THE 7TH ANNUAL

CONFERENCE ON TEST SECURITY

OCTOBER 10-12, 2018 | PARK CITY, UT

PRESENTED BY CAVEON TEST SECURITY

THIS PROGRAM BELONGS TO:

PLEASE RETURN IF FOUND

WELCOME

TO COTS 2018!

Dear Colleague,

Welcome to the 7th annual Conference on Test Security – the only event dedicated entirely to test security. It is my pleasure to have you with us in beautiful Park City, Utah. In this appealing town, you will find every outdoor recreational and social activity your heart could desire.

When COTS was first started in 2012, it was called the Conference on the Statistical Detection of Potential Test Fraud, and focused primarily on statistics. In 2014, the scope of the conference broadened to encourage dialogue on all test security capabilities and enhancements. Now in its seventh year, COTS has rightfully gained a reputation for facilitating the best atmosphere encouraging open and honest test security discussions amongst professionals like yourselves.

While the ever-constant threats to the security of our tests continue, it is your zeal for wisdom and innovation that keeps those dangers at bay. It is my sincerest hope that this year's conference will be of the utmost benefit to you as we learn and network together. With over 50 sessions, and attendees from over 75 organizations worldwide, we hope you'll find information and resources that are helpful to you and your program's needs.

If you're new to COTS, we hope that you'll join us for the Cocktails and Conversations event in the Sundial Pavilion on the evening of October 11th. Here, you'll find the opportunity to converse one-on-one with some of the brightest minds in testing and participate in our poster presentations. Dubbed as "the favorite event of COTS", you won't want to miss meeting with leaders in the field on this up-close and personal level.

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 can gather together and forward our joint goal to protect the validity of test results and brand integrity.

We hope that you enjoy your time in Park City, Utah at the 2018 Conference on Test Security.

Warmest regards,

David Foster
CEO & President
Caveon Test Security

COTS
Conference On
Test Security

A SPECIAL THANK YOU TO OUR CO-HOSTS

caveon™

Test Security

THANK YOU TO OUR FRIENDS

CONFERENCE AGENDA

Wednesday, October 10th

9:00 A.M. - 6:30 P.M.	REGISTRATION & INFORMATION WHITE PINE LOBBY
12:00 - 5:00 P.M.	WORKSHOPS WHITE PINE BALLROOM
5:00 - 6:00 P.M.	MEET & GREET WHITE PINE LOBBY
6:30 - 8:00 P.M.	COTS EXECUTIVE COMMITTEE MEETING THE CABIN BOARDROOM

Thursday, October 11th

7:00 A.M. - 5:30 P.M.	REGISTRATION & INFORMATION GRAND BALLROOM LOBBY
7:00 - 8:00 A.M.	CONTINENTAL BREAKFAST GRAND BALLROOM LOBBY
8:00 - 9:30 A.M.	OPENING KEYNOTE KOKOPELLI PARLORS II & III
9:30 - 9:45 A.M.	BREAK
9:45 - 10:45 A.M.	SESSIONS WHITE PINE, ARROWHEAD, KOKOPELLI I
10:45 - 11:00 A.M.	BREAK
11:00 A.M. - 12:00 P.M.	SESSIONS WHITE PINE, ARROWHEAD, KOKOPELLI I
12:00 - 1:00 P.M.	LUNCH UMBRELLA BAR
1:00 - 2:30 P.M.	SESSIONS WHITE PINE, ARROWHEAD, KOKOPELLI I
2:30 - 3:00 P.M.	EXTENDED BREAK

