CAVEON

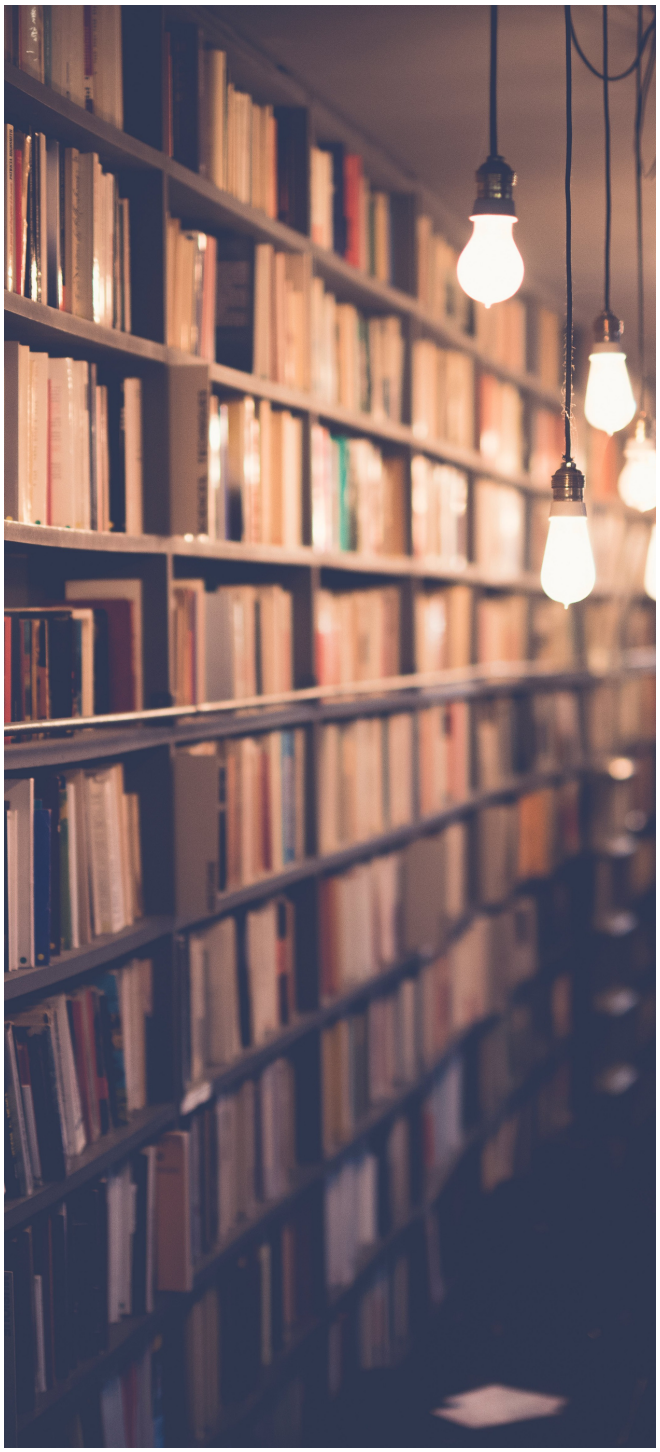
THE LANGUAGE OF SECURITY AND TEST SECURITY

BY DAVID FOSTER | CEO | CAVEON TEST SECURITY

MARCH 2021

INTRODUCTION

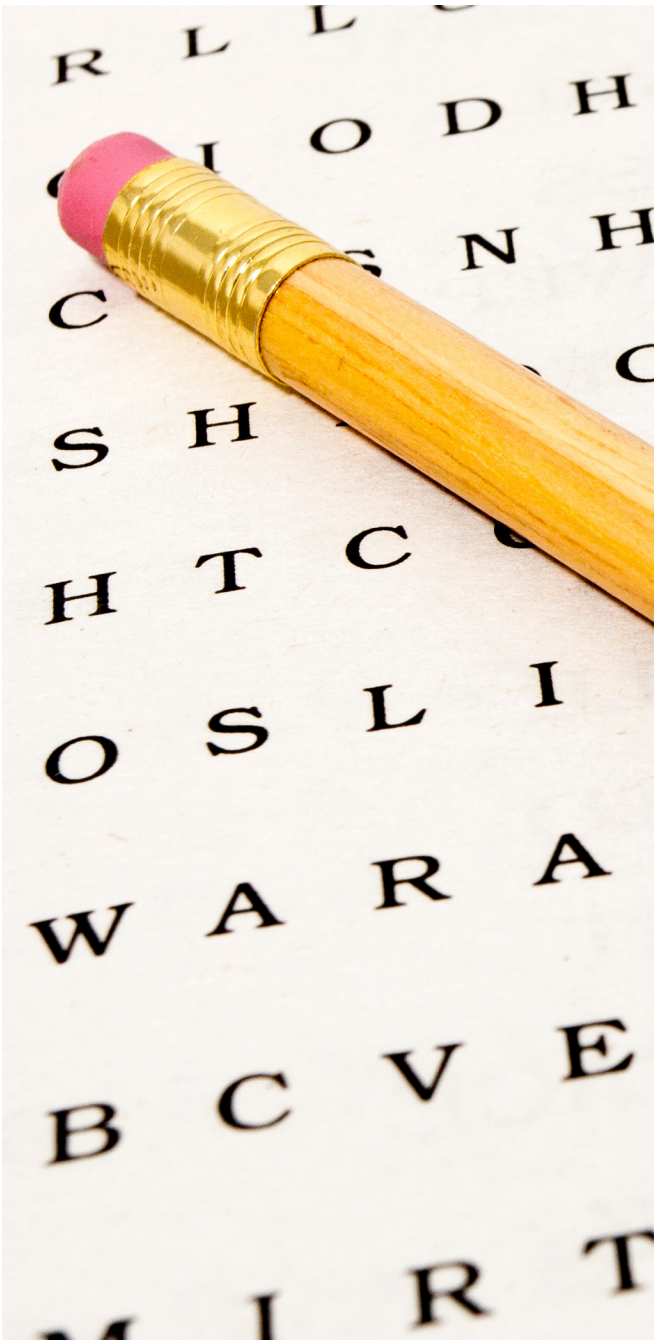
Let me begin this paper with a story. When a few colleagues and I began Caveon in 2003, it was the only area of testing where I could safely work. I had just left my job at a company. My employment contract had non-compete clauses that restricted me from competing, and therefore from working in virtually every area of testing. However, there was a single area in the field of testing that wasn't covered by the non-compete clause; and that area was "test security." Also fortunately for me, I had friends who seemed to believe that test security was a good bet for their future as well. They joined me in the new venture, Caveon.



The point of this brief story is that we started Caveon knowing a bit about test security, but as a focus, it was very new for us. Our combined hundred-plus years of experience counted for something; that was sure. We had designed, built, administered and analyzed more high-stakes tests than any similar group. Despite our considerable experience, we were not security experts. We had not been trained in security specifically. Instead, our security training came indirectly throughout our varied experiences in the testing industry.

Back when Caveon was started, I recall some very wise advice given to me by one of the group's members, Dennis Maynes. He stated on more than one

occasion that he felt we needed to be trained in “security.” He didn’t mean in test security specifically, but in the general principles of security as they may be applied to other fields where security is needed. Obvious examples of these fields include banking, information systems, transportation, military, law enforcement, and even the casino/gambling industry. I heard his advice, but wasn’t convinced enough to take it to heart. It was a few years before I began a serious study of general security principles and tried to understand how they might apply to testing. It was from that study that I learned a new and important language and began using security terms more carefully. It was then that I felt I was able to make real progress and contributions to Caveon and the field of test security.



I am no expert in how words come to be defined and used; but, I have come to appreciate the value of being as exact as possible in my conversations and writings. This may have initially been strongly influenced by my scientific training on the way to obtaining a doctorate in Experimental Psychology. Through the process of publishing scientific papers, including my dissertation, I learned the lessons, often painfully, of choosing my words carefully and communicating clearly. Some individuals in the non-testing fields I listed above have been careful to understand the principles of security, at least for their particular field. While I have not been able to find a good source, document, or book that provides general principles of security, it is not hard to find them described in the writings of individuals in one or more of these security-related fields.

