in generating and analyzing the evidence that comes out of their daily work. Network participants identified six primary causes for high failure rates, and then tested improvement hypotheses. They used evidence "to get better at getting better," and thereby dramatically improved outcomes -- tripling the student success rate in half the time.

3. Copy Carl Wieman. I would also urge requiring CAEs to learn from Carl Wieman, a Stanford physics professor and Nobel Prize winner who did a stint at the White House Office of Science and Technology Policy a few years back. Wieman is doing rigorous research to identify better methods for teaching college-level science and engineering, using control trials in the classroom to validate carefully formulated hypotheses about better teaching approaches.

One of the beauties of Wieman’s work is that it requires, first, clarity of educational objectives. What do we want students to learn in operational terms; what will they actually be able to do? Federal agencies should make clear to the schools that, within a specific period of time, they will only work with schools that set clear, public learning objectives for each course and for each class, share indicators about changes in student knowledge with the federal government and other CAEs in a CLIC, and participate in appropriately rigorous control trials of the sort Carl Wieman is doing to find comparatively more effective educational approaches.

4. Look for other means to find good cybersecurity employees. I hope that, even as efforts are being made to find good metrics to qualify and improve school-based CAEs, a complementary question is being asked and answered, “Are there other ways to find people likely to be good at this work?” I have a cousin who graduated from Sarah Lawrence with a degree in philosophy who has worked in the cybersecurity business for at least 10 or 15 years. He is obviously good because every time he tries to leave, they woo him back. I suspect he is good at this business because he likes to think about how people think. I hope federal agencies (and others) are working to figure out the different brain types and majors that produce people with the full set of inclinations and instincts needed to do this work well. Perhaps they are in anthropology or sociology departments, or in physics and biology, as well as in computer science and engineering. I suspect it will be valuable to test individuals to screen for the way they think, in addition to developing metrics to help schools teach (and possibly screen for good candidates).

4. **Cross-agency internships and recruitment**. Are federal agencies working together to build successful internship and recruitment programs? Are you building cohort networks to support the new hires so they want to stay? If not, why not? The law allows it so isn't it time to make this happen? (As one of its first projects, The Volcker Alliance debunked the myth that agencies cannot hire each others' interns without re-competing the jobs.) It is time to start working across government to find the schools and types of student that prove most useful to agencies. This will necessitate the identification of leaders for a cross-agency working group and agency commitments to make this work happen, but the potential return on this collaborative effort is enormous. Isn't it time to make this happen?