CONFERENCE AGENDA

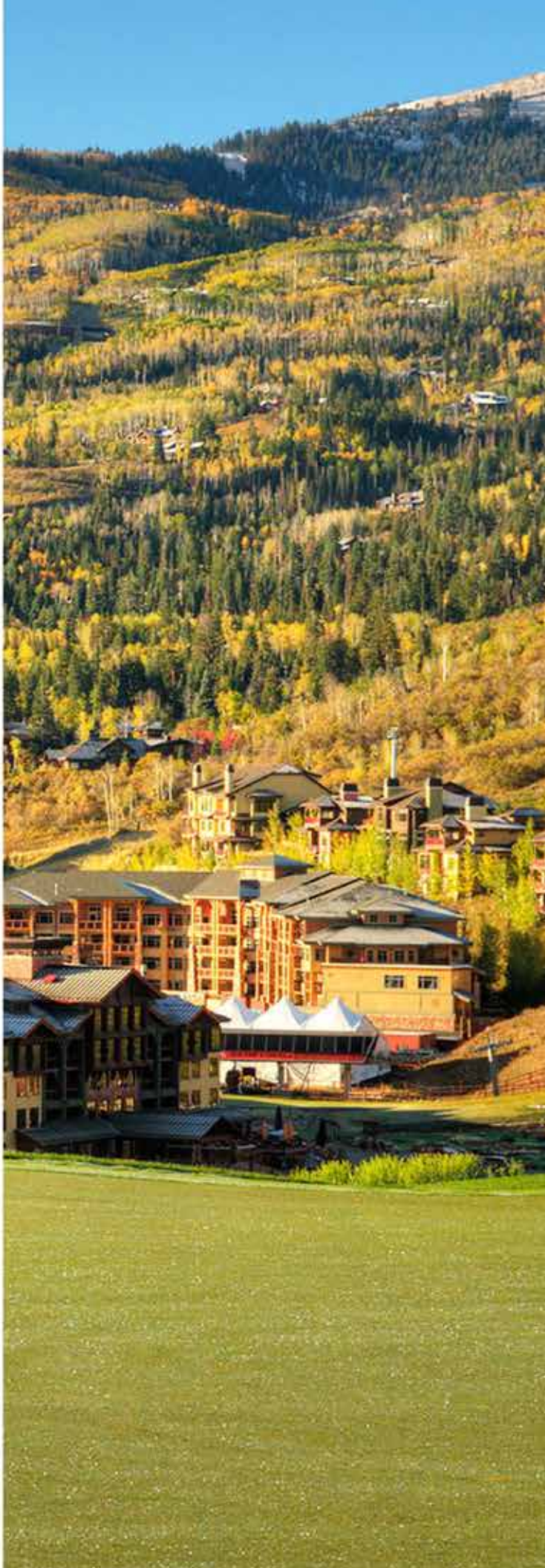
CONTINUED

3:00 - 4:00 P.M.	SESSIONS WHITE PINE, ARROWHEAD, KOKOPELLI I
4:00 - 4:15 P.M.	BREAK
4:15 - 5:15 P.M.	SESSIONS WHITE PINE, ARROWHEAD, KOKOPELLI I
6:00 - 8:00 P.M.	COCKTAILS & CONVERSATIONS POSTER EVENT SUNDIAL PAVILION

Friday, October 12th

7:00 A.M. - 2:15 P.M.	REGISTRATION & INFORMATION GRAND BALLROOM LOBBY
7:00 - 8:00 A.M.	CONTINENTAL BREAKFAST GRAND BALLROOM LOBBY
8:00 - 9:30 A.M.	SESSIONS WHITE PINE, ARROWHEAD
9:30 - 9:45 A.M.	BREAK
9:45 - 11:00 A.M.	CLOSING KEYNOTE KOKOPELLI PARLORS II & III
11:00 - 11:15 A.M.	BREAK
11:15 A.M. - 12:15 P.M.	SESSIONS WHITE PINE, ARROWHEAD
12:15 - 1:15 P.M.	LUNCH UMBRELLA BAR
1:15- 2:15 P.M.	SESSIONS WHITE PINE, ARROWHEAD

CONTENTS



WEDNESDAY, OCT 10
12 P.M. – 8 P.M.