At this point in my career, I can say that I have learned more about test security from the casino, transportation and information systems industries, than I have from the testing industry. I have come to realize that what I learned about security from more than 30 years in testing was a fairly narrow set of steps or "procedures" (e.g., how to proctor exams) that were generally accepted and largely unevaluated in any critical way. Since I took my good friend's advice to look outside the testing industry for security insights, I have been able to understand test security in a deeper way, to see the strengths and weaknesses of what we do, and to conceive of new possibilities. Key to this personal growth was understanding of the language of security.

Properly using the language of security seems obvious to me. We accept that premise in other areas of testing. Consider the terms we use in psychometrics. For example, can we be casual about describing the statistical properties of items using terms such as "p-value" and "bi-serial correlation?" Is it okay to use those terms interchangeably, as if they are synonymous? Of course not—they refer to two very specific statistical outcomes. Even if you don't know what they mean, you know they are specific terms with specific meanings and need to be used consistently and exactly. Likewise, should we use the term, "assessment" as a synonym for "test," which is often what happens; or should it be reserved for when we are trying to describe a broader concept? In order to use the set of test security terms properly, we need to be clear as to their meaning. This leads me to the main purpose of this paper.

SECURITY TERMINOLOGY AND DEFINITIONS

Psychometrics is Not the Same as Test Security

First of all, **psychometrics** is not the same thing as **test security**. Test security is not even a sub-field of psychometrics. It does not serve a useful purpose to try to combine them in any way. They are both professional areas of work and development within the field of testing and measurement. They have different goals, although sometimes those goals may overlap. For example, efforts in both psychometrics and test security contribute to providing evidence for the valid use of test scores, although that evidence is different and is produced through unique ways. One way I have found useful in differentiating the two concepts is that psychometrics has the ultimate goal of creating valid and useful test scores, whereas test security has the ultimate goal of protecting those test scores—and the items that make up the tests—so that they remain useful for as long as possible. Of course, a psychometrician can be trained to be a test security professional, and vice versa. Also, I believe that each should know quite a bit about and appreciate what the other does, although those aspirations are not required.

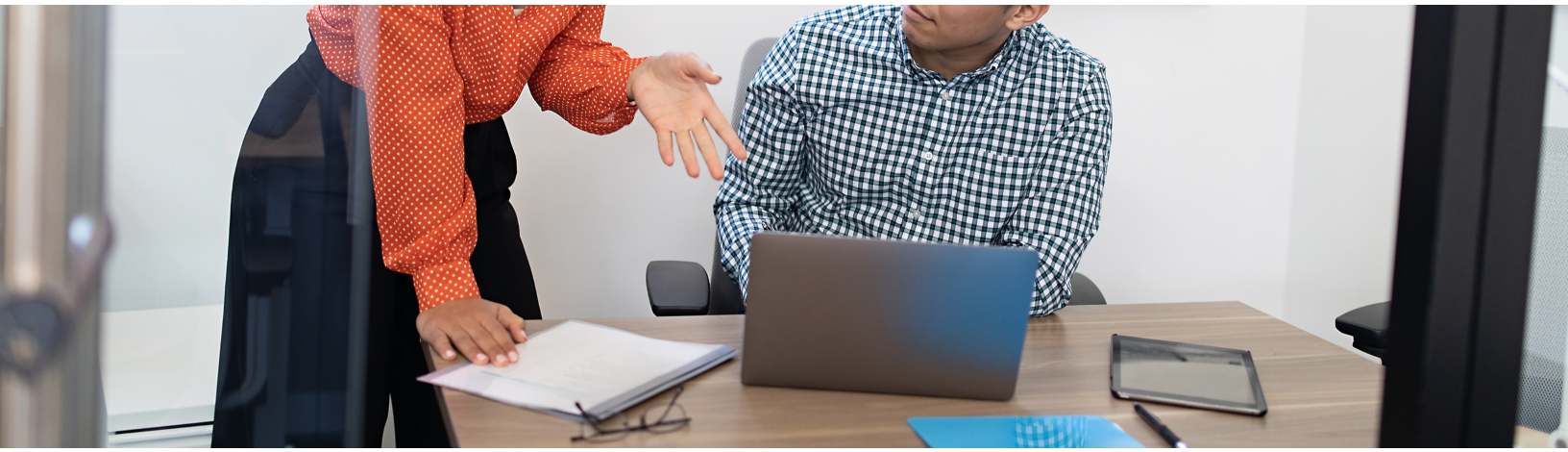
Test Security and Test Integrity are Not the Same

