Standardized Scoring for Security and Risk Metrics

With breaches and hacks in the news every day, information security is now firmly on the board's agenda. While certainly difficult to do, measuring security is fundamental to understanding it. Technology risk metrics monitor the accomplishment of goals and objectives by quantifying the implementation, efficiency and effectiveness of security controls; analyzing the adequacy of information security program activities; and identifying possible improvement actions.[1] Most security metrics programs are typically based on two assumptions: There is a secure way to manage any system, and the task of security management is to maintain that state.[2]

Measuring Security

The quantification of technology risk is an idea that continues to captivate. Parallels are drawn with credit and market risk, both of which allow currency-based means of risk quantification. There have been many attempts (including some that have been regulator prodded) where the concept of value-at-risk has been sought to be applied to operational risk, of which technology risk is a subset.

The measurement of information security risk is challenging. Realized outcomes for IT risk tend to be clustered toward the extremes, and the most likely outcome for a company is generally no losses, with a tiny probability of very high losses. Efforts at quantification have involved either black-box logic, such as modeling loss distributions based on extreme-value theory,[3] or the combination of various security metrics (often using weighted averages) as a composite metric. No approach has been successful enough to win any level of widespread adoption, and nearly all have failed to create a measure that correlates with and is predictive of realized operational losses. In fact, the Advanced Measurement Approach for operational risk,[4] which requires modeling operational risk using mathematical models akin to those used for market and credit risk, will soon be scrapped in favor of a simpler approach—an admission that operational risk, of which technology risk is a component, is structurally different from financial risk.

Most security metrics and risk quantification programs have, therefore, ended up focusing on building dashboards and scorecards that cast a wide net, mostly looking at control compliance. Technology risk reporting at most organizations almost always consists of tables of security metrics, often highlighted using a traffic-light convention.

Metrics relating to different information security areas use a diverse set of units of measure, and the numbers often need an interpretation unique to a given measure. For a senior executive who may not be well versed in the technical details of what each metric represents, the interpretation of how good or bad a number is can be a challenge. This article proposes an approach to assess and interpret security and risk metrics using standardized scores.

Interpreting Security Metrics

Security metrics for any corporation generally tend to be numerous, often numbering in the dozens, if not the hundreds. The sheer quantity of metrics often overwhelms the task of messaging. To confound matters, metrics come in different forms. Some metrics are absolute numbers, e.g., the number of vulnerabilities discovered in an application. Some metrics are averages, e.g., the mean time to repair. Others may be percentages (or ratios of some sort in a generalized form), e.g., the percentage of workstations not patched. Metrics may also be ranked statistics, such as league tables comparing divisions or regions.

Do you have something to say about this article?
Visit the *Journal* pages of the ISACA® website *(www.isaca.org/journal)*, find the article and click on the Comments link to share your thoughts.

Mukul Pareek, CISA, ACA, ACMA, PRM
Is a risk management professional based in New York, USA. He is copublisher of the Index of Cybersecurity *(www.cybersecurityindex.org)* and the author of the risk education website *www.riskprep.com*. He has more than 25 years of experience in audit, IT and information security and has been published on multiple topics relating to risk measurement in the *ISACA® Journal.*

The choice of whether a performance measure is expressed as a percentage or absolute number is generally based on an analyst's choice, driven by judgment and common sense in the context of the measurement being performed. This means there is a certain arbitrariness to how a measurement is expressed. To make a metric relative and allow for contextual interpretation, often a denominator is sought. This denominator is generally the total population to which a particular defect or attribute may apply. For some metrics, no practical denominators exist. For example, the number of security incidents may be best represented as an absolute number, for all possible denominators that can be imagined for this metric would dilute the message that the metric conveys.

> **"To make a metric relative and allow for contextual interpretation, often a denominator is sought."**

Standardized Metrics

When looking at metrics, a risk manager sees a wide range of numbers—some large, some small—and the numerical ranges vary. Interpreting and consuming such metrics can be a difficult task, particularly for someone who does not deal with them regularly.

Consider a representation where all metrics are stated on a common scale, e.g., 1 to 10, so adverse metrics that need attention would quickly stand out and those that are under control would be equally visible. This type of representation would make the task of risk communication significantly simpler.

