Conference on Test Security

THE 10TH ANNUAL

CONFERENCE ON TEST SECURITY

OCT 6-7, NOV 3-4, DEC 1-2, 2021

VIRTUAL

WELCOME TO COTS 2021!

Dear Colleague,

Welcome to the 10th Annual Conference on Test Security – the only event dedicated entirely to test security. It is our pleasure to have you join us virtually to discuss methods and best practices in and around test security. In recent years, cheating scandals have added to an already increased skepticism towards the testing industry. Now, as the pandemic has upended many of the standard practices in the spheres of education and assessment, public confidence in test scores continues to wane as test takers have had new opportunities to artificially inflate their exam scores. This year's program includes many sessions that demonstrate how testing organizations have addressed security concerns associated with the abrupt changes in testing conditions necessitated by the pandemic. While much has changed since early 2020, the need to protect the validity of test scores and brand integrity remains the same.

We have an exciting and diverse program lined up across three sets of dates: October 6-7, November 3-4, and December 1-2. Each day consists of four sessions and includes a mix of panel presentations, facilitated roundtables, coordinated symposia, demonstrations, or standard sessions. In addition to the content of the conference presentations, one of the biggest draws to COTS over the years has been the opportunity to connect with and learn from colleagues in the test security sphere. We hope this year's conference will afford a similar experience as attendees interact with each other through the chat function, Q&A during sessions, and informal conversation in the Virtual Lunch Room.

Finally, a special thank you to our sponsors for making this year’s conference possible. We appreciate their unyielding support. It is because of them that we have been able to gather together – both in-person and virtually – for a decade now to forward our joint goal to protect the validity of test results.

We hope that you enjoy your time at the 2021 virtual Conference on Test Security.

Warmest regards,

COTS 2021 Executive Committee

A SPECIAL THANKS TO OUR SPONSORS

HOST

ETS®

CO-HOSTS

ascend LEARNING

caveon™
Test Security

MEASURE LEARNING

proctorU yardstick

Pearson VUE

A SPECIAL THANKS TO OUR SPONSORS

FRIENDS

duolingo
english test

WEDNESDAY, OCT 6
10:00 A.M. – 2:15 P.M. EDT

10:00 - 10:45

Security Implications for Multi-Modal Delivery

Camille Thompson, College Board | Ray Nicosia, ETS | Faisel Alam, Law School Admission Council | Rachel Schoenig, Cornerstone Strategies

Many testing programs today are embracing multi-modal delivery, providing test takers the option of testing either in test centers or at remote locations such as homes or offices. Unfortunately for practitioners, exam security policies used in one modality may be less appropriate for use in another. Further, changes to deliver modes may raise questions from test takers and other stakeholders regarding exam security and fairness. How do testing programs re-evaluate long-standing security policies designed to protect test content and score validity to adapt to a multi-modal world? What methods can programs use to select and implement policies and procedures that are right for test takers in both modalities? Join testing professionals who have successfully navigated multi-modal delivery to discuss what you can do to help your program protect its exams across different modalities. Together, we will explore security implications of dual-modal delivery and considerations for ensuring new policies are appropriately implemented and expectations are aligned across the testing ecosystem. Join us for an engaging and information session that explores key considerations in crafting multi-modal policies.

10:45 - 11:00

Break

11:00 - 12:00

Standard Presentations I

Reducing Score Bias Through Real-Time Rerouting of Examinees with Anomalous Responses: Impact of Relaxing the Assumption of Identical Forms

Merve Sarac, University of Wisconsin-Madison | James Wollack, University of Wisconsin-Madison

We leverage a method for real-time anomaly detection within the context of item preknowledge for detecting anomalous responders early in real-time to reroute them to believed-to-be secure items for the rest of the test. We assume that rerouting examinees with anomalous responses to secure content once flagged in real-time would reduce the bias in their ability estimates at the end of test administration. Our motivation behind this operational decision is to obtain more accurate ability estimates of examinees with preknowledge rather than waiting for post-exam forensics and taking actions about their scores and possible score uses after the test administration. In an extensive simulation study, we found that the accuracy of ability estimates of EWP noticeably improved when rerouted, while it stayed the same for honest examinees. Our simulation results further showed that rerouting in the context of a credentialing test noticeably reduced the bias in the pass rates, with earlier rerouting producing the largest reduction.

A Practical Approximation of Response Similarity

Russell Smith, Alpine Testing Solutions

This presentation will describe a straight-forward proxy for response similarity index analyses based on reasonable estimates of expected matching responses, and discuss the advantages and disadvantages of each approach. Response similarity index analyses can be challenging for multiple reasons. It relies on a nominal response model to accurately estimate the probability of a specific response across the ability continuum. For each pair of test takers, the joint probability of matching responses is calculated and compared to the actual number of matching responses. The model can require substantial sample sizes, can be computationally intensive, and the current software and model fit can be finicky.

The expected count of matching responses should be less than the expected count of matching scores. Further, a test taker's score on the exam should be an accurate estimate of their average expected score on any given item. For example, on a 100-item dichotomously scored test, a candidate with a score of 70 would be expected to score 0.70 on each item on average. Total scores can, therefore, be used to estimate the expected number of matching scores. The proposed method uses total scores to estimate expected number of matching scores for pairs of test takers. Then, the number of matching responses are compared to the expected number of matching scores. Though the model does not have a specific underlying distribution, the ratio of actual to expected is a useful surrogate that is easy and relatively efficient to calculate.