Test security and test **integrity** are not synonyms; they should not be used interchangeably. I'm not convinced test integrity should be used at all in a security context. Test integrity seems to have been introduced recently as a more politically correct term, almost a euphemism for test security—perhaps trying to avoid the perceived harshness of the overall concept of security. But the word integrity misses the mark. In this context, it refers to the "soundness" of a result. If the result is a test score, it has much more to do with how
the score was produced than how it was protected. A strong test security implementation may add to the ability to extol the integrity of a test score, but it is not sufficient.

Here are a couple of other terms that I believe we should avoid using in the direct context of security: anomalies and irregularities. That is, we should not refer to a security incident as a synonym of "test irregularity" or anomaly. Both terms refer to something that happened that is rare or unusual, but are more general by definition.

In test administration procedures, proctors or test administrators may fill out a test irregularity report, a report covering any unusual event occurring during the administration of a test. It can be filled out for relatively unpredictable rare occurrences when the computer crashes, or the power goes out, or an examinee gets angry and destroys the computer. It can also be used to report a security incident, such as the use of a cell phone during the test. Because a test administration irregularity is clearly a broader concept, it should not be used to mean the same thing as a security incident.

I have seen the term anomaly used when reporting the results of data forensics analyses. Given a particular decision threshold, the statistician may report a statistical anomaly—a strange pattern in the data that may be indicative of a test security problem. This is not a confirmed security incident, but a statistical anomaly. Follow-up investigative work may later determine that a security breach is related to the anomaly.



Calling it what it is: Test Security

Let me take a minute to say why it is important that we use the term test security as part of the language in our writing, our presentations, and our conversations; why it is best that we don't apologize for those terms or create and use euphemisms to make test security seem "softer," "more subtle," or more acceptable.

We are up against individuals, some of whom are intent on committing all types of test fraud for no good socially acceptable reasons. They want to gain personally from the fraud, without regard to the harm it causes to others or to the testing program. These are not nice people; they understand what they are doing is wrong, but they do it anyway. Their methods are different, but they are no different, in a moral or ethical sense, from the people who would steal from your home, or take your car, or use skimmers on ATMs, or steal your identity. Others may commit test fraud unwittingly or even with "good intentions" (e.g., trying to help a friend). Regardless of the motives, if we don't adopt a professional and serious approach to security as an industry, then no one else will.

Now, let's get back to more test security terms and definitions.

Types of Test Fraud

Test fraud is a set of activities that are illegal, inappropriate, or against the rules. Cheating and test piracy (i.e., theft or stealing) make up the bulk of test fraud that testing programs can expect, although there are others.[1]

Cheating is probably the most common type of test fraud. Any threat that is cheating can be quickly identified by its goal of increasing a test score beyond what would normally be earned.[2] There are literally thousands of individual variations in ways to cheat, but it is useful to distill them into six categories. Those categories are described in detail in other Caveon papers and in at least one set of industry guidelines (ITC Guidelines on Test Security).

1. Using preknowledge of test content
2. Receiving assistance during the test
3. Using cheating aids
4. Using a proxy test taker
5. Tampering with or hacking into a scoring system
6. Copying answers from other test takers

1. This section doesn't cover the entire set of security threats. There are at least two other types of test fraud that should be considered. The first is a form of cheating where the goal is to increase an average test score across a number of test takers. An example of this is the method used by a school district superintendent who manipulated the test administration rules and used schemes such as forcing low-performing students to remain home on test days, preventing them from taking the test. The second type of test fraud is covered in the footnote for the next section.
2. Cheating does not actually have to result in a higher score. Many attempts at cheating are ineffective or prevented.

Test theft, and its associated threats, on the other hand, does not have the goal of increasing test scores. It represents activities intended to steal, capture, harvest, or otherwise obtain test content illegally. While they may ultimately lead to cheating by others, any increase in a test score due directly to test theft activities by an examinee is improbable. Like cheating, test theft methods number in the thousands; however, they can also be categorized into six[3] basic threats.