Any security metric can be interpreted by decomposing it along three dimensions: the velocity, or the rate of change, of the metric toward (or away from) a desired secure state; the distance of the metric from such a secure state; and the persistence of the control failures counted by that metric (or the turnover of the insecure elements of the population explained by the metric). This article proposes a common numerical language for information security metrics in which security metrics of all types are expressed along a common scale, allowing for comparisons across controls and organizations over time. The article considers the practical aspects of such computations and the difficulties in interpretation and using metrics for decision support to deal with such synthetic derivations.

Desired Properties of Scaled Metrics

Scaled metrics convert security metrics to a bound range. For the purposes of this article, the desired properties of a scaled score include it being:

- Scaled in a defined range, e.g., 0-10
- Disaggregatable. It should be possible to identify, with precision, what each of its components is contributing to the score so decisions can be supported.
- Directionally consistent across measures, regardless of the original metrics. For example, a higher score should always be consistently good or consistently bad under the scheme.
- Similar to others, to enable aggregation using averages. In other words, it should be possible to combine scores to get higher-level scores, thus supporting a hierarchy of scores.

The rest of this article uses a hypothetical metric and the data reflected in **figure 1** showing the number of machines missing operating system

patches. It relies on the premise that the score is constructed in such a way that a larger score indicates lower risk and a lower score indicates higher risk or worse performance. This metric is based on a scale of 1 to 10, with 10 being a perfect score. A score closer to zero would indicate inadequate control performance or higher risk.

Figure 1—Number of Machines Missing OS Patches

Month	Raw Metric
January	534
February	257
March	337
April	436
May	278
June	420
July	60
August	321
September	331
October	260
November	318
December	189

Source: M. Pareek. Reprinted with permission.

Converting Metrics to a Score

Interpreting a metric, i.e., deciding whether the metric represents a good state or a bad state, generally requires the consideration of a number of factors based on the metric, the context, and the intuition and judgment of the risk analyst. Much of this human interpretation is actually quite straightforward. Metrics, whether expressed as a number or a percentage, require the following considerations:

1. What was the number in the periods prior, i.e., what is the rate of change in the metric compared to the past?

   The first consideration represents the rate of change, or the first derivative. A person comparing a point-in-time metric to its value at a previous time is thinking about whether the rate of change is too fast or too slow and if the change is in the right direction. The rate of change provides that information. Its sign, positive or negative, provides the direction.

2. How does the number compare to a threshold, or a desired good-state number? In other words, what would it take to cover the distance from the current state to the desired state?
   The second consideration is more complex and requires thinking about how the number compares to a threshold. In the hypothetical example shown in **figure 1**, the metric value for December is 189. If the desired threshold for this metric is 100 or lower, the distance of the metric from the desired threshold is an adverse variance of 89.

   Theoretically, even with all the possible data that could be identified, it would probably still be necessary to know how long it takes to remediate each of these exceptions. For example, if the average time required to fix each exception is one man-day, it could be said that there are, theoretically speaking, 89 man-days of work needed to get to the desired good state. This could then be considered in the scoring of the metric as the distance-to-controlled state (similar to the concept of distance-to-default used for credit risk). But such data are difficult to come by and are subject to individual perspectives and debate. If credible time-to-repair data are available, they could be used in a fairly straightforward way, but for the moment, this line of thinking will not be pursued.

3. What is the extent of persistence over time in the unfavorable elements represented by the metric?

   This represents the extent of churn or turnover in the constituents of the metric. Persistence relates to the aging of the security attribute measured by the metric. When 189 machines are reported as missing patches in December, it is probably also useful to know if these were the same machines

missing patches in November or earlier in the year or if they represent new machines that only recently went out of compliance.

The following describes the mechanics of how each of the previous considerations can be computed in a practical way.

Velocity Measurement

The velocity should be calculated as the rate of change, i.e., the first derivative, with reference to the previous measurement period.

$$Velocity\ Measure = \frac{Prior\ Period\ Metric\ Value\ -\ Current\ Period\ Metric\ Value}{Prior\ Period\ Metric\ Value}$$

This formula is simpler than it sounds, and here is an illustration: If the metric measurement for November is 318 and the December number is 189, the rate of change is equal to (318-189)/318, which equals 0.406. There is no theoretical upper or lower limit to the result from this calculation. For example, if the November metric was 2, then the rate of change would be 93.5.

