Assessment Security Options: Considerations by Delivery Channel and Assessment Model

January 23, 2013
ATP Security Committee (ATPSC)

Five years ago, the establishment of an ATP Security Committee (ATPSC) highlighted the ever-growing concern about intellectual property (IP) theft that compromises assessments developed by and used within the testing community. Since that time, the attention on security concerns within the industry has grown and expanded with the ongoing enhancements to technology available to assessment providers and the test-taker population.

The ATP Security Committee provides a forum to encourage assessment organizations to collaborate in addressing test security concerns. The goals of the Committee are to: (1) identify methods to improve test security; (2) establish and disseminate security best practices; and (3) protect the integrity of assessments and assessment programs. By focusing on industry needs and providing solutions, we are creating a resource center for ATP members to access information and assistance for implementing or improving security efforts and programs.

One of the ATPSC’s most recent initiatives grew out of requests by the ATP members for information on security options available for various assessment programs, based on the delivery channel(s) and assessment model(s) employed. This document is designed to address this need by providing an overview of available delivery channels and assessment models, along with the procedures and best practices currently in use to enhance security across the different program options.

Nikki Eatchel
Chair, ATP Security Committee, 2012
Acknowledgements

ATPSC Test Delivery Sub-Committee. This document was developed by the ATPSC Test Delivery Subcommittee. We want to acknowledge their contributions and efforts to create this report for the ATP community. We also want to thank the ATP staff for the support of this project.

ATPSC Test Delivery Sub-Committee

Liz Burns
Juniper Networks

Aurora Hamilton
American Registry of Radiologic Technologists

John Dight
Kryterion, Inc.

Greg Stephens
Microsoft

Cathy Donath
Donath Group, Inc.

Beverly van de Velde
Symantec Corporation

Chuck Friedman
Association of Social Work Boards

Lauren Wood
American Registry of Radiologic Technologists

DISCLAIMERS:

Not all sections or subsections of this document may be applicable to a particular user. Consequently, this document is intended only to provide general, high-level information to any user on available security options for assessment programs.

While the ATP has made every effort to ensure that the information contained in this document has been obtained from reliable sources, all information is provided “as is” and neither the ATP, nor any participating publishers or service providers, make any warranty, express or implied, nor do they assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, product, or process described in this document. In no event will the ATP, its agents or employees thereof be liable to any user for any decision made or action taken in reliance on the information in this document or for any consequential, special or similar damages, even if the ATP has been advised of the possibility of such damages.

The information in this document is provided with the understanding that the ATP, and the individuals who participated in the preparation of this document, are not engaged in rendering legal, technical, or assessment and psychometric advice and services. As such, this document should not be used as a substitute for consulting with competent legal, technical, psychometric or assessment advisers.

Additional ATP Resources

This document refers to various security considerations that should be taken into account within an overall security plan. A number of additional resources are available for ATP members regarding security and best program practices, including:

Document Overview

This document contains the following four sections:

Section 1: Introduction
Section 2: Exam Delivery Models
Section 3: Exam Delivery Channels
Section 4: Practices to Enhance Security

The first three sections are intended to provide an overview of general security considerations, available exam delivery methodologies, and typical exam delivery channels. They are not intended to be comprehensive or to provide the full set of information needed to inform every program decision. These sections are included to outline general security considerations and to provide context for the security discussions in Section 4. Throughout this document, the term organization is used frequently. The term organization can also represent a sponsoring organization or company.

Section 1: Introduction

Security Threats and Risks

Basic terminology for a successful security effort includes the concepts of threats, risks, vulnerabilities and breaches. Knowing that specific security threats exist, and that they can be associated with an estimated amount of risk, a sponsoring organization, test publisher, or program administrator can develop and implement a successful program. Given circumstances at a specific point in time, risk can be calculated informally as: (1) the likelihood a threat will be successful; (2) the ease with which program vulnerabilities can be exploited; (3) the amount of damage a threat might cause as it becomes a successful breach; (4) how prepared the program is to stop the breach and repair the damage; and (5) the likelihood that the organization can learn from the breach and mitigate similar breaches in the future.

To establish an effective security plan it is important to understand the nature of current security threats and risks. A security threat is any source of potential cheating or test theft. For example, the use of cell phones to receive text messages during an assessment is considered a cheating threat. In the area of test theft, a threat would be the potential for someone to hack into a test storage device and capture the content of the entire test. A security risk is the likelihood of a threat and the amount of damage that specific threat would cause. As the variety and multiplicity of particular threats and risks for a testing program are better understood, it is possible to put in place an effective security plan. Properly managed, that security plan will reduce the security risks and mitigate damage from breaches.
Types of Cheating and Test Theft

There are various categories of cheating and test theft threats that should be considered by testing organizations. The two tables below provide categories of cheating and test threats respectively.

Table 1. Categories of Cheating Threats

<table>
<thead>
<tr>
<th>Cheating Threats</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using Test Content Pre-knowledge</td>
<td>Test-taker obtains actual test questions prior to the test administration.</td>
</tr>
<tr>
<td>Receiving Expert Help While Taking the Test</td>
<td>Test-taker receives help from a teacher, proctor, or other source during the test.</td>
</tr>
<tr>
<td>Using Unauthorized Test Aids</td>
<td>Test-taker uses non-authorized aids during the exam, such as cheat-sheets, cell phones, headphones, programmable calculators, etc.</td>
</tr>
<tr>
<td>Using a Proxy Test Taker</td>
<td>Test-taker uses a professional proxy testing service or simply has a friend or colleague take the test.</td>
</tr>
<tr>
<td>Tampering with Answer Sheets or Test Results</td>
<td>Following the completion of test, a person tamper with answer sheets, changing incorrect answers to correct ones, hacks into the test scoring database in order to raise test scores, or forges score reports with inaccurate (higher) score information.</td>
</tr>
<tr>
<td>Copying Answers from Another Test-taker</td>
<td>Test-taker copies answers provided by another during the test.</td>
</tr>
</tbody>
</table>

Table 2. Categories of Test Theft Threats

<table>
<thead>
<tr>
<th>Test Theft Threats</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stealing Actual Test Files or Booklets</td>
<td>At particular stages of test distribution when digital test files are stored on a server at a test site or test booklets are physically stored on location, test content is most vulnerable to theft. Poor access controls allow unauthorized individuals to capture entire test content along with answers.</td>
</tr>
<tr>
<td>Stealing Questions in Test through Digital Photography or Copying Devices</td>
<td>Test questions are captured as they are displayed during a test. The test-taker uses a hidden or otherwise undetectable high-resolution digital camera or other copying devices (e.g., pens that can).</td>
</tr>
<tr>
<td>Stealing Questions by Recording Test Content Electronically</td>
<td>For technology-based tests only, the entire test session (including all test questions) is captured with an automated procedure by using a recording system connected to one of the computer’s output ports.</td>
</tr>
<tr>
<td>Memorizing Test Content</td>
<td>Test-taker memorizes questions to be recalled at a later time. As part of an organized effort, this kind of theft is termed “harvesting.”</td>
</tr>
<tr>
<td>Transcribing Questions Verbally</td>
<td>Test-taker uses oral or written recordings of test item content during the exam. This may involve audio recording devices, text recording devices (e.g., cell phones, or notepads/scratch paper). Test-taker uses two-way radios or cell phones to capture and transmit the content of the questions.</td>
</tr>
<tr>
<td>Obtain Test Material from Program Insider</td>
<td>An employee or contractor of a testing program gains unauthorized access to test content and distributes it without permission.</td>
</tr>
</tbody>
</table>

With the growing number of security concerns and finite time and resources available to address them, it is important for organizations to understand the options available to enhance program security. Those
options vary based on the type of exam model and the types of delivery channels used by the organization. The remaining sections in this document are designed to cover the following:

- Overview of the major exam models currently in use within the assessment industry, including the advantages and disadvantages associated with each.
- Overview of the major delivery channels currently in use within the assessment industry, including the advantages and disadvantages associated with each.
- Current best practices in place to enhance security, described in the context of the exam models and delivery channels in which they can be used.

Section 2: Exam Delivery Models

When planning an assessment there are many factors to consider prior to development and delivery of the exam to test-takers. One of the first key questions relates to the overall test design that will be used, as this will drive (among other things) the development process, the delivery channels available for use, and the psychometric models to be employed. The following section is intended to provide a fundamental overview of the major exam models available for consideration. The information is included to provide context for the subsequent discussions regarding available best practices to enhance security. This section is not intended to provide comprehensive documentation on each methodology or to substitute for the organization’s extended research and consideration needed prior to determining the best model to be employed for a specific assessment.