6

CATEGORIES OF TEST THEFT THREATS

1. Stealing digital test files or test booklets Receiving assistance during the test
2. Recording content during the exam by digital capture devices
3. Capturing content by electronic recording of the screen
4. Memorizing content
5. Transcribing content verbally (on paper or recording device)
6. Receiving content from test program insiders

Cheating is often inappropriately used as an overall term encompassing **test fraud**, **cheating**, and **test theft**. However, these are different concepts and we must be careful to use the term, cheating, specifically for what it is.

The general process of test fraud

It's most important to understand the basic terminology of the security process. This process uses terms that include threat, attack, breach, and vulnerability. A threat is a potential source of an attack or a breach, a potential source of damage. For example, proxy testing is a general threat, and a particular proxy testing service is a specific threat, whether you know about them or not. Human beings are always behind every threat; they have created the threat and are looking for ways to succeed.

An **attack** is an actual attempt by a person behind the threat to capture your test content or to cheat on your tests. This is where the threat becomes real; actual test fraud is put into motion. Detection systems (see below) should already be in place. The best such systems will detect the attack while it is just starting or in progress. However, it is also valuable to detect that an attack has happened in the past. An attack can be as simple as:

- a test taker peeking over the shoulder of another test taker in order to copy an answer
- one person agreeing to take the test for another person, or
- an attempt to illegally access testing files on a testing center server.

3. A type of threat closely related to test theft occurs when an organization shares, or offers for sale, or disseminates copyrighted test content that is not theirs. Braindumps are the best example of this, although it could occur at a simple level by one student telling another student what he or she saw on the test. Tweeting a picture of a test item to a Twitter audience would be another example.

An attack means that a threat is no longer just a threat. Hopefully it is clear to all those reading this document that many, perhaps most, attacks end up being successful; they end up as a breach. The distinction between a threat, an attack and a breach may be made clearer by comparing them to a hurricane off the coast. Initially, the hurricane is just a threat to people on the coastline. However, as it hits land, it has begun its attack. If it succeeds in doing damage, then a breach has occurred.

A **breach** is defined as a successful attack. Any security defenses have been successfully bypassed without immediate detection:

- Test questions have been captured using a cell phone
- Cheat sheets have been used during an exam.
- A test booklet has been stolen.
- Test questions have become available on the Internet.

A breach can be small, almost ignorable, or can be large enough to set a testing program back a few months or years. Whether small or large, a breach must be taken seriously. Small breaches have a way of turning into large breaches if ignored or undetected.

It is this stage when a breach occurs, that results in damage to a testing program. A test score is higher than it should be, resulting in a bad decision (e.g., hiring, admission to college, grade promotion, etc.). Test content is disclosed, resulting in the cost of creating new test items. The media reports test fraud, resulting in loss of reputation. If there is good news in these scenarios, it is that the damage can be limited by a good detection system, followed by a quick and effective response. Additionally, lessons can be learned and changes made to a test security plan in order to reduce the chance of similar breaches in the future.

A **vulnerability** is a weakness in a security plan or defense:

- Lack of proctoring or monitoring test takers for an important exam is a vulnerability
- Having a single form of the test for thousands of test takers is a vulnerability.
- Not having a policy for confiscating cell phones during test administration is a vulnerability.

A **threat** will exploit a **vulnerability**, resulting in a greater likelihood of an **attack**, of a successful attack, **or a breach**. A vulnerability that is unknown or not strengthened by a testing program will embolden the person behind the threat, increasing the likelihood of a successful attack.

RISK AND RISK ANALYSIS

Risk is a term associated with security, including test security, and is generally used inappropriately. It is usually as a synonym for threat (“Proxy testing is a risk") or vulnerability (“Deciding to use untrained proctors is a risk”). But the term, risk, is accurately defined as the amount of damage a breach might cause times the likelihood of the breach. Risk is evaluated for each threat, which is the source of a future breach. For example, consider the threat that teachers may tamper with answer sheets, changing wrong answers to right answers on a statewide math assessment.

The first question to ask is:

1. What is the amount of damage we can expect from a breach?

The answer can be qualitative (High, Medium, Low) or quantitative (choosing 1-5 on a scale from High to Low), probably based on a previous breach or on what other states have experienced.