12:00 - 12:30

VIRTUAL LUNCH ROOM

Caveon Test Security invites you to join in a celebration of the contributions to the test security field by Dennis Maynes as he retires and pursues his next adventure. In his 18 years as Chief Scientist, Dennis has performed or supervised analysis of more than 40 million test instances in the certification, education, and military/government industries. Dennis has contributed significant advancements to many areas of test security, including the computation of similarity statistics, creating innovative algorithms to analyze clusters and groups of tests, analysis of unusual response time data, and designing items and tests to increase test security and detect various types of test fraud. Please join us to share your favorite stories about Dennis and wish him well.

12:30 - 1:15

Using the Caveon Advanced Search Function in Scorpion Deliveries to Locate a Cheater on an Exam Created Using Dynamic Forms

Donovan Allen, Caveon | Tara Williams, Caveon | Liana Maharaj, Slack

After discovering an exam breach, Slack used their exam design to determine the culprit. Explore various ways to create exams to minimize theft and determine the source of a breach.

1:15 - 1:30

Break

1:30 - 2:15

Making Testing Condition Decisions: Risk Assessment and Security Examples

Kimberly Swygert, NBME

The COVID-19 pandemic of 2020 caused serious disruption to almost all testing programs, and the exam programs supporting the medical education and medical licensure pathways in the US were not spared. Schools of medicine in the US sent their students home and transitioned to remote education platforms almost immediately, which meant that exams could no longer be administered at medical school testing centers with in-person proctors. In addition, brick-and-mortar testing centers closed in the spring of 2020, which meant large-scale licensure exams were temporarily halted. During this time, the NBME had to decide whether certain high-volume exam programs should be moved to a remote administration setting or remain closed until medical schools or testing centers re-opened. An important part of this decision-making process focused on risk assessment, which included threats to the security of the exams and the validity evidence of the test scores. This presentation will provide information about two exam programs that varied greatly in terms of the security concerns and potential risks, where both exam programs were considered integral parts of the pathway of becoming a physician within the US. In addition, there will be discussion of the risk assessment findings, the resulting exam administration decisions that were made, and specifics of additional data analyses, lessons learned, and security aspects to consider.

THURSDAY, OCT 7
10:00 A.M. – 2:15 P.M. EDT

10:00 - 10:45

Using Data Forensics to Manage Item Disclosure

Marcus Scott, Caveon

Item disclosure remains one of the greatest threats to test security and the validity of test scores. Many testing programs employ web patrol to find disclosed items on the Internet, but that is not the only tool available. This demonstration will show how data forensics analysis can provide useful information for detecting disclosed content and crafting an appropriate response.

The demonstration will begin with a discussion of how item disclosure affects p-values, response times, pass rates and mean scores, and other test result data. Next, data forensics methods designed to detect anomalies in these types of data will be presented. This section will cover currently used methods for detecting disclosed items and outline some research topics that will potentially yield additional useful techniques.

The remainder of the presentation will focus on using information from data forensics analysis to manage occurrences of item disclosure. To effectively manage item disclosure, efforts must be made to devalue the disclosed content and measure the extent to which the disclosed content was used. Disclosed test content is devalued when the disclosed items are removed from service and replaced by other items. Data forensics analysis can inform a testing program's schedule for rotating items and republishing exams. Monitoring item performance, score differences, and cluster sizes over time can give valuable insights into how usage of disclosed test content spreads among a testing population.

10:45 - 11:00

Break

11:00 - 12:00

Standard Presentations II

Detecting Cheating Behaviors Through Joint-Modeling Responses and Response Time Based on SEM

Qingzhou Shi, University of Alabama | Kaiwen Man, University of Alabama | Qiwei He, ETS | Mengya Xia, University of Alabama

Response time (RT) is a representative, popular, and valuable index for test-taking assessment that has been modeled to detect aberrant testing behaviors, select items in computerized adaptive testing, and improve modeling parameter accuracy. This study examines whether applying Structural Equation Models (SEM) to joint-model item-level responses and RTs in a computer-based unidimensional test can detect cheating-related behaviors across groups of test-takers with versus without pre-knowledge of the test items. Responses and RTs are from a controlled experiment containing 10 items similar to the actual test items with categorical outcomes. Of all the 200 participants, 93 did not have pre-knowledge of the test, and 107 had. A bivariate model of responses and RTs was applied: the normal ogive model (Samejima, 1974) was utilized to model item responses, and item-level RTs were estimated by an SEM-based Response-RT joint model (Wang et al., 2019). In the SEM-based Response-RT joint model, person abilities and working speed were treated as endorsed latent constructs, and test anxiety was treated as a covariate. Data were analyzed via Mplus Version 8. Results indicated that 1) the most time-intensive items in the non-cheating group were less time-intensive in the cheating group; the least time-intensive items in the non-cheating group were the most time-intensive in the non-cheating group; 2) person abilities demonstrated correlation with working speed only in the cheating group; 3) anxiety was significantly associated with working speed only in the cheating group; and 4) path and latent construct estimates demonstrated significant differences across cheating and non-cheating groups.