12:00 - 1:30

WHITE PINE BALLROOM

RESPONSE SIMILARITY ANALYSIS USING R

Cengiz Zopluoglu, University of Miami
Workshop

In an era of high-stakes testing, maintaining the integrity of test scores has become an important issue and another aspect of validity. A search on the web with words "test fraud" and "cheating" reveals increasing numbers of news stories in the local and national media outlets, potentially leading to less public confidence about the use of test scores for high-stakes decisions. The scholarly community has developed a variety of best practices for preventing, detecting, and investigating testing irregularities to address the increasing concerns about the integrity of test scores.

One aspect of investigating testing irregularities is screening item response data for unusual response similarity among test takers. There are many methods that are proposed in the literature to identify unusual response similarity such as the Omega index (Wollack,1996), the K index and its variants (Sotaridona & Meijer, 2002, 2003), the Generalized Binomial Test (GBT; van der Linden & Sotaridona, 2006), and the M4 index (Maynes, 2017). This workshop will introduce an R package and complementing R code to compute the response similarity indices listed above (Omega, GBT, M4, K, K1, K2, S1, S2) on an empirical dataset provided by Cizek and Wollack (2017).

Participants will have hands-on experiences in:

- computing the response similarity indices for a pair of examinees
- computing the response similarity indices for all pairs of examinees in a test center
- flagging test centers with exceeding number of pairs with unusual response similarity

1:45 - 3:15

WHITE PINE BALLROOM

THE HISTORY AND FUTURE OF THE SELECTED RESPONSE FORMAT AND ITS EFFECTS ON TEST SECURITY

David Foster, CEO, Caveon Test Security
Workshop

The traditional multiple choice question was created in 1915 as an invention of Frederick James Kelly in order to reduce scoring errors and bias, and to standardize the testing process. It was then picked up by the team that created the Army Alpha test in 1917. Since then, a few varieties of the MC item type were tried out, but the original version, created more than 100 years ago, remains the dominant type today, for better or worse. With the introduction of computing devices for testing, many new possibilities now exist for modifying the multiple choice question, better understood as a selected response variety, in order to solve important problems with today's assessments. Security of exams is certainly one of the most important topics today, as exams continue to play an important role in all areas of society. How can the selected response format be modified to improve security? How can computer technology assist in that effort? How do such changes impact the quality of assessments? What research exists to inform our efforts? How do we balance the trade-off between changing how we create and administer exams and improving the validity of our use of test scores?

This workshop explores these issues while providing new learning experiences, including some hands-on work in developing new types of items, experiencing innovative selected response items such as SmartItems, participating as research subjects, and reviewing research evidence. You will need to bring a device such as a smartphone, tablet or laptop.

Come and join this exciting look into a possible future of tests, testing, and test security.

WEDNESDAY, OCT 10
12 P.M. – 8 P.M.

3:30 - 5:00 *WHITE PINE BALLROOM*

WHEN BAD THINGS HAPPEN TO GOOD PROGRAMS

Rachel R. Watkins Schoenig, Cornerstone Strategies | Camille Thompson, The College Board | Jennifer A. Semko, Baker McKenzie
Workshop

Even the most secure programs will eventually need to grapple with exam security incidents and, more specifically, the individuals involved. Responding to such incidents raises a whole range of questions for testing organizations. What type of notice should an examinee be given? Should other parties involved in an incident be named or kept confidential? Is there a specific process that should be afforded the examinee? Is there an appropriate range of consequences that can be applied? The answers to these questions can have both legal and practical implications. Join us for an engaging and educational workshop, as seasoned legal professionals provide meaningful guidance on current practices and considerations in this area.

5:00 - 6:00 *WHITE PINE LOBBY*

MEET & GREET

Light hors d'oeuvres provided

6:30 - 8:00 *THE CABIN BOARDROOM*

COTS EXECUTIVE COMMITTEE MEETING

ENJOY PARK CITY

THURSDAY, OCT 11
8 A.M. – 8 P.M.