> **"If two identical organizations have an identical measure for a metric, it may not mean that the state of their controls is identical, too."**

This calculation may need an adjustment for directionality to align it with the initial premise—a higher score represents a good state, and a lower score represents a bad state. In this situation, a decrease, such as the one seen from November to December, reflects an improvement. Consistent with that, the trend measure is positive. A metric in which a high number represents a better state can be accounted for by multiplying the result by -1. An example would be a metric that measures the number of applications or infrastructure elements that have been successfully tested for disaster recovery. In such a case, the metric would ideally be higher, not lower. A direction adjustment would be necessary in such a case, which would require the result to be multiplied by -1.

But coming back to the hypothetical example of machines missing patches, the computed measure for the trend for this metric is 0.41, as a positive number represents a favorable change. If there is no change, the trend measure will compute to zero.

Distance Measurement

The distance measurement is also a straightforward calculation; if the threshold for the metric is, say, 100, then the distance is calculated as follows:

$$Distance\ Measure = \frac{Threshold - Current\ Measured\ Value}{Threshold}$$

This calculation has the property that it provides a negative number if the threshold is exceeded. It is a measure of distance from threshold as it expresses the current value as a multiple of the desired threshold. A positive number of, say, 0.40 would mean the organization is 40 percent away from the threshold value being breached. When exactly at the threshold, the value is 0, i.e., this measure is centered at zero. Anything above zero is a good thing, and anything below zero is not good.

For the month of December in the hypothetical metric, the value of the distance measure is -0.89.

Again, as before, if the metric is such that a larger number represents an adverse state, it can be multiplied by -1 to adjust for directionality.

Persistence Measurement

The persistence element considers the aging of the items included in a metric and the length of time each of the constituent control failures have been open.

Figure 2—Aging of Machines Missing Patches				
0-30 days	31-60 days	61-90 days	>90 days	Total
65	42	35	47	189
34%	22%	19%	25%	100%

Source: M. Pareek. Reprinted with permission.

If two identical organizations have an identical measure for a metric, it may not mean that the state of their controls is identical, too. Continuing the example of machines missing patches, if the hypothetical organization has 189 machines missing patches in December, but these are the same machines that were missing patches six months ago, it indicates that the business-as-usual process to remediate patches is not working effectively. But if these 189 machines are all machines that went out of compliance in the month prior and all other machines that were missing patches at the end of the last measurement period are now compliant, it represents a completely different state of control compliance.

The persistence measure seeks to quantify that scenario. One way to do that is to look at the aging of the constituents of the metric. An example aging profile of the 189 machines missing patches in December is shown in **figure 2**.

If the expectation or the service level agreement (SLA) for addressing missing patches is 30 days, it would mean that about two-thirds of the machines were lagging behind. In many cases, security metrics will report only what is beyond the SLA. But that does not change the essence of what needs to be measured, which is the shape of the aging distribution. The more statistically minded may actually choose to measure skewness, though for many, simplicity trumps mathematical elegance, and knowing the percentage of the metric's aging that is below a desired level is sufficient.

For the purposes of this article, 34 percent will be the simple representation of the persistence measure. Alternatively, the proportion for more than 90 days could have been used as the measure if it were more relevant. In that case, since a higher number represents a worse situation, the number could be adjusted by subtracting that proportion from 1. In other words, 1 – 25 percent = 75 percent could be used as the measure of persistence.

Persistence measure
= Proportion of the metric constituents aged less than a threshold (30 days in the example) [or 1 – proportion aged greater than a threshold]

Combine Velocity, Distance and Persistence Into an Interim Score

Since the metric should be represented as a single number, it is necessary to combine the previously mentioned three calculations into a single number. To keep things simple, it is recommended to use a simple average. Depending on what is more important to an organization, weighted average scores could be used as well, applying a different weight to velocity, distance and persistence.