A) Linear Fixed Exam Forms

General Description of the Design

Linear fixed exam forms refer to a single collection of items (both scored and pilot items) that constitute an assessment. In this model all test-takers receiving the designated exam form are presented with the same set of items. This model can be used with either paper-based or technology-based delivery. In the purest sense, a linear fixed form design would begin with the initial item of the exam and each test-taker would proceed through the exam, one item at a time, until the last item is completed. Items may be presented in the same order (a true linear test design) or the presentation order of the items may be randomly presented (if using computer-based testing). The key feature of the linear fixed form test design is that all examinees who take a given exam form see exactly the same set of items over the course of the exam.

Although all test designs have subject matter experts involved in the review, approval, and selection of items, only the linear fixed form test design includes subject matter expert review of the exact combination of items that every test-taker will see as an intact assessment. This means that all test-takers given a specific exam form will be seeing an assessment that has specifically been reviewed and approved as an intact form (not a subset of preapproved and reviewed items in a pool or item bank). The specific item content and difficulty represented in each intact form is reviewed and approved as an administrable unit. Basing all test-taker scores on the same set of items means that the test scores (within a given exam form) are directly comparable across examinees, both from a content and psychometric perspective.
Features/Considerations of the Design in Regard to Security

Alternate linear fixed forms of an examination are frequently developed as a response to exam security concerns. Test-takers are randomly assigned a fixed exam form from among a number of alternative (equivalent) forms. The alternate forms generally contain some subset of items that are in common as a psychometric anchor item set, but otherwise the specific items are different (although efforts are made to balance the content and the psychometric properties of the items across exam forms as much as possible). Because any given test-taker does not know which exam form s/he will receive prior to the administration, this lack of predictability or advance knowledge represents an enhancement to exam security. In addition, test-takers receiving alternate exam forms will find it more difficult to inappropriately try to compare answers following the exam administration. Alternate exam forms are also useful if unsuccessful test-takers have the potential to re-take the examination within a fixed period of time because they will not see the same exam form on subsequent attempts to pass the examination.

Available Test Delivery Methodologies within the Design

Paper-based tests (PBT) comprise a large portion of linear (fixed form) test design. Security concerns include training of proctors; space and seating design of the test administration environment to discourage cheating (seating test-takers far enough apart to discourage copying of answers); secure printing and distribution of test materials; and collection of secure test materials for scanning, storage or disposal. A thorough and well-written test administration manual is one of the key methods to ensure strict compliance with materials handling, exam administration, and collection and disposal of materials. Procedural details are complex and there are many points at which the logistical aspects of the process may have an impact, or potential impact, on exam security.

Computer-based testing (CBT) is another option for the linear (fixed form) test design. CBT delivery can reduce some of the exposure considerations of secure materials (e.g., hard copies of exam forms), but depending on the characteristics of the CBT delivery platform, there may still be a number of security concerns. One advantage of implementing a linear (fixed form) test design via CBT is that the test developer has the potential to implement randomized ordering of sections, items and item options, such that each test-taker (within an exam form) receives the same set of items in a uniquely randomized order of presentation. Randomizing the order of item presentation is primarily done to enhance exam security.

Pros and Cons (Security) of the Design

Pros
Having alternate (equivalent) exam forms available so that test-takers are assigned an exam form at random helps to enhance exam security. The use of alternate (equivalent) exam forms with re-testers is also a useful security feature since it can ensure that they do not receive the same exam form as on their previous attempt(s).

Cons
Even with a number of alternate (equivalent) exam forms available, the presentation of fixed forms (whether randomized in order or not) is less secure than some of the test design alternatives that provide for specific controls on item exposure and a more thorough utilization of items in a large item bank.
Pros and Cons (General Business) of the Design

Pros
A linear fixed form model is an advantageous test design for an initial transition from PBTs to CBTs. It capitalizes on the test development expertise and available content that the organization has in place. In addition, the linear fixed form test design is less technically demanding than other test designs (compare with LOFT and CAT). And depending on the type of assessment program, the technical sophistication of more complex test designs may not be needed for an organization. Lastly, the quality control burden is less with linear fixed forms than what it would be with more complex test designs. Based on psychometric considerations and sample size, the linear (fixed form) test design is a viable option for most assessment programs; this is particularly true with low-volume exams.

Cons
With a linear fixed form model, regular replacement of the exam forms to reduce item exposure is a critical component of maintaining exam security. Frequent replacement of linear forms requires ongoing development, which must be accounted for within staffing, schedules, and program budgets.

B) Linear on the Fly (LOFT)

General Description of the Design
Unlike a traditional paper-based or linear form, a Linear-on-the-Fly (LOFT) methodology is considered a bank-based computer delivery model. In this model, each test-taker is delivered a relatively unique form generated at the time of administration, based on the content and psychometric requirements established for the program. The actual level of uniqueness will depend on a number of factors, including, but not limited to:

- Total item bank size;
- Total candidate pool size;
- Distribution of items within the available bank across the content domains and sub-domains; and
- Distribution of items within the available bank across the range of statistical characteristics targeted for delivery.

Based on the use of pre-tested items with stable statistics, LOFT forms are not only equivalent based on the content matter covered, but they are also statistically pre-equated to ensure equitable levels of difficulty for every candidate. Pre-equating is typically supported by item response theory (IRT), which looks at such parameters as difficulty (how easy or hard is the item), discrimination (how well does an item differentiate among candidates), and chance (the likelihood a candidate can simply guess the correct answer).
Pros and Cons (Security) of the Design

Pros
A LOFT approach has several security advantages. First, because a LOFT exam is a bank-based exam, item exposure is more broadly spread across a larger group of items (the LOFT bank) and therefore, exposure of each individual item in the bank is reduced. This increases the shelf-life of the items and decreases the number of times an item is seen by test-takers during a given administration window. Second, because each test-taker is delivered a virtually unique exam, it is far more difficult for test-takers to memorize items in a way that will be useful or impactful to them other test-takers in future administrations.

Cons
The negative security aspect of the LOFT model is the potential for full bank exposure. Depending on the CBT mode used for the LOFT delivery (server-based or application based) the risk for breach of the entire item bank varies and should be considered carefully.

Pros and Cons (General Business) of the Design

Pros
As noted previously, a LOFT exam is a bank-based exam, which reduces the exposure of individual items within the bank. This increases the shelf-life of the items and allows an organization to obtain a better return on investment, because each item can typically be used for a longer period of time prior to retirement.

Cons
The disadvantage to the LOFT approach is the fact that it requires a larger item pool than a linear fixed form model. It is typically suggested that a LOFT pool be, at a minimum, 10 times larger than the number of items delivered per form. So, for example, a 100 item test would require 1,000 items in the LOFT bank. Developing such large item banks has cost implications for an organization employing this methodology. In addition, an organization may not have staff with the appropriate technical expertise to successfully plan for, implement, manage, and oversee more psychometrically-complicated test designs. Finally, it is more challenging to monitor that LOFT exams are properly functioning within the test delivery system of the exam administration vendor. The process of verifying technical details like appropriate implementation of content balancing, item selection, item exposure controls, more sophisticated scoring, and many other details is more complex and may require different staffing plans and longer time-lines for assessment development and implementation.

C) Computer-Adaptive Testing (CAT)

General Description of the Design

Computer adaptive testing (CAT) is another method based on use of a calibrated item bank, similar to LOFT. Unlike a LOFT model in which exam forms are created independent of test-taker responses, a CAT model is based on an algorithm which selects items for a test-taker based on the test-taker's responses to previous items. For example, a CAT usually begins by administering an item of average difficulty to the test-taker. If a test-taker responds correctly, a more difficult item is selected from the
item bank. If the examinee responds incorrectly, an easier item is selected for the next presentation. This process continues and at each successive step, the algorithm calculates a preliminary total score and selects an item that is optimal to measure the examinee’s preliminary ability level. Because the test is tailored to a test-taker’s ability level, test-takers do not waste time responding to questions that are too easy or too difficult. For credentialing tests, CAT administers items until it has measured a test taker’s ability with a sufficient degree of precision and the appropriate content specifications to determine if the candidate is above or below the designated cut score.