Conference On Test Security 2018

9:30 - 9:45 *GRAND BALLROOM LOBBY*

CONTINENTAL BREAKFAST

THURSDAY, OCT 11

OPENING KEYNOTE

8:00 – 9:30 AM | KOKOPELLI PARLORS II & III

"MAKE HER STOP COPYING ME!"

(AND OTHER THINGS TESTING PROGRAMS SAY TO THEIR MOTHERS)

Jennifer A. Semko, Baker McKenzie

They say imitation is the sincerest form of flattery. But when it comes to test content, "imitation" can be downright insulting. Speaker Jennifer Ancona Semko will take a look at an important topic—using the U.S. copyright laws to protect your program's intellectual property. Her presentation will cover the basics of copyright law, including an examination of some of the most interesting copyright infringement lawsuits of recent years. The session will also provide an update on the U.S. Copyright Office's efforts to modify its "secure test" registration process and will share anecdotal tales of testing programs' most recent efforts to register their test content.

THURSDAY, OCT 11
8 A.M. – 8 P.M.

9:30 - 9:45

BREAK

9:45 - 10:45

WHITE PINE I

TEST SECURITY: THEN AND NOW

Jamie Mulkey, Caveon Test Security | Victoria Quinn-Stephens, Cisco | Jennifer Semko, Baker McKenzie
Panel

In 2005, best practices in test security were in their infancy. Many programs were having test security issues, yet they were not well-equipped to manage them. At the ATP conference that same year, a workshop called, 'DIY Test Security' was conducted to an interested group of testing industry professionals, eager to learn and share what they knew to be best practices in securing exams and mitigating test theft. Fast forward to 2018. As an industry, we know so much more about how to protect our assessments. New tools, strategies and technological advancements have enabled testing programs to outwit and outsmart cheaters and test thieves. Developed protocols and processes have allowed testing organizations to prioritize test security activities and establish resources to manage the test security function. Join us as we look at the advancements of test security over the last 13 years. We will discuss what's changed, how test security has matured as a discipline, and what those changes mean to you as a testing professional.

WHITE PINE II

ADAPTING YOUR TEST SECURITY PROGRAM FOR NEW PRODUCTS AND MODES OF DELIVERY

Emily Scott & Jennifer Geraets, ACT, Inc.
Demonstration

The Assessment Industry is changing, and it can be hard to keep up. Your Test Security Program must be agile enough to adapt to new products and new delivery modes. This presentation will advise you on considerations and process recommendations to grow and develop your security program to encompass new products or modes. We will walk you through how to leverage your current test security procedures as a starting point and what questions to ask that might necessitate broader or more fundamental changes. We will also advise on how to create a test security program framework that allows for the flexibility needed to adapt to ongoing changes in products and modes.

THURSDAY, OCT 11
8 A.M. – 8 P.M.

9:45 - 10:45 *CONTINUED...* *KOKOPELLI PARLOR I*

A TALE OF TWO INVESTIGATIONS: WHEN SEEMINGLY UNRELATED INVESTIGATIONS CONVERGE INTO ONE LENGTHY COURT BATTLE

Alana Chamoun, Conference of State Bank Supervisors (CSBS) | Benjamin Hunter, Caveon Test Security | Cary Straw, Caveon Test Security
Panel

Many Investigations begin with a small triggering event. In this session, presenters from Caveon and the Conference of State Bank Supervisors will delve into the curious case of how routine, ongoing web monitoring detected what appeared to be item harvesting and distribution of test content by a candidate, but turned out to be something far more than anybody ever imagined!

ARROWHEAD

IMPROVING ANSWER-COPYING INFERENCES THROUGH BAYESIAN ANALYSIS

Dennis Maynes, Caveon Test Security

Answer-copying inferences are typically made using source-copier statistics, such as Wollack's omega. These statistics statistically measure the extent of anomalous agreement between the source (i.e., the individual who was presumed to be copied from) and copier (i.e., the individual who was presumed to copy) conditioned upon the null hypothesis (e.g., the tests were taken independently). While these approaches have been very beneficial for test security practitioners, it has been suggested that these probabilities are only indirectly related to the hypothesis of interest, which is the hypothesis that the copier misbehaved.