Analysis of Item Response Time Patterns through the Lens of Profile Similarity and Model Fit Metrics

Greg Hurtz, California State University, Sacramento and PSI Services LLC | Regi Mucino, PSI Services LLC

Analyzing test-takers' item response times is an increasingly common practice because it can help identify patterns that may be associated with fraudulent activity. For example, fast responding can indicate rapid guessing or preknowledge, and slow responding may be indicative of item harvesting. Metrics that assess how well test-takers' response times conform to what was found when establishing item parameters have been developed to provide additional information beyond speed. Not fitting the normative model may suggest response times are associated with abnormal test-taking behavior. However, existing measures of misfit fall short of providing information about the pattern of misfit involved.

In this presentation we frame the analysis of response time fit through the lens of profile similarity to identify multiple angles on the exploration of response times. Profile similarity metrics compare a person's values across a set of variables to a target profile, and describe 3 distinct features: dispersion, shape, and level (Nunnally & Bernstein, 1994). We evaluated 9 measures that explore response times through this perspective and frame some from the perspective of absolute and comparative fit concepts seen in the confirmatory factor analysis and structural equation modeling literature.

In both "clean" (N = 3316) and "compromised" (N = 6776) datasets, 3 distinct types of response time measures emerged through principal components analysis that correspond to dispersion, shape, and level. Existing measures assess only dispersion and level. Approaching response time analysis through profile similarity, we provide 3 alternative shape measures to add to existing methods for evaluating item response times.

A Multilevel Approach for Identifying Test Takers with Aberrant Patterns of Item Response Time

Yi Lu, Federation of State Boards of Physical Therapy | Yu Zhang, Federation of State Boards of Physical Therapy | Lorin Mueller, Federation of State Boards of Physical Therapy

Test response time provides information about test takers' ability and response behavior as well as about item and test characteristics (Marianti, Avetisyan, & Veldkamp, 2014). Response time has been modeled as sole dependent variable, or together with other variables as dependent variable, or as independent variable (Boeck & Jeon, 2019). The purpose of the current study is to identify test takers with aberrant patterns of item response time using a multilevel approach. The aberrant response time pattern considered in this study is defined as a response pattern that is different from the majority of test takers with a similar ability. It may indicate test performance with item preknowledge or guessing due to out of testing time. In this study, item response time was treated as sole dependent variable in three multilevel models: an unconditional means model which does not contain any predictor, a conditional model including item difficulty as the level-1 predictor, and a conditional model including item difficulty as a level-1 predictor and scale score as a level-2 predictor. The model effect is used to evaluate the aberrance of response time patterns. Specifically, test taker who have the largest negative random effect on intercept have the shortest response time. The largest negative random effect on slope indicates test taker's pattern of response time is the most unrelated to model fit. Results based on real data and simulation will be discussed in the final paper.

12:00 - 12:30

VIRTUAL LUNCH ROOM

12:30 - 1:15

How Do YOU Define Security?

Jarrod Morgan, Meazure Learning (ProctorU)

"How do YOU define Security?" is a roundtable discussion focused on how testing programs define secure exam delivery across both in-person and online invigilation. How do you know that your exam content is secure? Is security different for an in-person exam compared to one delivered remotely? What metrics should be reviewed to assess security? Testing programs and other industry leaders will be encouraged to discuss how their testing programs have changed during the pandemic and what impact these changes may have on exam delivery and ultimately exam security. This discussion will also include an overview of different metrics that can be used to measure security and gather feedback from the group on what works and maybe more importantly, what doesn't.

1:15 - 1:30

Break

1:30 - 2:15

Data Forensics as a Multi Tool: Statistical Analysis for Monitoring or Investigations

Sarah Toton, Caveon | Kevin Arndt, UBC Sauder School of Business

The goal of data forensics is to use statistical evidence to inform and support test security decisions. However, the use of data forensics may look dramatically different for routine security monitoring versus investigation of a known or suspected breach or incident. For testing programs to get the most value from data forensics, their goals and test security situation should inform the analysis plan and selection of statistics. In this session, we will discuss how the design of a data forensics analysis, the statistics used, and the resulting inferences may differ based on a program's security situation. Several situations will be considered, including 1) a program that desires regular monitoring, and does not know of any particular security issues, 2) a program that has become aware of a potentially widespread breach (e.g., test content has been disclosed on the internet), and 3) a program that has become aware of a potentially limited breach (e.g., a proctor has reported something regarding particular examinees).

WEDNESDAY, NOV 3
10:00 A.M. – 2:15 P.M. EDT

10:00 - 10:45

Lessons Learned Protecting Test Content and Personal Information in High Stakes Testing Programs

John Fremer, Caveon | David Foster, Caveon

Challenges to the security of our high stakes testing programs have been intensifying. Growing numbers of thieves are stealing and selling test questions, miniaturization technology is supporting undetectable recordings of testing sessions, and cheating technology is being sold on the internet. The ongoing validity of our test scores is being seriously threatened.

In response, the last decade has seen significant advances in data forensic science, web monitoring models, secure item designs, and test delivery safeguards. While these measures have improved test security in many assessment programs, the threat remains clear, present, and dangerous.