For example, test-taker A could demonstrate consistent ability and will only receive 75 items. However, test taker B may show variance in ability through his/her responses to answers and therefore will receive many more items before a final ability determination is made. Thus, each test is tailored to a test-taker’s ability and the number of questions a test-taker receives varies. The test developer sets the stopping rules for the CAT delivered exam based on the level of proficiency precision desired. No matter how expert or below ability a test-taker is, each test-taker receives a challenging exam that delivers items through CAT that are predictive and adaptive to each test-taker’s level of ability. Unlike linear tests, CAT generally does not allow test-takers to skip questions or review questions and change answers. This is sometimes noted as a concern from the test-taker population.

CAT relies on the psychometric application of Item Response Theory, the production of a large number and well-designed pool of items, pretesting (calibration) of scored items, a large enough sample of examinees to pretest the items to develop the item pool (e.g., best practice analyzes each item 500 times), and software algorithm to deliver items and make decisions in real time. Another application of CAT is “multistage testing”, which uses an algorithm to select groups of items (testlets) in stages. There are more complicated variations of CAT not described here (e.g., performance-based testing). One source for additional information about CAT is the International Association for Computerized Adaptive Testing (IACAT), www.iacat.org.

**Pros and Cons (Security) of the Design**

**Pros**

As compared to linear fixed-form tests (presented by a computer or on paper) where all test-takers are administered the same set of scored items, CAT greatly reduces item exposure risk because test-takers typically receive different and fewer items. Overall, CAT balances items in all areas and selects equivalent items with decreased item exposure.

**Cons**

CAT is generally used with large assessment programs due to the robust item banks required and the more complicated test drivers needed to support the delivery. A significant security issue with the CAT approach is the fact that the calibrated CAT item pool resides on a content storage computer (server). If all items within the pool are on a single server and that server should ever be compromised, the test sponsor is at risk for losing the entire item bank. One mitigation strategy is to break the full item bank into different pools and rotate the pools so that the full bank is never exposed in one administration. In addition, in the unlikely (or undesirable) situation in which an item in the bank is found to be problematic after it has been delivered operationally, exam results cannot be re-scored as easily as in traditional, linear models of administration.
Pros and Cons (General Business) of the Design

Pros

Similar to LOFT, CAT exams are bank-based, which reduces the exposure of individual items within the bank and can extend the life of the items within the bank (assuming the content of the items remain relevant). One major financial advantage of CAT is that it can greatly reduce the proctored “seat time” cost charged by test delivery vendors. In some cases, application of CAT can reduce the number of items and seat time by 50%. Consequently a four–hour linear test, may be reduced to two hours or less, decreasing the associated hourly costs for proctored delivery.

Cons

The major disadvantage is that CAT requires large item pools of pre-test items, using IRT calibrations with large sample size. Developing such large item banks has cost implications for organizations employing this methodology. In addition, the organization must have psychometric expertise on staff or regularly available for the evaluation of item readiness and design, and to initiate evaluation of the efficacy of the organization’s CAT algorithms. A test sponsor that implements CAT must also invest in a “communications plan” (handbooks, video, public relations campaign to schools and examinees, etc.) to educate test-takers about the fairness of CAT, explain why test-takers cannot skip or review items, and document why some test-takers receive a pass/fail status report after responding to a smaller number of questions than other test-takers. The CAT model is often difficult for test-takers to understand and it places a larger burden on the organization to fully explain the process. This may also lead to an increased level of customer service needed to respond to test-taker questions and to help ensure that test-takers understand that the CAT model is fair to all test-takers.

Section 3: Exam Delivery Channels

As with the determination of exam models, organizations have choices in the delivery channels that they will utilize for their assessment. The following section is intended to provide a fundamental overview of the major delivery channels available for consideration. Delivery channels are sometimes used independently and are sometimes used in combination, depending on the assessment program. The information is included to provide context for the subsequent discussions regarding available practices to enhance security. It is not intended to provide comprehensive documentation on each available channel or to substitute for the extended research and consideration needed prior to determining the best channel or channels to be used for a specific assessment.

A) Paper-Based Test Delivery

General Description of the Methodology

Paper-and-pencil instruments refer to a group of assessments in which candidates read questions and respond in writing. Examples of question types include multiple-choice, true/false, matching, and constructed response (essay or short answer). These multiple question types can be used independently or combined in the same testing session. Questions can include maps, illustrations, scenarios, figures, tables, etc. For most high-stakes purposes, a paper-based assessment is an examination administered to a
test-taker using a test booklet and a scannable (“bubble” optical mark reader) answer sheet. Essays and short answers may be collected in a separate booklet.

Features/Considerations of the Methodology in Regard to Security

PBT testing can be most effective for large group administration or in situations that lack technical infrastructure for computer-based testing. PBT provides a viable alternative; however, there are multiple points in the process where the chain of custody of materials can be broken, threatening overall test security. Documentation of these processes, clear delineation of roles and responsibilities, and process redundancies are all essential in setting up a secure PBT program to prevent the loss of a test booklet and the compromise of test questions. The test sponsor must address instructions to the printer, secure storage during printing, shipping and receipt at the test site, distribution to candidates at the test site, and return and final destruction of materials. Considerable efforts must also be made to ensure that the established procedures are being followed by all parties and included in the administration contracts.

Available Test Designs within the Methodology

PBTs are either a single, linear fixed form with items presented in the same order or multiple versions of linear fixed forms with alternate versions containing the same items presented in a different order. This is an approach to deter item recall based on sequential item number and to prevent individuals sitting next to one another from “copying” answers.

Pros and Cons (Security) of the Methodology

Pros

Administering the same test on a fixed date can mitigate widespread sharing of test content. All test-takers see items only on that day (though examinees in different time zones can conceivably share information if steps are not taken to protect against that or to monitor whether such sharing occurs). Same day testing can also prevent one impersonator sitting for the test at multiple sites. While same day testing is also possible with computer administration, PBTs do not typically have the same infrastructure constraints or capacity limitations.

Another advantage to the PBT delivery methodology is a psychometric one. If the psychometric analysis (equating and calibration) is conducted after the actual test administration, this can increase the percentage of new items presented on each test form. By minimizing common items (overlap) from prior exams, the test can include a majority of new items and limit the potential of previously exposed or potentially compromised items from appearing on future exams.

Cons

The major weakness of PBT testing is the multiple points in the printing, distribution, administration, and collection process (which can lead to loss of a test booklet as noted above) and the level of effort needed to standardize those processes. A strong and often extensive staff is needed to achieve appropriate chain of custody throughout these processes. Moreover, the loss of a single, physical test booklet (or many booklets in a shipment) at any step can compromise the entire testing program.
General Pros and Cons (General Business) of the Methodology

Pros
PBTs are an efficient method of assessment and can be a less expensive methodology than computer-based testing because many test-takers can be assessed at the same time or on the same day in multiple locations. PBTs can be administered in many different types of facilities (e.g., conference centers, hotels, and classrooms). A PBT is typically used when a large number of people are together in one location and are being tested at the same time, such as an annual conference, large meeting, state-wide school-based assessment, large-scale admissions testing (e.g., ACT, SAT), or at the conclusion of a classroom-based training session, where computers or Internet access is not readily or reliably available. Also, some smaller volume testing programs also use PBT to minimize costs.

Cons
The administration of paper-based assessments raises a number of logistic challenges, which can be viewed as difficulties or restrictions. The most common approach for PBT administration includes hiring individuals (or using teachers) to serve as proctors and securing locations for administration. Thus, practices and procedures must be detailed in a well-written and comprehensive test administration manual (e.g., check-in procedures, seating requirements, physical security of testing booklets throughout the administration). The manual must also describe test site selection and appropriate room set-up at sites not designed specifically for test administration; selection and training of test center personnel; printing and sealing examination booklets under secure conditions; shipping, receipt, handling, storage, and distribution of materials before, during, and after the administration; and destruction of materials. Extensive time and effort are needed for the development of such procedures, as well as for the training of the personnel involved in the process and auditing of their conformance with such procedures.

All aspects of shipping and document processing are also an important cost component for PBTs. The additional time and effort needed to ensure the security of test content during transport and for tracking and retrieval of materials can be significant and can also increase processing time. The type of shipping used within the administration cycle also adds cost and may range from use of standard or specialized freight companies, to express shipment requiring secure signature, to use of armored car transportation. Printing all the materials is also a base cost component. Should an error (typo, problematic item) be detected after materials are prepared, it may be necessary to reprint all test booklets or use other corrective actions (e.g., placing corrective labels on each page of material with issue). The use of alternate paper forms also adds to the cost printing and complexity of scoring.