For the month of December, the scoring calculations would work as follows:

- **Velocity, or trend measure**—(November value – December value)/November value = (318 – 189)/318 = +0.41
- **Distance measure**—(Threshold – December value)/December value = (100 – 189)/189 = -0.89
- **Persistence measure**—Proportion under 30 days = 65/189 = 0.34

The average of the three measures is -0.0468, but this is not the final score that meets the criteria established earlier. There is one more transformation to complete.

Converting the Interim Calculations to an Absolute Score

Now that the interim score has been calculated, the scaled score can be calculated. Such a conversion can be performed using a mathematical function

that would take these numbers as an input and provide an output that varies between a certain range. There are a number of mathematical functions that can do this. For example, a score can be calculated as $\chi/(\chi + 1)$, where χ is the raw number that needs to be converted to a range-bound score.

The remaining part of this article uses the logistic function[5] (also called the inverse logit function) to convert these measures to a number that varies between 0 and 10. The logistic function has the property that for a given input, it provides a result that varies between 0 and 1 and is very linear for a range around 0, except around the extremes where it gets close to 0 or 1. Once a scored number between 0 and 1 is found, it can be scaled to a range, e.g., 0-10, by multiplying the result by 10.

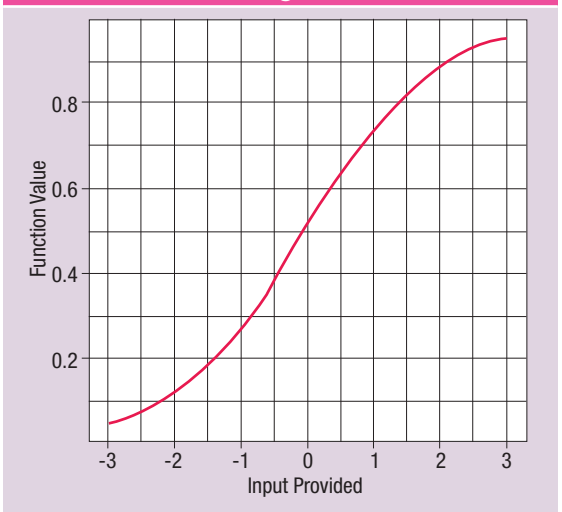
$$\textit{Logistic function-based score} = 10 * \frac{\text{Exp}(\textit{average of velocity, distance and persistence scores})}{\text{Exp}(\textit{average of velocity, distance and persistence scores}) + 1}$$

Figures 3 and **4** show the behavior of the logistic function. The function is near linear for small numbers and gets close to a maximum or minimum value fairly quickly as the departure from 0 becomes large. This is desirable for security metrics, so if a metric depicts a very unfavorable or a very desirable situation, it immediately stands out.

Since the logistic function results in a number between 0 and 1 and because the number needs to be between 0 and 10, it should be scaled by multiplying the score by 10.

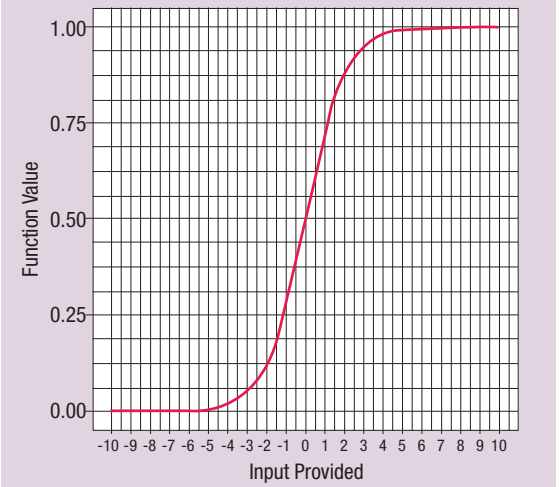
Even then, there is a remaining problem. A score of 0 returns a logit score of 0.5 or, per the scaled measure, a score of 5 on a scale of 0 to 10. But an interim score of 0 is a good score, i.e., it means the metric is on target. Therefore, representing it as a 5 on a scale of 0 to 10 is misleading, and it should be closer to 10 according to the intended scheme.

Figure 3—Inverse Logit Function of the Range -3 to +3



Source: M. Pareek. Reprinted with permission.