The presenters will highlight test security responses to the growing security challenges, in the areas of prevention, deterrence, detection, and follow-up actions. Specific and actionable ideas will be provided for test program planners and managers to enhance security in every aspect of a high-stakes assessment program. Major lessons learned from dealing with security challenges in international testing programs will be shared, offering ideas that have the stood the test of time.

Many testing programs are looking for new solutions. Often these solutions move beyond the century-old reliance on static multiple-choice items, traditional proctoring methods, and conventional test administration models. The panelists will look at research underway and new technologies being developed to help programs transition from traditional approaches to more secure testing environments. Future directions for test security will be considered, including the promise of cheat resistant item design and delivery.

Test security threats will continue to evolve. We will all have to keep learning.

10:45 - 11:00

Break

11:00 - 12:00

Standard Presentations III

Technology and Creativity: Leveraging AIG and SmartItems to Develop Great, Secure Exams

Steve Addicott, Caveon | Janet Lehr, HPE

New technologies enhance the creativity of test developers in profound ways, and creativity can go a long way in protecting the validity of test scores. With large item banks, programs can reduce the impact of test fraud (cheating and item theft) by leveraging innovative test designs, including multiple forms; linear on the fly (LOFT); computer adaptive testing (CAT); rapid republishing of forms, and others. Long considered a luxury affordable to only the largest, well-funded programs, large item banks may now be realized by any program. Automated Item Generation (AIG) transforms item pools into "item oceans." Though AIG has been around for years, more and more programs are now realizing its promise as new AIG engines are simpler to integrate and use.
While the designs mentioned have great security impact, "one and done" tests (where an item and/or form is administered only once before being retired) represent the ultimate in test security. However, incorporating "one and done" on a large scale represents cost and item bank complexity that has been simply unachievable. Until now. "SmartItems" represent a "treatment" to an item that dynamically renders each item during an exam so every test taker's experience with the exam is truly unique. Your form has been stolen? No matter when those compromised versions of items will never be presented to another test taker. As programs grapple with growing test fraud challenges, AIG and SmartItems are utilized by evermore programs to not only meet test security challenges but prevent them in the first place.

Machine Learning Approach to Prevent Item Preknowledge

Dmitry Belov, Law School Admission Council

The objective of item response modeling (IRM) is to predict statistical parameters of an item (e.g., difficulty) based on features extracted directly from the item (e.g., number of words). Recently, most high stakes testing programs worldwide had to migrate from offline to online testing environment. Online test proctoring cannot beat test collusion and modern technology for stealing test content. Therefore, item preknowledge may occur caused by (a) using the same test over different time slots due to limited number of live proctors and by (b) pretesting new items. Larger number of tests is needed to avoid (a). However, creating new items (by a test developer or by an automated item generator) without controlling their statistical parameters may unbalance item pool and, thus, limit the assembly of more tests.
A deep neural network (DNN) model has been developed to predict parameters of three parameter logistic model. With minor modifications this DNN can predict item characteristic curve (ICC), which allows supporting empirical ICCs and multiple IRT models. The 3548 items from a high-stakes testing program were used to train and validate the DNN. Major stages of this development will be demonstrated including mining features and identifying hyper parameters of the DNN in order to minimize variance of the bias while keeping its distribution symmetric about zero. The latter is critical when a test with multiple items is assembled using predicted statistical parameters of the items. Detailed results of a cross-validation of the developed DNN model will be presented.

Robustness of a Statistic Based on the Neyman-Pearson Lemma to Violations of its Underlying Assumptions

Sandip Sinharay, ETS

Drasgow, Levine, and Zickar (1996) suggested a statistic for detecting preknowledge on a known set of items. The statistic is based on the Neyman-Pearson lemma and is the most powerful statistic for detecting item preknowledge when the assumptions underlying the statistic hold for the data (e.g., Belov, 2016; Drasgow et al., 1996). The first goal of this paper is to demonstrate using real data analysis that one assumption underlying the statistic of Drasgow et al. (1996) may not hold in practice. The second goal of this paper is to examine, using simulated data, the extent of the robustness of the statistic to realistic violations of its underlying assumptions. Together, the results from the real data and the simulations demonstrate that the statistic of Drasgow et al. (1996) may not always be the optimum statistic and occasionally has smaller power than another statistic for detecting preknowledge on a known set of items, especially when the assumptions underlying the former statistic do not hold. The findings of this paper demonstrate the importance of keeping in mind the assumptions underlying and the limitations of any statistic or method.

12:00 - 12:30

VIRTUAL LUNCH ROOM

12:30 - 1:15

Exam Security: An All-Inclusive Discussion on Key Considerations for your Assessment Program

Ashely Norris, Meazure Learning | Kim Brunnert, Elsevier | Ray Nicosia, ETS | Faisel Alam, Law School Admission Council

Testing programs have been challenged from a number of angles over the past year and a half considering the impact of COVID. They had to quickly adjust to ensure sustainability and continue to demonstrate value to test-takers. From changes to exam administration modality or schedule, review and redesign of exam development processes, or modification of policies, it required creativity, flexibility, and resilience. At the center of all these changes was the importance of maintaining the integrity and security of their program, for all aspects of the assessment life cycle.