B) Computer-Based Test (CBT) Delivery (Server-Based)

General Description of the Methodology

Server-based, CBT testing is a common delivery method used by many organizations to avoid the necessity of directly accessing the internet for test administration. The sponsoring organization delivers the exam content and requirements to an exam delivery vendor to administer exams on its behalf. The vendor then loads that content into its corporate servers and then, at the appropriate time, electronically pushes the exams to the test center servers within its delivery channel (network of test centers) which enables administration without providing test-takers with direct access to the internet.
Each of the test centers has a server that hosts (caches) the content and redistributes the content to client machines when a candidate is scheduled to take the exam. Although tests are typically downloaded to the test center servers on a daily basis, it can take several days for a test to show up across the entire channel. This approach provides organizations with the flexibility of leveraging the delivery vendor’s systems to warehouse content, manage item and exam performance, build exams, schedule test-takers and coordinate the updating of the exams at each center as necessary.

Features/Considerations of the Methodology in Regard to Security

For programs with global audiences this is a popular method due to the sheer size of the delivery channel. Large test delivery vendors have thousands of test centers located throughout the world across the globe, all of which can be made available to a program to achieve the maximum possible coverage. However, with this broad reach comes the challenge of securing thousands of domestic and international test centers. Minimizing this risk is a significant effort, and ideally includes strong user agreements, intelligent test design, frequent renewal of items, use of data forensics and constant monitoring. Additionally, the potential risk associated with delivering exams globally means that the selection of the test delivery vendor is critical. Ensuring that the security policies, protocols, systems, and enforcement approach align to your program should weigh heavily in the selection process. Lastly, additional risk includes the diversity of test centers that can exist, test centers can exist that are not part of the broader vendor network, but rather are specific, contracted test sites set up for a particular program. This scenario poses an added potential risk since these contracted sites may not have direct vendor oversight or may lack equally well-trained proctors. Due to this risk, it is critical that the vendor’s security manager takes measures to ensure all locations follow the established and agreed upon policies and procedures.

Available Test Designs within the Methodology

Because this is arguably the most established test delivery method, it also has the largest availability of test designs including advanced adaptive testing and it supports a wide variety of traditional and technology-enhanced items.

Pros and Cons (Security) of the Methodology

This is generally considered a secure method of testing but, like all channels discussed, it does have security issues that should be considered. The security concerns vary from the threats of decrypting files, reverse engineering, or collusion with test administrators. These tactics are typically used by organized entities that employ individuals to steal the exams with the intent of selling the test items or full exams to future test-takers. There are other types of security threats used by individuals, such as memorizing items and the use of technology to capture test items, but these are problems that are inherent in every test delivery method.

Pros and Cons (General Business) of the Methodology

While this approach offers the potential benefit of global distribution, it is not without risk. First, the assessment content is replicated and stored across a number of systems, including those of the assessment organization, the delivery vendor, and the test center. Second, the content is distributed across the globe. These facts mean that the content is only as secure as the weakest link. This requires that the assessment organization’s and the vendor’s Electronic Data Privacy (EDP) policies, processes,
and tools used to secure the exam content, must be scalable and consistently applied. Some of the most common methods of content protection include encrypted downloads of content on secured sites, strict proctoring and identification methods such as photo taking, ID checking, as well as security barriers to item harvesting threats such as hidden cameras, metal detection devices, and videoing of test sessions.

C) Computer-Based Test (CBT) Delivery (Application-Based)

General Description of the Methodology

Application-based test delivery involves computer-based testing over the internet and utilizes a desktop application to deliver exams on test-takers’ personal computers (PCs) and/or laptops with a controlled proctored environment of the sponsoring organization’s choosing. The exam is downloaded as an encrypted file that the test-taker cannot open until the assigned exam time. Once at the exam venue (e.g., school, training room, hotel) the proctor can issue a code to all the test-takers for that session to allow them to launch the exam in a locked down environment. Once the exam has been submitted, the test is uploaded back to a main server, scored and individual test-takers are given the result. The exam is then automatically wiped off the test-taker’s PC or laptop.

Pros and Cons (Security) of the Methodology

This is an appropriate delivery method for a program that wishes to conduct multiple testing events in different private locations (schools, offices, hotels, etc.) simultaneously, or nearly simultaneously, without the use of expensive server-based systems. However, security considerations apply because many of the same security risks exist with this methodology as exist with CBT server-based delivery model (see Table 2 for categories of security threats). This is not a very useful model for open enrollment or continuous testing (e.g., offered in traditional testing centers via server-based or internet-based delivery methods).

Pros and Cons (General Business) of the Methodology

This delivery method does not require significant investment in technology because it uses the test-takers’ devices to access the internet. It also allows for flexibility in the proctoring location and does not require massive bandwidth for one location as the exams can be downloaded and responses uploaded in a reasonable timeframe after the test administration is completed. This does require the test-taker to complete and submit the exam within the designated administration timeframe (even though the results are not being uploaded immediately) to ensure the test-taker does not continue to have access to the test content after the administration is concluded. This also allows large numbers of test takers to be tested at once in multiple locations without the risk of a single point of failure on the administration day.

Internet-Based Testing (IBT)

General Description of the Methodology

IBT delivery, in some ways, is similar to CBT server-based and/or application-based exam delivery. One major differentiator is that the exams are delivered through an internet browser in real time via a remote server on a password protected website. The IBT test delivery methodology is traditionally used for un-
proctored exams that test-takers can administer themselves from their own home or office. In recent years, however, new technology and enhanced administration processes also have enabled the secure, remote proctoring of internet delivered tests.

IBT lends itself to multiple delivery methods including:

- un-proctored examinations;
- test-sponsor proctored examinations (where the test sponsor manages/owns the computers and proctors exam administrations within its own facilities or designated location);
- third-party proctored examinations at testing centers; and
- remote-proctored examinations at a site of the individual test-taker’s choosing (e.g., home or study) using a webcam to monitor the test-taker.

IBT can be used to support various testing needs from post-course testing, distance education, delivery of practice tests, pilot (beta) testing, or administration of published high-stakes licensure and certification exams. IBT also can potentially expand the number of testing locations available to an assessment organization and provide flexible exam scheduling. Some remote proctoring is tied to a specific testing engine. Other uses are system agnostic and simply require that the test sponsor provide a launch code that a proctor can enter to start the test.

Features/Considerations of the Methodology in Regard to Security

IBT exams are centrally stored through a remote service/hosted software model and are delivered in real time. This type of security is similar to the way a bank delivers encrypted, centrally stored content to an ATM, for example. The most significant impact of this method is that with IBT the answer key and scoring logic is never downloaded to the testing computer or local server, but rather remains on a single, secure server that is always under the control of the test owner or distributor. In addition, IBT does not require downloading of the exam item pool prior to the actual delivery of the assessment.

There are several remote proctoring solutions available which vary in the level of security offered. The range of delivery options and monitoring all play a role in regards to security. Most systems have the capability to use a lock down browser during a test so that no other programs can be accessed by the test-taker during the test administration. Other options include: variable levels of authentication of the candidate; webcam monitoring; technology forensic applications to monitor key strokes, timing, and test-taker responses; and recording of testing sessions and real-time, remote monitoring during the session.

The cameras and/or the testing system itself may allow the proctor to occasionally view the test-taker's screen and to take a picture of the test-taker to verify the actual individual taking the test. Monitoring of the test-taker during the exam can be done live, with a proctor-to-test-taker ratio of up to around 20 test-takers per proctor, depending on the vendor. In other systems, the session can be recorded and reviewed at a later date. If the test is being monitored live, communication with test-taker during the exam can either be through live video and audio or via pre-programmed standard alerts, text chat or by phone.
Available Test Designs within the Methodology

IBT can deliver virtually all of the same item types and test designs that traditional CBT offers, although the options are dependent on the vendor's specific offerings. For instance, the availability to support CAT capability and rich item type (e.g. video) vary by testing vendor.

Pros and Cons (Security) of the Methodology

Pros
One of the benefits of IBT is the range of testing solutions enabled at different costs. Programs operating on tight budgets are typically able to offer at least some level of identification and security where previously they may have been unable to include any level of security or proctoring. IBT can also be used with testing centers to provide an alternate, secure testing model to traditional CBT (because the answer key is never downloaded to local servers). Remote proctoring options available vary in the level of security offered. For exams that are historically un-proctored, this results in a cost effective way to increase exam security. With remote proctoring through a webcam, potential, inappropriate interactions between a test-taker and a proctor also are reduced or removed.