Figure 4—Inverse Logit Function of the Range -10 to +10



Source: M. Pareek. Reprinted with permission.

It is, therefore, necessary to add a constant that biases the determination of a scaled score closer

to 10 for an interim score of 0. Through trial and error, a reasonable correction can be provided if a constant of 1 is added to the score before the logit score is computed. This constant can be varied according to the needs of the organization and is akin to adjusting a weighing scale to 0. Therefore, the actual computation of the logit scores becomes:

$$\text{Logistic function-based score} = 10 * \frac{\text{Exp}(\text{average of velocity, distance and persistence scores})}{\text{Exp}(\text{average of velocity, distance and persistence scores}) + 1}$$

This article uses a constant of 1. Using the hypothetical example described earlier, **figure 5** shows the score calculation for different months.

As is shown in **figure 5**, this scoring mechanism corresponds to the properties described as desirable previously. In the month of July, which has a good trend (the metric falls from 420 to 60) and a good absolute number (60 machines, lower than the trend seen in earlier months), there is a high standardized score (8.2). August shows a decline in the standardized score to 2.6, as is to be expected given the five-fold increase in the number of noncompliant machines compared to the last month.

Using Z-scores as the Interim Calculation Mechanism

The approach described previously allows for a consideration of the various factors that go into interpreting metrics. The previous approach assumes there are thresholds available for all metrics, which is easier said than done. In situations where thresholds have not been established, an alternative and simpler approach that relies on z-scores can be adopted. This approach, while not as sensitive and precise as the one described

Figure 5—Scaled Metric Calculation Based on Trend and Distance

Month	Raw Metric	Threshold	Number of machines within 0-30 days	Step 1			Step 2	Step 3
				Velocity, or trend measure (previous month – current month)/previous month	Distance measure (previous month – current month)/ previous month	Persistence measure (number 0-30 days, divided by raw metric)	Equi-weighted average of trend and distance from Step 1, plus a constant = 1	exp (interim score)/(exp(interim score) +1)
Jan	534	100	166	NA	NA	31%	**Interim score**	**Final score**
Feb	257	100	134	0.52	-1.57	52%	0.82	**6.95**
Mar	337	100	84	-0.31	-2.37	25%	0.19	**5.47**
Apr	436	100	113	-0.29	-3.36	26%	-0.13	**4.67**
May	278	100	100	0.36	-1.78	36%	0.65	**6.56**
Jun	420	100	244	-0.51	-3.2	58%	-0.04	**4.89**
Jul	60	100	20	0.86	0.4	33%	1.53	**8.22**
Aug	321	100	144	-4.35	-2.21	45%	-1.04	**2.62**
Sep	331	100	139	-0.03	-2.31	42%	0.36	**5.89**
Oct	260	100	57	0.21	-1.6	22%	0.61	**6.48**
Nov	318	100	191	-0.22	-2.18	60%	0.40	**5.98**
Dec	189	100	65	0.41	-0.89	34%	0.95	**7.22**

Source: M. Pareek. Reprinted with permission.

previously, can provide results that come fairly close and are highly correlated to those obtained by the more elaborate process described earlier.

Z-scores are standardized scores calculated as the distance from the mean, expressed as a multiple of standard deviation. Standardized scores are based on multiples of standard deviation, an approach not too different from that used in financial risk where value-at-risk is a multiple of standard deviation.

$$Z\text{-score} = \frac{Metric\ Value - Mean}{Standard\ Deviation\ of\ Metric\ Value}$$

Normalized z-scores offer a number of advantages. They are easily computed, are foundational for a number of statistical techniques and are easily explained. Though they have no theoretical maximum or minimum, they are more likely to be small numbers than large numbers. Chebyshev's Rule (which states that that no more than $1/k^2$ of a distribution's values are more than k standard deviations away from the mean[6]) makes it difficult for the probability of a single observation to be too many standard deviations away from the mean.

The values of the z-scores for the example used earlier are shown in **figure 6**.