Participants on this panel will discuss how they have tackled these challenges and solutions they have put in place to maintain and enhance security for their programs. They will discuss key considerations related to exam delivery modalities, especially as it relates to online proctoring. They will discuss how they use data to monitor for testing integrity concerns as well as how to maintain security and integrity in providing testing accommodations. They will also address test development and assembly strategies to minimize risks of content exposure. Finally, they will discuss how they use data and reports to identify trends in suspicious behaviors and how to stay on top of their game as it relates to cheating and potential threats to the integrity of their exam program. Walk away with a clear understanding of how these high stakes programs adapted during COVID and what changes they have adopted permanently.

1:15 - 1:30

Break

1:30 - 2:15

Automated Item Generation: Security Through Variation

Brooke Dresden, PSI | Tomer Gotlib, PSI | Xin Li, PSI

Automated Item Generation (AIG) is a process which uses cognitive and psychometric theories to produce an unlimited number of item variations from a single item template utilizing computer technology. AIG serves the need to enhance test security as it creates unique variations of items and response options while maintaining high-quality items. However, a good foundation is needed to properly design a model to generate and prepare these items for practical use.

This session will provide a brief introduction to AIG, a demonstration of AIG item construction through a graphical user interface, and steps for implementing AIG items in testing programs to improve test security. In addition, results from a research study aimed at measuring statistical performance data on AIG item variations from candidates completing both high-stakes (scored) and low-stakes (practice) exams to evaluate the statistical performance of item variations will be reviewed.

THURSDAY, NOV 4
10:00 A.M. – 2:15 P.M. EDT

10:00 - 10:45

Don't Just Tell Me You Know It...Prove It! Security Considerations in Remotely Delivered High-Stakes Performance Exams

Benjamin Hunter, Caveon | Jim Lucari, HPE | Kathy Murphy, CDI

For many years before the COVID-19 Pandemic forced a wave of exams to migrate online, IT Certifications had been plagued by theft that challenged exam validity. Traditional methods of securing exams still allowed for massive losses of content with major impacts to exam validity and the trustworthiness of test results. Through a combination of innovative item design and performance testing strategies, HPE has designed a test that combines multiple elements such as: one-time use multiple choice questions; Discrete Option Multiple Choice items (designed to limit content exposure); performance tasks completed on real hardware and automatically scored; remotely proctored; and accessible by any test taker world-wide; with a lower delivery cost over traditional performance exams. This session will explore the pitfalls of traditionally designed IT certification exams; and discuss how the methodologies employed on the HPE performance exams eliminate the security, validity, and measurement problems presented when delivering standard multiple-choice format questions. The secure-by-design nature of these exams make the remote proctor primarily responsible for authentication and observation, while also providing technical support to the candidate if needed and appropriate for the exam design.

10:45 - 11:00

Break

11:00 - 12:00

Standard Presentations IV

The Utility of Nonfunctioning Distractors for Credentialing Tests

Merve Sarac, University of Wisconsin-Madison | Richard Feinberg, National Board of Medical Examiners

While literature on constructing quality multiple-choice questions is vast, the focus has primarily been on development of the stem and correct option (Gierl et al., 2017). The incorrect options or distractors have received relatively little attention (Thissen, et al., 1989), though recent research suggests that many distractors are nonfunctional prompting practitioners to question the value and necessity of including five response options (Rodriguez, 2005; Raymond et al., 2019). For instance, a rarely chosen response option is not decreasing the chances that an examinee guesses correctly, as intended, but is contributing to test development costs and the time an examinee needs to read all the response options. However, nonfunctional distractors (NFD) may still prove useful across the myriad of ways an operational testing program investigates potential threats to score validity. Thus, the purpose of the present study is to explore the utility of NFDs in the context of licensure/credentialing examinations and their implications for detecting aberrant response behavior. Discussion of results will focus on possible distinct examinee behavior patterns that could be reflective of low motivation, insufficient time limits, population knowledge deficiencies, or malicious content harvesting, all of which may warrant policy consideration and action such as stopping rules or limited number of future attempts.

Simulating the Security Benefits of SmartItems

Andrew Marder - Caveon

This presentation will address three questions:

1. What is a SmartItem™?
2. How do SmartItems improve test security?
3. How much do SmartItems improve test security?

The bulk of the presentation will focus on the third question, quantifying the security benefits of SmartItems. A numerical simulation will provide concrete numerical answers. Like all simulations, the conclusions from this simulation depend upon its assumptions. We aim to make the simulation flexible, accommodating a range of assumptions, to illustrate how our conclusions change with our assumptions. This simulation study compares the effectiveness of an exam built with SmartItems to an exam with a fixed set of items. We explore how exam length, item difficulty, item discrimination, size of security breach, and usefulness of pre-knowledge impact reliability and accuracy of rankings for the two exams.

The Importance of Unpredictability in Security Measures for Standardized Tests

David Foster, Caveon

While its use may seem a contradiction, world-class security, regardless of the field, relies to a major extent on randomness and its major effect, unpredictability. Examples of the security value of unpredictability will be provided for such diverse fields as information technology, finance, policing and the world of gambling. And, of course, high-stakes testing.

The session will then cover how limited forms of randomness are commonly used in today's test designs and test administration procedures to thwart cheaters and intellectual property thieves. The session will then describe testing technology that utilizes randomness in more extensive and innovative ways, with the goal of completely preventing most forms of cheating. A case study will be described where such technology is routinely and profitably used.