The second question is:

2. What is the likelihood of the breach?

Again, the responses can be qualitative or quantitative. Risk is the product of the two answers to the questions. Obviously, a quantitative answer makes it easier to mathematically calculate risk, but a strong evaluation of risk can be obtained by a qualitative approach as well. The process that was just described is called a risk analysis. A risk analysis is best used as a formal process covering all of the threats, as this comprehensive process is less likely to be influenced by preferences, emotional reactions, or memories from an actual recent breach.

The outcome of a risk analysis will have three great benefits. First, the most important threats will rise to the top, allowing the testing program to focus on and create solutions for the threats that can hurt the program the most. Second, security budgets, when available, are not unlimited; this will allow security resources to be allocated appropriately, targeting the threats that carry the greatest risk. And third, test security solutions can be crafted specifically for the most serious threats.

TERMS FOR SECURITY SOLUTIONS

Security solutions come in three broad categories: prevention, deterrence and detection. These are each very different solutions that can be implemented to affect the security threats and reduce levels of risk. The three types of solutions can be identified by their primary goals:

Prevention solutions have the direct goal to prevent test fraud.

Deterrence solutions are psychological with the goal to persuade or convince test takers, and others who would commit test fraud, not to do so.

Detection solutions have the goal to detect an attack in progress or a breach that has already occurred.

If a breach occurs and has been detected, there may be damage that needs to be repaired, such as:

- A compromised test may need to be replaced.
- Test scores may need to be cancelled.
- Cheaters may need to be punished.
- Civil or criminal legal action may need to be started.

Eventually, some of these actions may have a deterrent effect. For example, if test scores based on confirmed cheating are routinely cancelled and the cheaters punished, and those actions are well-publicized, future test takers will experience less of an incentive to cheat because they are worried about getting caught and experiencing the same consequences.

It is important to note that both prevention and detection solutions can have deterrent effects in addition to their primary purposes. As an example, research has shown that the DOMC item type (see a description below), an item design prevention solution, is viewed by test takers as an item type that makes it more difficult to cheat or to steal.

There are two ways to look at prevention solutions. One is absolute. The other is statistical.

An **absolute prevention** solution clearly prevents all attacks by a threat, effectively removing it from the list of likely threats for certain segments of high-stakes testing. One obvious example of absolute prevention is the use of computerized testing to prevent answer copying. Answer copying is the well-known and well-practiced threat of copying answers from a neighbor taking the same paper-based test. Computerized testing with its various methods of test design and randomization of test questions and answer options has likely had the most impact on preventing answer copying. Testing programs worried about this threat simply randomize the order of questions as they are presented. Two examinees taking the same test and starting at the same time would see the questions in a different order. The cheater would gain no significant advantage from seeing how another test taker answered the questions. Add to this other features of computerized testing, including different start times, the randomization of answer options (for multiple-choice questions), and test designs that create a unique test for each examinee (e.g., computerized adaptive testing), and it is clear why the threat of answer copying is no longer considered a serious threat in most testing programs. It remains on the list of threats because there are testing efforts, particularly in secondary or higher education classrooms, where students still take tests using a paper format in rooms where they sit close to each other. For these testing situations, the threat remains a valid one.

A couple of other examples of absolute prevention might be helpful. A program is concerned that its employees, working in the test development group of a large testing program, might copy the test questions and give or sell them to outsiders. To prevent this threat, the program provided the employees with diskless computers with no access to the Internet or printing resources and access only to a local server where their work was stored. With those restrictions, it became impossible to print paper copies of the items or to copy the questions to a USB drive or other device.

As a final example, if a program were concerned about proxy test takers (defined as a person who takes a test on behalf of another person), that threat can be easily prevented by using one or more biometrics, such as a photograph, vein scan, or voice recording. The logic is simple. When an examinee registers with a program, he or she supplies a biometric along with all other application requirements. That may be a photograph, an oral recording of a phrase, a keystroke pattern, a palm scan, or others.

With that biometric result connected to the applicant, it must also be provided by the examinee on the day the test is scheduled.[4] A strong authentication method prevents proxy test taking on test day.