There are numerous ways that the sponsor or the vendor can authenticate the test-taker as the person who initially signed up for the test. Keystroke analysis uses a test-taker’s typing pattern to confirm the individual’s identity. Fingerprint technology and facial recognition can also be used, but some organizations have raised privacy concerns about these technologies and the equipment for performing them can be costly. Some programs require the test-taker login to a company’s learning management system; others require the test-taker to provide a photo ID or match the test-taker to a photo previously provided by the individual to the test sponsor at registration. Others also scan photo IDs through a kiosk where the exam is delivered in a remote proctored environment. Another popular method is to ask personal identification questions to verify a test-taker in order to provide access to the test-taker’s score report, or to identify the test-taker prior to (and sometimes during) the assessment. Finally, some programs place a photo of the test-taker on the exam.

Cons
Like any test delivery channel, the IBT option has security challenges. These include test-taker authentication and potential system vulnerabilities including locking down the browser and back end hacking. The greatest concern with this methodology (when partnered with remote proctoring) focuses on test-taker authentication so that an organization is able to confirm the identity of the test-taker prior to the start of the exam. Service Providers (vendors) have developed numerous ways (described above) for test-taker identification. The assessment organization must evaluate the available options and decide if the identified processes and procedures are sufficient as compared to live proctor verification in a traditional channel delivery model. A second concern with the IBT channel and remote proctoring revolves around the fact that different types of webcams may not be sufficient to capture all potential instances of cheating. Even when checking the full room it may be possible for the candidate to receive some assistance off camera during a remote administration.

Based on the concerns discussed, it is important for the assessment organization to verify if the IBT vendor offers a system lockdown browser ways to address backend hacker risks, and any real-time monitoring of exam delivery (both from the examinee and backend delivery standpoint).
Pros and Cons (General Business) of the Methodology

Pros
One potential benefit of IBT is the ability to use multiple delivery channels from one system. IBT enables test delivery in a variety of settings: at traditional testing centers for test-takers who live near one; remote proctoring to individual locations for a test-taker who has a fast internet connection and a quiet place to test; and to large groups at conferences and tradeshows. IBT enables world-wide exam delivery as these centers can be set up to meet minimum PC and internet connectivity standards.

Cons
One primary issue with IBT is connectivity. A reliable and preferably fast internet connection is required. This is not always available in all locations but as the internet matures, so does reliability. If there is an interruption in internet connectivity, then the test is typically paused and can be re-started where the test-taker left off, the clock is stopped and re-started automatically with any lost time taken into account so the test-taker is not penalized. In addition, vendors may offer different levels of support. Some may bundle test publishing and delivery while others may provide remote proctoring only, requiring the test sponsor to engage with multiple providers.

Section 4: Practices to Enhance Security

As security issues continue to grow, so do the options for enhancing security and reducing cheating and test theft. There are many aspects of a testing program that impact the security of the assessment and the intellectual property of the program. For example, the development of robust item banks, the creation of multiple test forms per administration, the random presentation of forms to test-takers, and the establishment of re-take rules are just a few assessment decisions that have significant impact of the security of an assessment program. The Operational Best Practices for Statewide Large-Scale Assessment Programs (authored by the Council of Chief State School Offices and the Association of Test Publishers) is a resource available that outlines key considerations throughout the test development and administration process, as well as for securing scores/reports.

Though many activities throughout the entire assessment lifecycle have an impact on the security of the assessment, there are some processes and procedures available to assessment programs that are specifically designed to enhance the security of intellectual property. The following section describes methodologies currently in practice that are designed to help reduce and detect issues with cheating and theft of assessment content. Each methodology is described in the context of the models and delivery channels in which they can be used.

The following methodologies are presented in this section:

- Data Forensics-Fraud Detection
- Randomization and Presentation of Items
- Item Pool Flooding
- Continuous Publication
- CBT/IBT blended Delivery Model
- Answer Key Holds/Delayed Scoring
- Fraud Detection Items
- Results Hold
Some of these approaches are not mutually exclusive and multiple methodologies can be used individually or together to further increase test security and should be customized depending on the needs of the assessment program and the nature of the security threats.

**Data Forensics - Fraud Detection**

Many testing programs have implemented data forensic methods to detect and investigate exam fraud and to enforce program policies. Data forensics is the use of statistical analysis to identify potential exam security violations. This analysis, when conducted in the regular course of monitoring test administrations or conducted on exam scores, and performed by knowledgeable professionals, can provide information for measuring and managing security risks that may be adversely affecting exam integrity specifically, and the assessment program generally. Data forensics is a tool that allows assessment programs to effectively detect potential exam security violations, providing objective and legally defensible evaluations of the probability of exam fraud.

**Analytical Tools/Services**

In the past few years, a number of data forensic methods and tools have been developed. The type of data forensic analysis that an organization may choose to use depends on its specific needs and may vary by exam model and test delivery channel. Regardless of the type of analysis, the intent is the same – to detect aberrant test behavior by identifying results that are statistically abnormal for a given exam across the entire population of test-takers. Aberrant results indicate potential cheating and/or other issues with test security that can then become part of a larger investigation into potential unauthorized or illegal behavior. However, it is recommended that data forensics should not be used in isolation to determine fraudulent or inappropriate activity.

Some types of data forensics may apply to a variety of programs; others are specific to an organization’s business rules and practices. An example of a common fraud detection rule that may be used to flag aberrant behavior is based on the length of time to complete an exam combined with the overall exam score. For instance, if a CBT or IBT delivered exam is published as a 90-minute exam and a test-taker completes the exam in less than X minutes (e.g., 15 minutes) with a score above Y (e.g., 95%), this combined result would be flagged for review due to aberrant results.

Some programs have developed their own data forensic tools and others contract with vendors who offer software and services in this area. Some vendors also sell data forensic analysis software packages, which may provide a more cost effective solution for some organizations. When implementing data forensics to enable enforcement against exam security violations, the organization must decide which testing behaviors it will measure – such as a fast test time with high score as noted above, extremely similar exam responses patterns to fraud detection items (see below), etc. In addition, the organization must set its “fraud detection thresholds” based on individual program goals and business needs. For example, if using a fast test time with high score category, the organization must define what a “fast test time” is for its tests, and what it will consider as a “high score.” Setting program fraud detection thresholds is critical to the ability to enforce against exam security violations objectively and consistently.
Data Forensics can be performed at various times during the examination process. This section addresses several common options.

1. After the Exam/After Scoring

   One approach is to perform data forensic analyses on exam results after exam administration and after a score and pass/fail decision has been communicated to the test-taker. In the event aberrant scores are detected post exam, a decision has to be made whether to revoke the “pass” decision given to the test-taker. At that point, the organization implements its policy regarding the handling of indeterminate or aberrant scores. This approach is being used widely, and is well supported in the assessment industry. However, in this post-scoring methodology, test-takers with prior access to exam content or assistance during testing initially receive a passing score and any associated designation/certification provided by the assessment organization, which can create a negative impact on the integrity of the assessment program if those scores and designations are later revoked. In addition, enforcement of exam security policies has to be carried out AFTER the designation is awarded – which often makes this approach a difficult process. Using this methodology also requires the organization to have a policy in place regarding revocation of scores and must ensure that the policy is implemented fairly and without discrimination. The revocation of scores and associated designations adds additional legal considerations.

2. After the Exam and Scoring/Before Reporting Scores

   Tools are now available that allow for automated analysis of exam results with very quick identification of aberrant results (i.e., potential exam fraud). Some organizations are leveraging this technology to perform data forensics after the exam is administered but BEFORE the exam results are validated and presented to the test-taker. If the test-taker’s exam results are identified as involving potential exam fraud, the test-taker does not receive a score on the exam until the results are reviewed and confirmed. In effect, the criteria to pass an exam become not only achieving the cut score, but also passing the fraud detection threshold. This approach represents a significant enhancement in the ability to protect the integrity of assessment programs and the certifications/credentials that are awarded. The pros and cons of this methodology are discussed in more detail below, under the Result Hold section.

3. During the Exam

   A newer methodology some organizations are exploring is the concept of “real time” data forensics. In this scenario, analysis would be performed on each question, as the test-taker works his/her way through the exam. At any point, if the fraud detection threshold is reached, the exam could be stopped and a failing score issued to the test-taker. This is an emerging area of investigation and represents a potential advancement in exam security processes.

Randomized (Stochastic) Presentation

General Description of the Design

Randomization is a tool that can be applied at various levels within an assessment presentation, from the of the exam presentation to the individual item options.
Generally speaking (assuming such general factors including a robust item bank), the more randomization applied to an assessment, the more unique the exam will be for each test-taker. Uniqueness in the test-taker experience is important, as the more unique a test experience is for each test-taker, the more protected a program is from the impact of a single security breach. A second benefit that is derived from a unique testing environment is that if a security breach does occur, it is typically easier to identify the source of the breach (as very few, if any, test-takers see the same set of items in the same order).

