Methodological Foundations:

Enabling the Next Generation of Security

The promise of breakthroughs in computer security—experimental test beds, insider-detection advancements, biometrics, and user interfaces that are robust to human error—will remain empty as long as methodological details trail the hype.

Roy A. Maxion and Rachel R.M. Roberts
Carnegie Mellon University

Methodology is like the weather; everybody talks about it, but nobody does anything. Most people seem to think that effecting proper methodology is either not important or not possible. Au contraire: discerning the causes of security threats, and forecasting the economic consequences of implementing bug fixes, require strict attention to experimental detail. Failure to do so can push products down the wrong path and subsequently bankrupt a company; it can leave consumers helpless against identity and data theft, system unavailability, and more.

People often make security decisions based on results that they read in journals and conference proceedings. Their understanding of these results depends primarily on a clear exposition of the method by which experimenters did their work, typically reported in the "experimental method" section of a research paper or presentation.

Comprehensive and lucid methodological explications have at least two major benefits: they facilitate replication of reported work, either conceptually or tangibly, and they help avert experimental errors—those of experimenters as well as those of readers. Although actual replication is not common in computer science, readers nearly always replicate experiments conceptually; that is, they imagine how an experiment *might* have been done. To do this with any reasonable degree of accuracy, the description of an experimental method must provide sufficient information to prevent readers from making erroneous assumptions. In addition to helping readers, a thorough methodological description also aids the experimenter in avoiding errors, simply through the act of writing down the procedure.

A few selected issues—operational definitions, reliability, internal validity, and external validity—limited by the small space available here, serve to illustrate common sources of experimental error that may be brought to one's attention by a written methodology. The first two issues regard valid measurement, and the second two regard valid experimentation. Together, they exemplify the importance of strong methodology in pushing the envelope of security.

Operational definitions

Dictionary definitions usually attempt to impart a conceptual understanding; an *operational* definition additionally includes an exact description of how to obtain an objective value for the measured characteristic. This is necessary to reduce ambiguity—to guard against the possibility that different experimenters will interpret and measure objects in different ways. Operational definitions ensure that everyone measures the same phenomenon in the same way; they prevent people from misunderstanding, for example, what threats were actually countered, and what results were actually achieved.

Example

Unauthorized users with access to a legitimate password can be detected when their keystroke-level timing rhythms are anomalous (with respect to normal). To build a system able to recognize temporally atypical typing patterns, researchers must decide what constitutes a timing anomaly. If the definition is absent or unclear, there is no way for other experimenters to replicate or address shortcomings in work done by others. Lack of clarity, or undue secrecy, impedes the development and comparison of diverse approaches to mitigating the problem, in this case, of stolen passwords.

Addressing the issue

There is no secret weapon for devising operational definitions for measures. However, the following steps should prove helpful:

1. Identify, or conceptually define, the characteristic of interest.
2. Select a measuring device, such as a tool or expert human judgment.
3. Describe the measurement procedure (or decision process) clearly, so that variability among repeated measurements is minimized, particularly when carried out by different people. If the measure consists of deciding whether or not an event occurs (detection of signal) or in what form (category of membership), establish a criterion by which to make the decision that puts the event into one class or another.
4. Test the definition (preferably among several peers) to ensure that it accurately reflects the concept being considered, and that it will convince potential critics.

Following the steps above, we could operationalize a keystroke-timing anomaly in the following way:

1. Conceptually define keystroke-timing anomalies as temporal latencies (time intervals between successive key presses) that happen only rarely.
2. Select a particular anomaly detector as a measuring device that is sensitive specifically to rare events. Simple mean and standard deviation might suffice for this example.
3. Control measurement variability by keeping detector parameters constant and background conditions (such as type of keyboard and system) consistent; establish a threshold for deciding that latencies are anomalous if they are more than, say, three standard deviations away from the mean.
4. Ensure that each timing anomaly detected in this manner is in fact a rare event, by appraising the outcome based on the decision threshold, and by soliciting critical input from peers about whether the definition has captured the concept. Setting the threshold will require consideration of the costs of failing to catch an intruder, and of keeping out authorized users who typed sloppily on one or more instances.

Unless we construct a description free of uncertainty in its interpretation, what is anomalous to one person may not be anomalous to someone else. Lack of ambiguity is the essence of an operational definition.