Statistical prevention is a little less obvious concept, but is prevention nonetheless. At its foundation, a method is preventative if it makes cheating or theft difficult to do. The use of multiple equivalent forms is a good example. The creation of multiple forms is a secure test design put in place to reduce exposure rates of items and to prevent individuals from seeing the same item if the test is re-taken. With that method in place, it is still possible, although more difficult, for all the forms to be stolen. An examinee can still cheat, but that ability to cheat is made more difficult because the examinee does not see the same form two times in a row.

Using the Discrete Option Multiple Choice (**DOMC**) item type is another example of statistical prevention. The DOMC item presents the options (i.e., answer choices) one at a time, accompanied by YES and NO buttons. If the test taker believes the option shown is the correct one, he or she would click on the YES button, otherwise the NO button. This simple design change allows the item to score as correct or incorrect without exposing many of the options. Because only about 50% of the item option content is displayed to a test taker, it prevents that test taker from seeing and then sharing the 50% he or she did not see. Plus, it prevents a new test taker, informed by the first one, from effectively using the bit of information that was presented. This is because the same item on a future test for the new test taker will likely present a different set of options. Certainly, over dozens or hundreds of test takers, all of the content will have been presented, but it would require a painstakingly organized harvesting effort to gather it all together in a useful way.

To finish up this section, let me provide examples of absolute and statistical prevention of security from another field where security is important. Consider the security a bank might use to protect its cash holdings. Absolute prevention would mean not keeping cash in the bank at all. Theft is prevented because there is no cash available. Statistical prevention would mean putting the cash in a very strong vault. The cash can be stolen, but doing so won't be easy and it won't happen often.

Deterrence is a test security method that affects the psychology of the person behind the threat. The deterrence measure actually causes the person to decide not to attempt to cheat or steal test content. Here are a few examples:

---

4. Authentication, Identification and Biometric. While a bit unrelated to the general nature of this paper, it's important to know the definitions of authentication and identification in order use them correctly and to propose and implement solutions. They are probably mis-used most of the time. Authentication is simply a match of a biometric given at the time a person signs up to be part of testing program with one presented on the day of the test. If this one-to-one comparison is a match, the person is authenticated, is allowed to take the test. A biometric can be anything that is a unique characteristic of the examinee (photo, typing pattern, facial pattern, and even a name and address). In the testing context, identification, or actually identifying a person, is rarely if ever done. Even presenting a government-issued identification (e.g., driver's license) to a proctor requires the proctor to simply compare the name and address on the ID with the name and address of the person who is on the schedule to take the test. The proctor's job is never to truly identify a person. Because ID's can be easily forged, biometrics that do not carry that risk should be preferred.

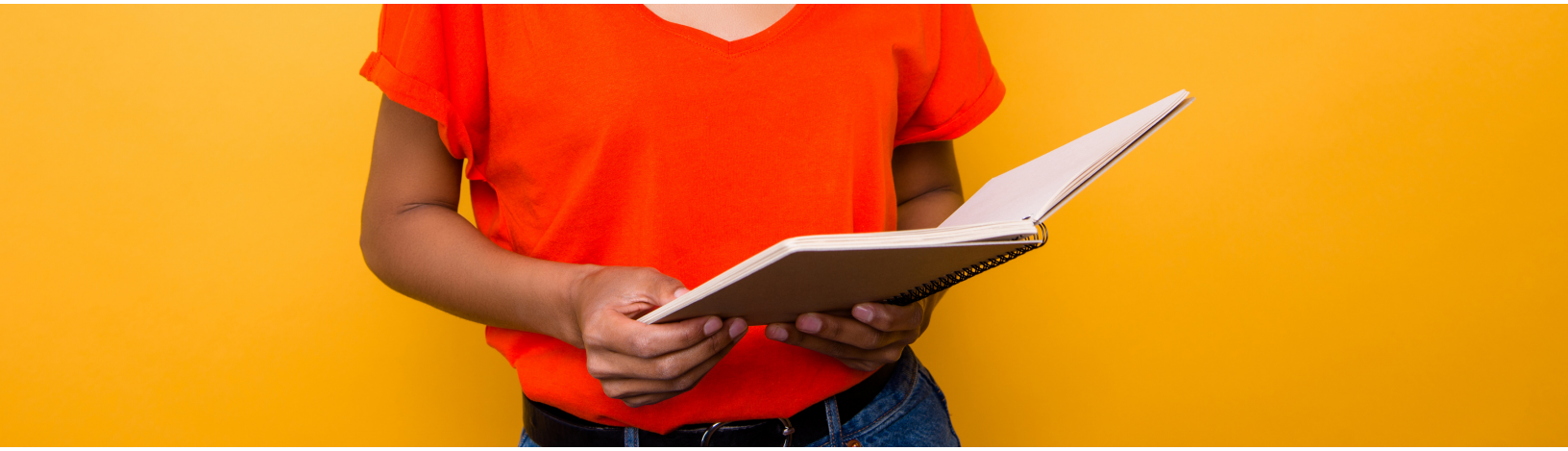
- The testing program begins a program to educate examinees and other stakeholders that test fraud of all types is bad for everyone involved.
- The testing program publishes and distributes test security rules and the consequences for breaking them.
- The program publicizes an aggressive and effective response to a breach, where the individuals were caught and punished. Salient details are helpful here.
- The program describes the systems put in place to easily detect or frustrate test fraud. If examinees felt that there was a high probability of getting caught or that cheating has just become more difficult, they may reconsider their plans.