Pros and Cons of the Methodology (Security)

Pros
The more randomization introduced into an assessment, the more the item presentation is essentially unique for each test-taker, thereby enhancing the security of the exam process. Answer key order cannot be memorized independently from the content, because there is no consistent pattern to the sequence in which items are administered to any given examinee. Thus, there are no predictable sequences or patterns with this stochastic process.

Cons
Any time that items are presented to examinees in a differing order, there is some potential for item order effects. Organizations should be aware of this potential and analyze the data from randomized presentation of test items both to evaluate the effectiveness of the stochastic process, as well as to determine whether there are any effects of item ordering. An ongoing process of monitoring exams for evidence of discrepancies that may be related to the methodology used during exam administration is recommended (particularly across other types of examinations). To date, however, random presentation of content (during the exam) generally does not seem to be excessively disruptive to a test-taker and/or to have any significant impact on the ability to demonstrate proficiency.

Pros and Cons of the Methodology (General Business)

Pros
Randomized presentation of test items is widely supported as an option in computer-based test administration software. It does not generally require larger numbers of items or more sophisticated measurement models to implement. In addition, the resources required for implementation are fewer than many of the other test security measures. The randomized presentation of test items required so
that the sequence is unique for each individual also tends to be easily understood by examinees as well as other key constituencies.

Cons
There is no single, correct way to implement a randomized presentation process into an examination. Programming a randomized presentation of test items involves a number of choices, decisions, and options for the organization. There is considerable data available on the implications of the different approaches in the available literature; information on this can be provided by the program’s psychometrician with expertise in this area. Depending on the choices made and the options employed, time and cost are likely to be added to the development cycle. If an organization is relying on a random presentation feature that is built into test administration software, it should independently evaluate the software to ensure its proper functioning – which should also be considered when determining the schedule and the budget for the process.

Item Pool Flooding

General Description of the Design

The purpose of item pool flooding is to overwhelm test-takers who want to circumvent the process of studying and/or gaining real world experience by purchasing and memorizing unauthorized exam materials from braindump sites. The phrase ‘braindump’ refers to copyrighted test items and/or answers which are unlawfully obtained by test-takers who memorize the information and then after the administration ‘dump’ (document) the information for use by others, often in an organized way. Braindumps also reference collections of exam items that have been stolen from high-stakes assessments which may then be illegally sold to other test-takers as a preparation tool. These types of braindumps contain actual exam questions and answers. Brain dump sites usually constitute theft because the information acquired was generally taken from material owned by the organization that must remain secure (i.e., information that is copyrighted or registered as trade secrets).

There are a numerous approaches to item pool flooding depending on how transparent an organization or company wants to be. One approach is to publish questions that align to the subject of the exam, purposely deceiving test-takers by publishing many more items than the number actually on the assessment (e.g., double, triple, or quadruple), so that individuals trying to memorize those items will become mentally fatigued and unable to memorize all questions, including the intended exam questions. For example, if a test sponsor intends to conduct an exam testing “Administration of Software 2012,” then it would publish retired questions from “Administration of Software” 2009, 2010, and 2011 and additional items not on the actual exam.

Another approach is to publish questions that don’t align to the subject of the exam and try to distract or stump test-takers by forcing them to memorize irrelevant content. Of course, organizations may also use this approach in the hope that candidates will throw away the braindump and actually train, study, and/or prepare for an exam. The flooding process may also consist of publishing legacy/retired decoy exams besides items intended to be on the exam that is delivered and scored. For example, if the sponsor intends to administer an exam testing the Administration of Software 2012, then it would also publish an exam with completely different content (i.e., the decoy exam) under the exam name of Administration of Software 2012. However, the test delivery vendor will be instructed to only deliver questions on the exam identified as testing Administration of Software 2012. Test-takers taking the exam will only see
questions testing Administration of Software 2012. But, test-takers who purchase Administration of Software 2012 from a braindump will see double the amount of questions (since the stolen braindump file includes a file with questions from Administration of Software 2012 along with the decoy items). In this way, test-takers attempting to memorize a braindump will either have to learn irrelevant content or they will have to decipher which questions actually belong on the exam that they purchased. The decoy questions will hopefully help the test-takers realize that they have purchased “bad exam data”, which will make it much more challenging to figure out which questions to memorize and which questions are completely irrelevant, or decoy questions.

It’s important to ensure that flooding is invisible to the braindump companies by “sprinkling” flooded items throughout the intended item bank. Within the exam export, the decoy item IDs should be revised to look similar to the intended, authentic item IDs. The purpose of making the decoy item ID’s similar but different from the authentic items is to ensure the braindumps aren’t able to identify the flooded decoy items within the file. It is critical, however, that the certification team and the test delivery vendor can identify the flooded decoy items to ensure that they decoy items are not delivered during the live assessment.

Finally, if the organization is going to use an exam of translated (i.e., transadapted) items, it should use a decoy item bank from a prior version of an exam that has already been retired and translated. If a program localizes its exam content, program administrators must ensure that the decoy item bank is also translated. If a program selects a decoy item bank that hasn’t been translated, program administrators will need to account for additional time and cost to localize the decoy items so that it is not obvious to test-takers who purchase braindumps that two or more different exams/items were published as one exam.

Pros and Cons of the Methodology

**Pros**
This process helps to discourage the use of unauthorized content to prepare for an assessment, and makes it more difficult for braindump sites to guarantee the usefulness of the materials being sold. If the illegal sites cannot provide such a guarantee, some candidates will be less likely to purchase the illegal materials.

**Cons**
It is very challenging, if not impossible, to accurately measure the success of item pool flooding. However, the limited time and nominal cost to implement this approach and the potential positive outcome often outweighs the limitations in the precise metric of success. Item pool flooding is an evolving approach. Industry experts continue to evaluate the process to augment the approach and to better measure the effectiveness of item pool flooding.

Continuous Publication

**General Description of the Design**

Mitigating the degradation of item performance associated with exposure is a critical element of managing a healthy and cost-effective assessment program. The number of times an item is seen has a direct impact on its effectiveness in discriminating between qualified and non-qualified test-takers. This
degradation can be caused by malicious and intentional acts, such as using unauthorized content on a braindump, or by more innocent actions (e.g., taking an exam multiple times and thereby seeing the same item multiple times). In either case, the test-taker obtains the opportunity to enhance performance on the assessment without any enhancement to his/her actual knowledge and skills.

One method used by some organizations to maintain the performance of items and reduce the impact of item exposure is continuous publishing. Continuous publishing is the process of publishing unique forms at frequent intervals (e.g. monthly, quarterly). By publishing unique forms more frequently, the organization has the ability to change out questions rapidly and without the test-taker’s knowledge. As a result, this limits the likelihood that the test-taker will know which items will actually show up on the next assessment and reduces the performance advantage that may be obtained by test-taker attempting to use inappropriate preparation materials or taking the exam multiple times.

The assessment should have an item pool with known psychometric properties, which is sufficiently large to allow for the frequent publication of unique and balanced forms containing minimal overlap with prior, current, or future forms so that each publication is equivalent.

There are two additional steps an organization can take to enhance the benefits of continuous publishing:

- First, it should align the number of forms published and the republication schedule with the retake policy. This helps to ensure that a test-taker will not see the same form regardless of how many times they take the exam.
- Second, it should align the development and seeding of content such that every form republished has some amount of new content on the form. By continuously seeding small groupings of items in every republication, an organization can have confidence that test-takers are unlikely to have used inappropriate preparation materials. Additionally, if an organization uses this approach consistently, not only will it be difficult for a test-taker to memorize the ever increasing pool of items, but it will be increasingly difficult for the test-taker to identify items that are not actually going to be scored.
If the test-taker has used inappropriate sources to study in a continuous publication environment, data forensics can more easily identify the situation because the test-taker will perform much better on the older items than the newly published content.

Pros and Cons of the Methodology

Pros

A key benefit to continuous publishing is the fact that it represents a proactive step by the organization. By implementing this method, the organization has the ability to refresh content before the impact of item exposure is evident. Continuous publishing also enhances the ability to detect aberrant test-takers and over-exposed items, reducing the proportion of test-takers scoring well on the exam and poorly on pilot items, as well as reducing the proportion of test-takers obtaining very high scores in short periods of time. If small numbers of forms are published often, this approach can also be used to limit the amount of content in the channel at any given time, which reduces the impact on the assessment program in the event of a security breach. Finally, this approach has the ability to be very disruptive to operators of braindumps and other IP pirates. This outcome is based on the fact that only a small amount of the overall item pool is in the channel at any given time, and the content is changing frequently. Because of this, pirates must constantly monitor these exams to ensure they obtain all of the relevant content. If they fail to do so, their customers will find the information to be ineffective, reducing the perceived value of the content they offer.