Good definitions are difficult to come by, for constructs as well as for specific measures. One group that has labored hard over precise definitions is the collaborative team comprising the IEEE Technical Committee on Fault-Tolerant Computing and the IFIP 10.4 Working Group on Dependable Computing and Fault Tolerance (www.dependability.org). Their 20-year effort has culminated in a recent article on concepts and definitions—including many that are security relevant—that sets a fine example (see A. Avižienis et al. in the "For further reading" sidebar).

Reliability

In the context of experimentation, a reliable measure is one that is stable; that is, if we repeatedly measure the same object or event, we will consistently obtain the same value (within measurement error). Reliability, manifested as low variability, helps us detect significant differences between theoretically distinct entities when they exist; it also makes running experiments less costly by requiring less data to be collected while maintaining the same statistical power.

Example

Researchers conducted an experiment to compare the timing performance of two distributed database transaction systems over a medium-sized local area network. Although the investigators calibrated the transactions carefully, repeated measures varied by an amount substantially greater than could be accounted for by any error in the measurement procedure. The measurements were not reliable enough to capture distinct transaction-timing patterns, hence preventing fruitful comparison between one system and another. Researchers later determined that the wide variation in results was due to time-of-day effects. Because they had taken measurements at different times of the day, during which the network's background traffic varied significantly, the uncontrolled variation in the background traffic made the transaction timings appear inconsistent. When the researchers repeated the experiment during the same hour of the day, the outcomes were as consistent as anticipated, making them sufficiently reliable for analysis purposes.

Addressing the issue

The best way to avoid unreliability is to control for as much unwanted variation (deviation from typical values) as possible. We can do this by correctly using the appropriate measuring instruments and procedures. We can also use explicit checks to ensure adequate reliability. One such check is called test-retest, in which we run the same test at two different times for the same phenomenon. In

For further reading

A. Avižienis et al., "Basic Concepts and Taxonomy of Dependable and Secure Computing," *IEEE Trans. Dependable and Secure Computing*, vol. 1, no. 1, 2004, pp. 11–33.

F.C. Dane, *Research Methods*, Brooks/Cole Publishing, 1990.

G. McPherson, *Statistics in Scientific Investigation*, Springer-Verlag, 1990.

D. Montgomery, *Design and Analysis of Experiments*, 5th ed., John Wiley & Sons, 2000.

the transaction-timing example, researchers could have done the test-retest check at the same time on different weekdays, with the expectation that traffic at a given time of day would be largely consistent from one weekday to the next (controlling, of course, for end-of-semester or holiday variations). Another reliability check is a parallel-form test, in which two closely related variants of a measurement procedure are used, and a consistency score is calculated to show correlation between the two; if the correlation is low, reliability is suspect. A final check lies in a form of inter-rater reliability, in which two experimenters conduct the same experiment with the expectation that at least 90 percent of their observations will agree. We can find other checks in the literature.

Poor experimental methodology can lead to critical mistakes.

Next, we move our discussion from the level of pure measurement to explanatory studies, which seek to discern the nature and extent of a potential relationship between two or more variables of interest.

Internal validity

In general, a *valid* experiment is logically well-grounded, and relevant to the purpose at hand. *Internal validity* concerns the extent to which the experimental outcomes are influenced only by manipulations in the experiment, and not by unanticipated or covariant factors. Its main requirement is the ruling out of plausible alternatives for explaining a given outcome. Internal validity is essential for knowing when the true cause of a problem has been isolated, or when the best solution for a security breach has been identified.

Example

An experiment in user identification attempted to discriminate among users on the basis of their biometric, idiosyncratic mouse activities (clicks, movements, scrolls, and so on). Participants browsed self-selected Web pages while instrumented software recorded their mouse activities into log files. Unfortunately, the variable of interest (mouse activity) was unexpectedly confounded with an extraneous variable (Web content). Because the particular Web content might have influenced a user's mouse behavior, researchers could not discriminate among users, solely and definitively, with respect to individual mousing style. Any apparent user-identification success might have been due either to user-specific mouse activities, or to differences among the Web pages that participants browsed. The experimental design did not provide enough control over influential factors, such as Web content, to determine whether or not the mouse-activity biometric would be successful (on its own) at uniquely identifying users.