Whenever randomness is used in testing, concerns about standardization are inevitably raised. The session will end with the surprising proposition introduced decades ago that adding more randomness to testing actually contributes to greater standardization.

12:00 - 12:30

VIRTUAL LUNCH ROOM

12:30 - 1:15

Important Test Security Lessons Learned From Quick Transitions to Remote Proctoring

Marc Weinstein, Caveon | Roger Meade, Prometric | Thomas Gera, The Enrollment Management Association | Patrick D. Watts, Project Management Institute | Liz Grater, HR Certification Institute (HRCI)

The pandemic forced many high-stakes testing programs that were no longer able to test in-person to make the difficult decision to quickly pivot their test administration and delivery operations to offer remote proctored online testing so that examinees could take tests from home. The change was even more dramatic for some programs that historically tested on paper and pencil and were now switching for the first time to computer-based tests delivered remotely to examinees in their homes. Managing these significant transitions over a very short time period presented numerous security, technical, privacy, and legal challenges for testing programs. This panel discussion will include panelists from several programs that managed these abrupt changes to their program and can speak to their experiences and lessons learned over the course of remote proctored online testing throughout the pandemic. The panelists will identify the greatest challenges they faced and the solutions they found to be most effective. The panel will also include the vendor perspective to address the unique challenges vendors faced in assisting programs in making the abrupt shift to remote proctored online testing. The focus of the panel will be real-world examples of specific security challenges and practical solutions that panelists found effective and solutions that panelists found to be less than effective. Panelists will share their own experiences for the first thirty-five minutes, and save the last ten minutes of the session for questions and comments from the audience.

1:15 - 1:30

Break

1:30 - 2:15

Framing Test Security as a Big Picture Issue: Communication Strategies

Lorin Mueller, ***Federation of State Boards of Physical Therapy*** **| Nyka Corbin,** ***Financial Industry Regulatory Authority (FINRA)*** **| Ray Yan,** ***Financial Industry Regulatory Authority (FINRA)***

One of the challenges to implementing test security measures is the difficulty many stakeholders have with seeing the direct contribution of security to the broader mission of organization. This presentation will focus on strategies that have been successful at linking test security to concepts to which many stakeholders can easily relate. One set of strategies involves making a clear line of sight between security policies and the validity of scores, and the importance of score validity to the organizational mission, brand, and relationships with other stakeholders (such as schools, regulators, and industry partners). These strategies make the rationale for security measures explicit to stakeholders in a way that makes it clear how these measures impact strategy and mission. Another set of strategies involves putting metrics in terms that are more relatable to stakeholders outside of the testing industry and providing more context around metrics. For example, showing trends over time (such as the frequency of incident reports or flagged scores) and relating those to policy changes, or showing factors that influence test scores and how those patterns have changed over time. Another way to frame metrics is to put them into financial terms many decision makers are more familiar with, like describing forensic analyses as "audits", putting dollar value terms on your item bank, and formatting item development reports to resemble financial reports. The presenters will show examples of stakeholder directed communications using these strategies throughout the presentation. Time will be allotted to encourage participants to share their own strategies.

WEDNESDAY, DEC 1
10:00 A.M. – 2:15 P.M. EDT

10:00 - 10:45

Latest Updates on Lessons Learned in Maintaining and Improving Test Security for State Assessments: A State Panel Discussion

John Olson, Caveon/OEMAS | Theresa Bennett, Delaware DOE | Sandra Greene, Georgia DOE | Jeff Holtz, Minnesota DOE | Walt Drane, Caveon

Over the past year, many lessons have been learned by states on improving test security and maintaining the integrity of their state assessment programs. The COVID-19 pandemic shut down many districts and schools in 2020-21 and often prevented normal testing from taking place, requiring states to rethink how their assessments were to be administered. States had to become much more proactive in their approaches for implementing stronger procedures and policies to prevent improprieties and ensure validity of test scores.

In this session, states will provide updates on the latest lessons they have learned during the pandemic and their plans for the future. Information will be provided from a variety of state perspectives, with each state describing how they maintained and improved the security of their assessments. Panelists from three states (DE, GA, and MN) will share details on their recent activities and describe promising and effective strategies/practices being implemented in the future. Presentations will address several important themes – (a) recommended methods/approaches/guidance to improve test security in states, (b) meeting federal requirements for peer review, especially for test security and monitoring of test administrations, and (c) latest lessons learned and best practices for preventing cheating and increasing the integrity of state assessments.

10:45 - 11:00

Break

11:00 - 12:00

Standard Presentations V

Detection of Possible Cheating at the Group Level, Using a Statistical Hypothesis Test Based on Meta-Analysis

Edmund Jones, Cambridge Assessment, U.K.

When the same test is run in a large number of schools, it sometimes happens that the testing organization suspects that in one school a large proportion of the test-takers might have cheated, or been able to cheat. A statistical hypothesis test is needed to judge whether the school's scores are abnormally high. Sometimes previous years' scores can be used as a reference, but this presentation is about the case where previous scores are not available. The hypothesis test needs to allow for legitimate variation between schools' average scores and their sizes. This presentation will describe a statistical model and hypothesis test that use ideas from random-effects meta-analysis (a method widely used in medical science). Under the null hypothesis, each school has its own normal distribution with its own mean, and the school means themselves also follow an overall normal distribution. Under the alternative hypothesis, the same is true, except that for the school under investigation the mean does not follow the overall distribution. The hypothesis test is a relative likelihood test. A p-value can be calculated by simulating from the null hypothesis model.