Cons

When considering this publication strategy, a program needs to consider the likelihood that it will need to increase investment levels due to the fact that the effectiveness of this publication strategy is directly related to the size of the item pool. The larger the item pool, the greater the likelihood that this approach will bolster the performance and “life expectancy” of the items. Additional funds, time, and resources are also likely to be required for localization, psychometric analysis, data forensics, publication, and enforcement. Continuous publication can also create some psychometric challenges including in areas such as: multidimensionality, validity, and misleading item statistics. Based on these factors, this approach is typically appropriate for high and/or high-stakes assessments.

Answer Key Hold

General Description of the Design

Many programs offering CBT provide immediate scores (pass/fail, scaled score, etc.) to test-takers at the end of the exam. As noted above, with CBT server-based delivery, the exam content files distributed to test center servers include both the exam items and answer keys. In the event of a security breach, all test content is readily available to steal, distribute and sell on the web. By contrast, with the “answer key hold” approach, in the event of theft or compromise of the exam files, the answers would not be available with the exam items. This step poses a significant hindrance to those who make it their business to steal exams for sale. Without the answer key, it could potentially delay the process of disseminating the exams for sale and could ultimately be a deterrent to purchases who cannot buy the answers.

Inherent with this approach is that test-takers will not be given an immediate score at the end of exam. The exam scores will be delayed and test-takers only receive their results on a delayed basis after exam
administration. This delay also allows the organization time to conduct analyses of test results to detect aberrant scores and response patterns prior to communicating the exam score to the test-taker.

Pros and Cons of Answer Key Hold-Delayed Scoring

Pros
Since protection of exam content is of utmost concern to testing organizations, this approach provides a solution which may slow the dissemination of exam content that is posted for sale and distribution, or even render it meaningless. If braindump sites want to sell complete exams with answers, they would need to hire individuals to provide the correct answers on all new exams that are published. Although the exams could be still be posted for sale without the correct answers, many prospective buyers will not be as interested and the sites that offer exams with a “guarantee to pass” as well as require additional investment if they want to continue to sell exams with the answers.

Cons
Since CBT offers the capability of providing immediate scores at the end of the exam, many testing programs have established this as a standard. To change the policy and offer delayed scoring constitutes a paradigm shift for those programs that may require major adjustments and also change the perception of the exam for test-takers.

The feasibility of this approach should be investigated by an organization together with its test delivery vendor. Instituting the new policy will require a comprehensive implementation plan including education and “marketing” of the new approach. Although there is a cost to the answer key hold method, it may be offset by the benefits of reducing the impact of stolen exams.

CBT/IBT Blended Delivery Model

General Description of the Design

Most international testing organizations are using a CBT delivery model. Because of the global audience, they need the reach of the vast test center networks that are available. However, using the network of test centers increases the potential for a security breach when the exam content is distributed to local test center servers. When test centers servers are compromised, the risk is that all the testing organization’s test files are stolen and the integrity of the program is negatively impacted. To address this issue, some organizations have pursued a blended delivery model to enhance the security of exam test items and answers. This process involves the administration of the exam through the test center network using an internet-based testing (IBT) engine to deliver the exam. With IBT, the test items and answer key are held in at least one secure remote location and are inaccessible to test center personnel.

Pros and Cons of CBT/IBT Blended Delivery Model

Pros
The approach provides a blended model that meets the needs of international testing organizations that require a broad testing network to deliver exams to their audience. At the same time, it reduces the risk of security breaches that may occur at testing centers.
Cons
This model will require review and test delivery vendor acceptance and support to plan and coordinate the program implementation. In addition, there continues to be an internet connectivity concern when IBT is utilized. A reliable and preferably fast internet connection is required, which is not always available in all locations; but as the internet matures, so does reliable connectivity.

Fraud Detection Items

One of the primary objectives of any exam security program is to prevent test-takers who do not meet the requirements— but have passed the exam by studying the live exam questions (with or without the actual answers) – from being certified or credentialed. Building exams that can reliably identify when a test-taker has had prior access to the exam content is a strong tool in protecting the program integrity by allowing a program to take appropriate action. There are two types of test items that organizations embed in their tests and use as part of their overall, comprehensive exam security process. These item types, so-called Trojan Horse and Embedded Verification Test (EVT) items, identify the statistical probability that a test-taker has had prior access to exam content AND may not truly understand the content of the exam. Each item type is focused on different issues, and can be used for multiple purposes.

Trojan Horse (Investigative) Item

General Description of the Design

The Trojan Horse (TH) fraud detection item methodology was developed to identify when a test-taker has memorized a test answer key AND does not truly understand the content of the test. These items can also be used to determine when the test answer key has been exposed by identifying where, when and which test-taker first shows knowledge of the answer key.

To develop a TH item, the test developer includes in the test a statistically significant number of “extra”, unscored items (typically six are enough) that are:

- Very easy relative to the content of the exam (i.e. qualified and below minimally qualified candidates would be expected to get the question correct);
- Have both viable, plausible distracters and one or more definitely correct answer(s); and
- Have an obviously incorrect distracter purposely keyed as the correct answer.

To illustrate the TH method, consider the question “What does two plus two equal?” The answer options would be A) 2, B) 4, C) 5, D) 6. In this case, the answer is purposely miskeyed as C) 5. Anyone with basic understanding of the content would select option B. Consequently, a test-taker who selects option C provides a flag to the program sponsor that the test-taker does not have command of the material. The level of relative difficulty of the TH item has been established during test development as very easy and that anyone who does have basic knowledge and understanding of the topic would know the correct answer was option B.
The assumption is that test-takers who memorize the answer key AND do not truly understand the content of the test will

1) select the purposely miskeyed Trojan answer because they do not truly understand very basic concepts; and
2) will perform well on the test overall because they had prior access to the answer key.

Statistical probabilities can be assigned to the combination of the number of TH items answered with the miskeyed answer option and the overall score on the test to determine the probability of exam security violation. These probabilities are put into a matrix and the organization assigns its “acceptable” threshold for enforcement. For example, a typical probability matrix might look like this (see graph on the following page):

In this example, the organization will not act on any incident with greater than 1 in 100,000 chances the result was NOT due to exam security violation (green zone). It may send warning letters if the probability falls between 1 in 100,000 and 1 in 1 million there is no exam security violation (yellow zone). Finally, it could potentially revoke exam results if the probability this is NOT an exam security violation is 1 in 1 million or below (red zone). This example shows a very conservative approach that errs on the side of false negatives - not enforcing against test-takers who have truly had prior access to exam content. Organizations can determine exactly what level they are comfortable with for their individual program and determine appropriate consequences based on the laws and business rules established for the specific industry and program.
Pros and Cons of the Trojan Horse Item

Pros
Trojan Horse items are very reliable and accurate in detecting if a test-taker has had prior access to the exam content AND is not truly knowledgeable about the exam content; because the TH items are so easy they work to identify candidates that would have failed without that prior access (i.e. the truly unqualified individual). With TH identifications the organization is able to focus on the highest risks—test-takers who passed the test but do not have the knowledge and skills that a passing score imply. In addition, TH items can be used to detect when an exam answer key has been compromised. When and if the organization starts to see test-takers responding to TH items as purposely miskeyed then it has a strong indication that the exam answer key is compromised and can take appropriate action.

Since TH items are not scored and not many of them are required to use the methodology, TH items have little to no impact on a test-takers assessment experience. They are simple to develop because the objective is to make them easy to understand and respond to for a minimally qualified/knowledgeable test-taker. The Trojan Horse methodology is also very easy to explain to stakeholders and test-takers that may question the reliability of the approach. People understand the low probability that a truly qualified test-taker would select an incorrect answer on a simple question multiple times, and instead select the answer identified as correct on the answer key.

Finally, the quantifiable nature of the TH items lends itself to automation. The entire process of identifying TH violations and the appropriate action to take can be automated and built into the scoring process BEFORE a test-taker receives a passing score. With the TH approach an organization must establish not only a cut score required to pass a test but also a TH threshold required. Test-takers who have had prior access to the test answer key and do not truly understand the test content will not receive a pass on their test. See additional discussion of this in the Result Hold section below.