Addressing the issue

Whenever there is more than one explanation for an observed experimental outcome, internal validity is jeopardized. Unwanted or unaccounted-for influences on the outcome of interest lead to ambiguity in experimental results. To avoid this, it is vital to anticipate and control as many sources of variation in (determinants of) the outcome measure as possible. Researchers must explicitly manipulate the most informative variables, as well as record the values these take, during the course of an experiment. In some cases, it is also wise to control ancillary factors, even though they may not be of primary interest. For instance, in a detailed experiment to test the mouse-activity biometric under different conditions, researchers could systematically vary Web content to see whether they could distinguish users under certain browsing conditions but not others.

After the major input variables have been accounted for, the researcher must control all remaining influences on the outcome by making background conditions uniform and/or by using randomization to neutralize residual factors. In this example, the experimenters could have made the Web pages the same for all participants, thus controlling for the influence of different Web pages that might have required a particular kind of mouse activity. Alternatively, the researchers could have employed a common mouse-based application and example task instead of relying on user-selected Web browsing. Keeping all factors constant except for one (here, the person) is the simplest approach.

A "controlled and randomized" experimental design can be very useful in showing internal validity. Control—holding all variables constant except one or more expressly manipulated variables—seeks to isolate the effect of the input variable(s) on the outcome. Randomization—ensuring equal chances that an experimental subject or specimen is assigned to one (or more) of the sets of influential conditions under study—seeks to neutralize potential sources of bias, or systematic errors. As a consequence of randomization, any "hidden" influences will tend to cancel out on the whole; some will affect the outcome by amplifying it, whereas others will affect the outcome by damping it. Using randomization, along with proper control, leads to solid evidence supporting the existence of *specific causal* relationships between variables of interest.

External validity

External validity refers to the ability of a study's results to generalize to settings beyond the circumstances of a specific experiment. Situations in which the characteristics of a study may (or may

not) match the characteristics of an applied setting include measurement procedure, background environment, application of the manipulated conditions, and the type of individual units studied (systems, people, widgets, and so on). Without external validity, there is little realistic hope of accurately assessing security levels, or of resolving security concerns that require measurement and experimentation.

Example

Out of convenience, designers may be tempted to sample host or network data in an ad hoc manner (or create it artificially) for use in evaluating intrusion detection systems. However, a biased sample that contains unrepresentative background traffic, artifacts due to traffic simulators or to injected malicious events, wrong base rates, and so on will make it difficult to reach valid conclusions about a detector's effectiveness in the wild. Researchers should carefully compose test data that reflect a realistic operational profile.

Addressing the issue

We can reduce or eliminate test-data bias through random sampling, although in many security-related contexts, the quantity and complexity of data may inhibit this. A general rule of thumb is to characterize the target domain, such as traffic on a particular network, over a time period substantially longer than that of the anticipated sample. Then the researcher must ensure that the characteristics of the data sampled over a shorter period still match the characteristics of the longer sample. Sometimes bias is unavoidable; in such cases, it may be best to make a clear statement about what biases *are* contained in the data, so that experimenters and practitioners can compensate for them or simply avoid them. For example, if intrusion base rates are not representative of real traffic, technicians can later tune their detectors to compensate for differences between the test data and operational conditions.

Although we illustrated only the hazard of unrepresentative samples, any dissimilarity between experimental conditions and reality—such as measurement procedure, manipulation of input variables, or background conditions—can weaken external validity. As a rule of thumb, it is best to compare only "like with like," or those scenarios that share commonalities at every key juncture.

We can glean further information about valid and reliable measurement and experimental design from the statistics and experimental methods literature.

With the proliferation of problems such as identity theft and cyberterrorism, stakes in security research are high. We can counter these threats by adopting a culture that continually tries to match the best practices and state of the art in experimental methodology so that we can run valid and cost-effective security experiments. This enables researchers and companies to create next-generation products, with the assurance that they will be highly effective in operation. ▢

Roy A. Maxion *is a principal systems scientist in the Computer Science Department at Carnegie Mellon University, where he is also the director of the CMU Dependable Systems Laboratory. His recent work has included anomaly-based intrusion detection and masquerader/insider detection. His research interests include broad aspects of dependability, security, user interfaces, quantitative experimental design, and autonomic systems. He has a PhD from the University of Colorado. Contact him at maxion@cs.cmu.edu.*

Rachel R.M. Roberts *is a research associate in the Computer Science Department's Dependable Systems Laboratory at Carnegie Mellon University. Her research interests include dependable human-computer interfaces, experimental methods, and model selection and uncertainty. She has an interdisciplinary BS in human languages and computer science from Vanderbilt University. Contact her at rroberts@cs.cmu.edu.*