The school means can be shown on a forest plot (another idea borrowed from meta-analysis). This shows for each school the mean, the standard error of that mean, and the size of the school. Simulations to evaluate the method are not straightforward, because calculating the p-value is slow, but small-scale simulations will be presented.

Two-Step Detection of Examinees with Preknowledge and Exposed Items

Merve Sarac, University of Wisconsin-Madison | Ting Xu, AICPA

In a previous study, we borrowed information on one item format (multiple-choice questions (MCQ)), which is known to be secure, to detect preknowledge on another format (task-based simulations (TBS)) within a test. The performance of two different methods for preknowledge detection were investigated and compared through a simulation study: a differential person functioning approach and a regression method with prediction intervals. The results showed that the differential person functioning approach yielded more power than the regression method with prediction intervals. For both approaches, power decreased as the percentage of examinees with preknowledge increased, and as the number of compromised items decreased. The current study extends the previous research by evaluating a two-step method on detecting examinees with preknowledge as well as specific items that are likely to be compromised. Specifically, differential person functioning is utilized in the first step to detect examinees with preknowledge by comparing their performance on operational versus pilot items. Given the group of examinees flagged in the first step, differential item functioning is performed in the second step to identify compromised items. A simulation study mirroring a high-state licensure exam will be conducted in which proportions of compromised items and proportions of examinees with preknowledge are manipulated. Type I error rate and power are computed to evaluate the results of each simulated condition. Real data analyses will also be included.

Mitigation and Monitoring Strategies of a Large-Scale Licensure Exam

Matthew Schultz, AICPA | J. Carl Setzer, AICPA

Each testing program has a unique set of test development, design, and delivery systems. Thus, there is no one-size-fits-all approach to test security. Nevertheless, exam developers must be aware of their potential for vulnerabilities, and program-specific mitigation strategies can drive decisions about where to shine the light. Broad categories of cheating such as answer copying, collusion, tampering, proxy testing, and preknowledge can potentially be minimized as a result of certain decisions, including those related to monitoring approaches, policy implementation and communications around the consequences of cheating. An implication of strong test security is that it becomes harder to detect small scale or localized breaches. This presentation will highlight the mitigation strategies utilized by one large-scale licensure organization. We will describe the test design and delivery strategies that enhance the overall security of the exam, as well as the implications on detection. In addition, the presentation will discuss the operational approaches used to establish baseline measures for monitoring and how those measures can be utilized in case of a test security issue. Limits to what these measures can (and cannot) capture will be discussed.

12:00 - 12:30

VIRTUAL LUNCH ROOM

12:30 - 1:15

Standard Presentations VI

Humanizing Virtual Proctoring

Mac Adkins, SmarterServices

The pandemic prompted the creation of a new proctoring modality known as Hybrid Virtual Proctoring (HVP) which combines the security and deterrent of a live, human proctor with the efficiency and scalability of virtual proctoring. Using HVP a candidate is informed that during their proctored exam a person from the organization from which they are taking an assessment may be watching them in real-time. Or there may be a person from a proctoring network monitoring them. Finally, they are assured that at a very minimum their testing session is being monitored by artificial intelligence that is flagging possible testing anomalies.

One of the criticisms that providers had about automated virtual proctoring during the pandemic was that the artificial intelligence when working autonomously does not have the power to stop an assessment when an anomaly is detected. This gap in response time could seriously endanger the assessment content. But with HVP, when a live person is monitoring an exam, they can make that judgment call and stop the exam if needed.

HVP can also achieve a scale of efficiency not possible when schools attempt to do their own form of hybrid proctoring using a tool such as Zoom. The HVP interface in SmarterProctoring allows one live, human proctor to simultaneously monitor up to a dozen test takers. HVP is superior to Zoom in that it couples artificial intelligence with live monitoring. It also provides 24/7/365 support to all candidates, performs device compatibility checks, and a lockdown browser.

Increasing Test Security and Combating Threats Through Technology

Shailu Tipparaju, Examity | Shari Lewison, University of Iowa | Rachel Schoenig, Cornerstone Strategies

As testing programs have increasingly incorporated more technology into their test development and delivery processes, testing staff, volunteers, and test takers have benefited from the enhanced flexibility and convenience it provides. Whether working from home, engaging in online item development and reviews, or participating in online testing, technology has provided testing staff, volunteers, and test takers with anytime / anywhere access. But along with these benefits comes new risks to our test items and scores.

The reality is cyber threats have become more common and more diverse, as hackers have created new ways to gain unauthorized access to information. From ransomware and social engineering to monitor mirroring and the use of virtual machines, new technology also exposes our programs to new cyber risks and concerns. In this presentation, Examity's Chief Innovation Officer, Shailu Tipparaju, and the University of Iowa's Chief Information Security Officer, Shari Lewison, will discuss current cybersecurity trends and emerging risks. How is technology being used to undermine secure testing? What tools should organizations be aware of to combat these threats and better ensure the ongoing security of their exams and validity of test scores?