Cons
There are some challenges with Trojan Horse items. The most critical problem is that they are totally ineffective if the content available to test-takers reflects the actual correct answer versus the purposely miskeyed, correct answer. This can happen if illegal braindump, test prep sites research questions and use subject-matter experts (SME's) to develop their own answer key, or to validate stolen answer keys. Similarly, they cannot distinguish if a test-taker only has had prior access to the test questions and not the answer key. In addition, this methodology will not prevent a test-taker from passing the exam unless the automated techniques described in the Result Hold section below are implemented.

Partial Exam Refresh with New vs. Old Item Analysis

General Description of the Design

A relatively new process for fraud detection entails a partial refresh of the exam, with subsequent analysis of performance on the new items vs. the old (or pre-existing) items within an exam. This process (“New vs. Old”) identifies test-takers who have prior access to the test questions, but who are not truly knowledgeable about the test content. New vs. Old analysis is effective in situations where a test-taker has purchased a complete exam from a website, or simply collected questions from other,
previous test-takers. As with TH items (see above), New vs. Old analyses are easy to understand and relatively simple to implement.

In New vs. Old analysis several new exam items are “swapped” into the exam (replacing other old or pre-existing items) after the exam has been live for a given period of time. Thus, the new items become a “test within a test” that can identify if a test-taker has been exposed to the older items in the exam, but not the new items that were added after the exam may have become compromised. The assumption is that test-takers exposed to the original exam items, who do not truly understand the exam content, will score well on the old items they have seen, but poorly on the new items to which they have had no previous exposure.

To implement New vs. Old analysis, a pool of extra items is created during the exam development process. The same standardized and defensible development methodologies should be used when developing these items as when developing the standard items for the assessment programs. This is critical to the success of the New vs. Old analysis as the new items are regular test items and are included as part of the test-taker’s score. When it is discovered the exam is compromised, or on a regular schedule such as every 8 to 12 weeks, the exam development team does a limited re-publication of the exam, replacing ten or more “old” items with the same number of equivalent new items. It does not matter where in the exam these items are replaced, but the replacement process will need to ensure that the updated exam forms are equivalent to those previously administered.

Test-takers who have prepared using stolen exam questions, and who do not understand the content of the exam, will perform better on the old items than on the new items. This difference in performance can be quantified and instances with large score differences can be flagged for further review of potential exam security violations. As with the TH items, individual organizations can set their own criteria for determining the statistical threshold that defines an exam security violation.

Pros and Cons of the New vs. Old Analysis

Pros
A key advantage of the New vs. Old analysis (as with TH) is that it allows an organization to identify unqualified test-takers. If a test taker truly understands the content of the exam, s/he will perform consistently across the old and new items and therefore will not be “flagged.” As a result, those test-takers that are the highest risk to the assessment program – those who pass an examination although they are not truly qualified or knowledgeable – are the ones identified for further investigation and potential action or enforcement.

The New vs. Old analysis works across a broad range of content theft and illegal access to content. This method does not require that the answer key be available to the test-taker, nor does it require that the questions come from theft of the entire published item bank. It simply requires that a test-taker have been exposed to a majority of the items on the exam, whether through purchase on the web or “farming” from friends, co-workers, forums, etc.

New vs. Old analysis is relatively inexpensive to implement. It requires:

1) A pool of items that should be developed at the same time the standard exam is developed;
2) Regular re-publishing of a small part of the exam – to “swap out” old questions with new questions; and
3) Qualified statistical analysis of results to identify exams with inappropriate differences in performance between old items and new items.

Depending on how many new items are included in the exam, as well as the number and composition of different forms, additional analyses may be required to confirm form equivalency and a valid cut score with the new items.

New vs. Old analysis provides intelligence on the vulnerability of exams. If there are “hits” after the new items are inserted, an assessment program will know that their exam content is compromised. In addition, New vs. Old analysis that initially get hits but over time stops showing significant differences in performance of the old versus new items, suggests that the new items themselves have been compromised (in addition to the older exam items), resulting in a need to go perform the “swap” process again, or update the entire exam.

Finally, the quantifiable nature of the New vs. Old analysis (as with TH) lends itself to automation. The entire process of using this process to identify violations automated and built into the scoring process BEFORE a test-taker receives a passing score. With the New vs. Old analysis an organization must establish not only a cut score required to pass a test, but also a New vs. Old flagging threshold. Test-takers who have had prior access to the test questions and do not truly understand the test content will not receive a passing score on their test. See additional discussion of this in the Result Hold section below.

Cons
Utilizing New vs. Old analysis requires a regular, proactive program of refreshing and republishing exams with new items as the existing items become compromised. New items are only effective until these items become publicly available or are shared broadly with the assessment audience. There must be a pool of new items that are equivalent to the exam content. To ensure equivalency the items should be developed as part of the overall exam development process. In addition, some additional analysis may be required to confirm cut score and form equivalency after the new items are “swapped in”. Finally, this methodology, if only using a small number of new items, will not always prevent a test taker from passing the exam unless the automated techniques described in the Results Hold section below are implemented.

Result Hold – Data Forensics before Final Score Awarded

General Description of the Design

Assessment programs that use technology-based testing often provide immediate pass/fail information to test takers. Organizations that conduct data forensic analysis of exam results after scores are awarded to the test-taker may have a challenge enforcing exam security policies when aberrant scores are detected. In this case, assessment programs must “revoke” the scores and/or the designations that were previously awarded when security violations are discovered.

In recent years, some assessment programs have addressed this issue by “holding” exam results until after the data forensic analysis is completed. Any test-takers with identified exam security violations do not receive a pass, so the assessment program does not grant any designations based on those results. In
effect, the requirement to pass an exam becomes not only meeting the exam cut score, but a test-taker must also achieve the fraud detection threshold developed by the assessment program.

**Pros and Cons of Result Hold – Data Forensics before Final Score Awarded**

**Pros**
Holding the exam result until after data forensics are performed allows an assessment program time to verify that a test-taker has not violated exam security policies before receiving a passing score, and consequently prevent awarding a score or certification to an unqualified test-taker. This avoids issues with having to retract a “pass score” that was given or designations awarded as well as saving time and effort required to revoke exam results after the fact.

**Cons**
Holding exam results until after data forensics are performed is relatively new and will require about a new definition of what is required to pass an exam. Instead of simply achieving the cut score, with this approach the candidate must also achieve the Fraud Detection threshold. Programs must also implement fraud detection items/methodologies and policies that allow for detection of fraud and provide guidance around how to enforce against the identified security violation. Finally, test-takers who are accustomed to getting immediate feedback on their exam will have to wait for final confirmation of a pass.

**Provisional Scoring**

**General Description of the Design**

The provisional scoring approach is similar in many respects to the Result Hold approach noted above but offers some differences that may be a preferred approach for some programs. In this case, test-takers will receive a score at the end of the exam. However, test-takers are notified in advance and again at the time of the exam that the scores they will receive are “provisional” and final scores only will be provided after the exam administration session, after the completion of multiple administrations conducted at or near the same time, or even after the organization has evaluated the fraud data and authorized final scores. The specific timeline and methods to communicate with the test-takers is program specific. As with Result Hold, it allows time for the organization to conduct data forensics of the results prior to granting a designation or certificate and issues with having to “revoke” these at a later date.

One of the benefits of this approach is that it sends a message to test-takers, in effect, that “things have changed.” Because the program has implemented new measures to review exam results and put test-takers on notice that the exam results will undergo further analyses and since they will not know what measures will be taken, it may be a deterrent for cheating. A key consideration is that this new approach can be implemented without any significant changes to the program.

**Cons of Provisional Scoring**

**Pros**
Organizations may find that implementing this new policy a more feasible alternative to implement versus Results Hold (i.e., giving no score at the end of the exam). Test-takers may be less resistant than receiving no score at the end of the exam. It also communicates a message to candidates that the organization will be conducting further analyses of the results and they need to wait for the final results.
Cons
Provisional Scoring will require that some assessment candidate audiences be “re-educated” that they will not be receiving a final score at the end of the exam. In addition, test-takers who are used to getting immediate feedback on their exam will have to wait for final confirmation of a pass.

Conclusion

Security methodologies continue to evolve and advance on a daily basis. Designs and methodologies outlined within this document should be reviewed and considered carefully prior to development and implementation. Regardless of the methodologies that are employed within assessment organizations, security policies and program activities should be in place to continue protect and monitor the security of the program’s intellectual property.