A testing program should make every effort to deter test takers from attempts at test fraud.

While not their primary security purpose, efforts at detection and prevention also have a deterrent effect. Knowing that the test is computerized with the randomization of questions, a test taker will be dissuaded from even trying to copy the answers provided by a neighbor, even if the opportunity to try is there. If test takers knew that very sophisticated data forensics analyses would certainly catch any attempt to use pre-knowledge to cheat, they would be afraid to get information from a braindump site. Deterrence should be considered a secondary security goal for all prevention and detection procedures.

The final set of security methods can be considered under the heading of **detection**. With these methods, there is no goal to stop an attack or to deter the person behind the fraud. The purpose of detection is to detect an attack in progress or a breach after it has occurred. If the attack can be detected, then perhaps it can be stopped before any damage has occurred. If it can be detected post hoc perhaps the damage can be repaired or mitigated, and steps taken to prevent or deter a similar attack in the future.

Here are a few examples:



Data forensics is a set of security methods dedicated to detection. Data forensics analyzes examinee responses in a set of test results, looking for unusual patterns that may be indicative of various types of test fraud. For example, a pattern of a number of tests with too-similar responses from one classroom may indicate coaching by the proctor/teacher. Or, an unusual rise in scores from one year to the next might indicate some type of cheating or that test content was stolen. Testing programs quickly move into action when such outcomes are discovered.

Another example of detection is proctoring. The primary responsibility of a proctor is to discover types of test fraud as they are occurring.[5] For example, the proctor might spot a fake ID presented by the examinee or they might catch a person using a cell phone to take a picture of the testing workstation screen. And of course, proctors are always on the lookout for the use of well-hidden cheat sheets. Having a proctor cruise the testing room serves as a deterrent as well, but this is not their primary responsibility.

A final solid example of a detection system is web monitoring. Caveon's own Web Patrol group scours the Internet looking for the disclosure and sale of copyrighted test content. Finding such content indicates that a breach has occurred and may provide information as to how and when it occurred. Web monitoring may not detect the theft in action, but certainly detects its outcomes and gives the testing program enough information to take action.

5. John Fremer, a colleague of mine at Caveon, argues that the primary responsibility of the proctor is deterrence. He may be right. Nevertheless, the point is debatable and therefore proctoring remains as a good example (also) of detection.

All three of these types of security solutions—prevention, deterrence, and detection, in as many specific methods as can be devised—should be used for every threat that poses a significant risk to the testing program. Overkill is probably a good thing when considering the full range of test security threats.

Parting Words...

Thanks for taking the time to read the paper. I hope you, like I, realize the value of standardizing our use of security terms. If we can make this effort, I believe we will make greater progress against the individuals behind the growing test fraud we seem to be experiencing. The understanding of our risks will be clearer, we will communicate our concerns and plans more effectively, and we will have more confidence in our decisions about the security solutions we use.

Overkill is probably a good thing when considering the full range of test security threats.