Join experts for an engaging conversation on these topics. Attendees will walk away with a better understanding of trending cyber risks that threaten testing and ways they can better protect their tests and the testing process. We hope that you will join us for this informative and timely session.

1:15 - 1:30

Break

1:30 - 2:15

When Words ARE Enough

Camille Thompson, The College Board | Jarret Dyer, College of DuPage | Rachel Schoenig, Cornerstone Strategies

Testing programs have received significant exam security benefits from technological advancements. From wanding and palm veins scans to online proctoring, artificial intelligence, and web monitoring, technology has helped to enhance protection of exam content and the integrity of testing events. Sometimes in the press to use the newest technologies, however, we can lose sight of one of the most powerful tools in our exam security toolbox: the power of the written and spoken word. While it may seem like a simple tool, the words we use to craft agreements, set forth rules, communicate nudges, and enforce policies weave a strong web of protection around our exams. During this session, we will discuss how testing programs can increase exam security by giving thoughtful consideration to the words we use in our contracts, honor codes, attestations, reminders, nudges, and other types of communications to key stakeholders. Join testing professionals as they share how words can provide some the strongest exam security tools we have to protect test content and ensuring score validity. It's a practical session you won't want to miss!

THURSDAY, DEC 2
10:00 A.M. – 2:15 P.M. EDT

10:00 - 10:45

Standard Presentations VII

A Strong Security Cocktail—Shaking up Fraudsters with a Splash of Technology and a Dash of New Techniques

Leah Hojem, UiPath | Amy Ressler, College Board | Janet Lehr, HPE

As all exam and assessment professionals know, protecting an organization's intellectual property from theft requires continual due diligence. Yet, technological advancements will continue to pose risks in the development of secure and fair assessments. In fact, the growth and popularity of social media alone can represent a major threat to test security. Let's not forget to include a large dash of collaboration with industry vendors to identify candidate misconduct and/or websites with potentially compromised content. We will share experiences implementing the policies and practices used to create secure and legally defensible exams, statistical analysis and services used to identify candidate misconduct, cheat site detection, and remediation.

Surprising Test Security Advice from Mid-Century Testing Experts

David Foster, Caveon

Cheating has been a part of testing for thousands of years, plaguing every important testing program and event along the way, and continues at uncomfortably high levels today. A major effect of cheating is to negatively impact the validity of test scores. Because cheating is usually hidden from view, it makes us question not just the scores of those who actually cheated, but all other scores as well. These problems associated with cheating were also recognized by testing experts from the 1950's through the 1970's. It was during this time period that the exciting promise of the computerization of testing first appeared. Clearly noted at the time, one of the benefits of marrying this new computerized technology with testing, were creative solutions to the problem of cheating. From out of such a time, significant contributions came from such individuals as Frederick Lord and Lee Cronbach, among others. With the prospects of exciting technological changes before them, a few of these individuals proposed innovative test theories, statistics, designs and procedures which would help prevent or discourage test fraud.

It is the purpose of this presentation to cover these security contributions, to evaluate them according to the technologies of today, and to describe how they are currently or might be implemented to solve the security problems of today.

10:45 - 11:00

Break

11:00 - 12:00

Statewide Testing During the COVID Pandemic: The Challenges of Conducting Secure and Valid State Assessments in 2020-2021

Timothy Butcher, West Virginia Department of Education | Lynn Schemel, Indiana Department of Education | David Ragsdale, Massachusetts Department of Elementary and Secondary Education | John Olson, Caveon | OEMAS | Walt Drane, Caveon

With COVID-19 suddenly impacting educational systems across the country, testing in Spring 2020 offered many new challenges and uncertainties in the form of a pandemic, which ultimately led to the cancellation of statewide testing. States had to pivot quickly to come up with plans for possibly conducting assessments differently in the coming year. With distance learning and remote testing increasingly being used in schools and districts across the US, the security of state assessments came into question along with how to maintain validity of the results. In a short period of time, many new approaches had to be designed/developed by states. Spring 2021 moved statewide testing forward but required states across the country to adapt new policies and procedures to allow assessments to be redesigned, and to possibly occur remotely (usually in the student's home) or in unconventional situations within school districts.

In this Coordinated Symposium, a large amount of data gathered from states on their plans for testing in 2020-21 will be shared. This session promises to inform participants of many new challenges and lessons learned by Assessment Directors in three states (Indiana, Massachusetts, West Virginia) who securely administered tests in 2020-21 during the ongoing COVID pandemic. Among the challenges discussed are secure testing in remote environments, monitoring test administrations, data forensics approaches to detect possible cheating, increased Web monitoring of Internet/social media for disclosed assessment materials, and how scores from statewide summative assessments were analyzed, validated, and utilized.

12:00 - 12:30

VIRTUAL LUNCH ROOM

12:30 - 1:45

COTS Keynote Debates

Organized and Moderated by Rachel Schoenig, Cornerstone Strategies, and Kim Brunnert, Elsevier

Once again this year we will be offering the much heralded and never-to-be-missed COTS Keynote Debates. Seasoned professionals will be discussing some of the industry's hottest topics. You'll learn from the experts, shape and inform your opinion, and gain the wisdom of the crowd. The COTS Keynote Debates: in a world where so many things have shifted, it's nice to know some traditions remain the same?