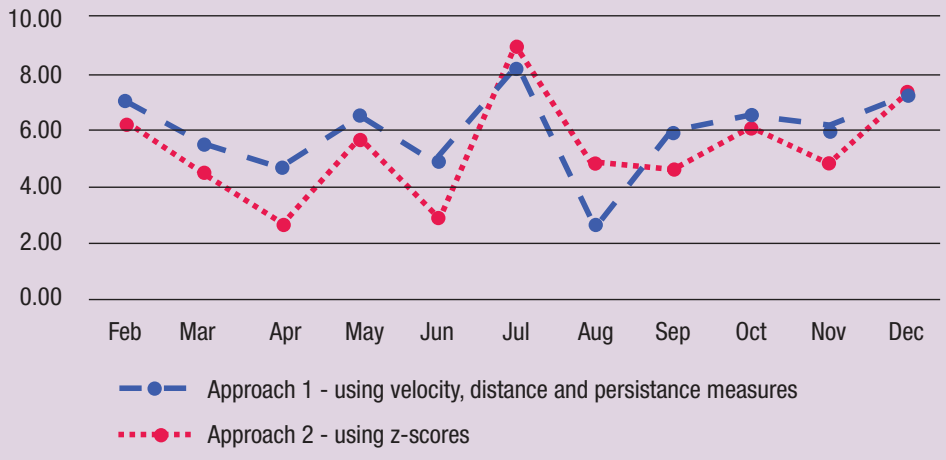
Figure 6—Scaled Metric Calculation Based on Z-score

Month	Raw Metric	Z-score -1 * (Metric value – Mean)/ Standard Deviation	Final Score = 10* Exp(Z-score)/ (Exp(Z-score) + 1)
January	534	-1.8	**1.4**
February	257	0.4	**6.1**
March	337	-0.2	**4.5**
April	436	-1.0	**2.6**
May	278	0.3	**5.7**
June	420	-0.9	**2.9**
July	60	2.1	**8.9**
August	321	-0.1	**4.8**
September	331	-0.2	**4.6**
October	260	0.4	**6.0**
November	318	-0.1	**4.9**
December	189	1.0	**7.3**
Mean	311.75		
Standard deviation	121.7		

Source: M. Pareek. Reprinted with permission.

Figure 7 shows a comparison of the two approaches: the standardized scores calculated according to the

Figure 7—Scaled Metric Values for Missing Patches Using Both Approaches



Source: M. Pareek. Reprinted with permission.

first, more sophisticated, approach and the second, more coarse, approach using z-scores.

As shown in **figure 7**, the scores calculated using the two approaches are quite similar. For the statistically minded, the correlation between the two for the hypothetical data set was 0.74. The second approach could be a cost-effective way to begin exploring standardized scores with only time-series data for a metric.

Conclusion

The standardizing scoring approach for security and risk metrics allows the risk manager to state a wide range of metrics in terms that use the same unit of measure, all owing for a comparison of items across time and control areas. While these calculation methods can be useful, they can contain limitations as well, and those constraints should be clearly understood. Composite scores are useful to highlight variations in a controlled process, but are not useful at more microscopic levels. Unless explained well, the logic behind such computations can come to be regarded as a black box, which can limit their adoption.

Endnotes

1 Chew, E.; M. Swanson; K. Stine; N. Bartol; A. Brown; W. Robinson; *Performance Measurement Guide for Information Security*, NIST Special Publication 800-55, USA, July 2008, *http://csrc.nist.gov/publications/nistpubs/800-55-Rev1/SP800-55-rev1.pdf*

2 Bayuk, J.; "Security as a Theoretical Attribute Construct," *Computers & Security*, vol. 37, September 2014, p. 155-175, *http://dx.doi.org/10.1016/j.cose.2013.03.006*

3 The Professional Risk Managers' International Association, *The Professional Risk Managers' Handbook*, PRMIA, USA, 2011

4 Basel Committee on Banking Supervision, *Operational Risk—Supervisory Guidelines for the Advanced Measurement Approaches*, Bank for International Settlements, June 2011, *www.bis.org/publ/bcbs196.pdf*

5 James, G.; D. Witten; T. Hastie; R. Tibshirani; *An Introduction to Statistical Learning*, Springer, USA, 2013

6 McClave, J.; P. Benson; T. Sincich; *Statistics for Business and Economics*, Prentice Hall, USA, 2000