This research addresses the hypothesis the copier misbehaved by forming the likelihood ratio test between the null and the copying hypotheses, conditioned upon the observed data. Wollack's omega statistic is used in this study because of its known power and acceptance within the research literature. To form the likelihood ratio test, a generative mixture model is proposed as a reasonable statistical model for the observed data conditioned upon the copying hypothesis. The properties of the likelihood ratio test using the models of independence and the generative mixture model are evaluated using simulation methods to demonstrate estimation bias of the copying effect and the distribution of the maximum likelihood ratio when each condition is correct.

THURSDAY, OCT 11
8 A.M. – 8 P.M.

Conference On Test Security 2018

10:45 - 11:00

BREAK

THURSDAY, OCT 11
8 A.M. – 8 P.M.

11:00 - 12:00

ARROWHEAD

DEVELOPMENT AND USE OF STANDARDS FOR THE PREVENTION AND DETECTION OF TEST FRAUD IN ONLINE TESTING ENVIRONMENTS

Rebecca Rust, Rosetta Stone
Facilitated Roundtable

An informal dialogue about the development of online tests in open testing platforms. We will discuss how to identify potential test fraud and integrate prevention techniques based on your previous test data. I'll share stories of my experience working in a fast-paced environment, and what it's like to have the ability to try out new ideas, evaluate tactics, and pivot in different directions when researching new test behaviors. Some common themes we'll talk about are timing tactics, randomization of answer options, viewing order of answer options, commitment questions, ways to cut down copying of test items, and creating standards for item viewing and test completion. This session strictly covers tests in an online environment, on desktop computers, and may benefit only those who have open testing platforms (by 'open' I mean the ability to change the UI/UX of the testing platform). But, let's chat about it to find out!

WHITE PINE 1

HOW A SIMPLE TOOLKIT OF ONLINE INVESTIGATIVE TECHNIQUES CAN DRIVE MULTIPLE OUTCOMES FOR INCREASED PROTECTIONS

Bryan Friess, Pearson VUE | Brent Morris, Cisco
Panel

If you've never seen the TV show "Catfish" on MTV, it's an episodic series where two videographers assist people to make connections with those whom they've met online but haven't yet met face-toface. The show usually takes a few crazy twists and turns as the hosts help their guests discover if their online friend is even real. But did you know there are teams of people in the test publishing industry who routinely conduct similar online research as part of their investigation strategy for program protection? In this session, two collaborating exam security experts will walk you through a case study showing how their coordinated strategies to acquire online information helped uncover harmful activity and the people tied to it for beneficial effect. After all, something as small as a message board post can snowball into an important investigation, leading to corrective action, enforceable outcomes and a host of program improvements that otherwise may not have occurred.

WHITE PINE II

30 SECONDS TO LAUNCH: HOW OKTA TOOK AN OUT-OF-THE-BOX IDEA AND BLAZED THEIR OWN CERTIFICATION TRAIL

Kpayah Tamba, Okta | Benjamin Hunter, Caveon Test Security
Panel

With a modest budget, executive support, and only the clothes on their backs, the education leaders at Okta struck out for certification gold in early 2016. Through research and happenstance, they were able to create partnerships pioneer a new approach to high-stakes certification testing using online remote proctoring, and an unfamiliar but security-focused question type. With a short runway of three months from program inception to beta testing at a live-test even in August of 2016, these industry leaders took control of their program and in April of 2018, were recognized for their testing program as the 2017 CEdMA Innovation award winners!

THURSDAY, OCT 11
8 A.M. – 8 P.M.

11:00 - 12:00 *CONTINUED...* *KOKOPELLI PARLOR I*

AN INTRODUCTION TO ITEM PRE-KNOWLEDGE DETECTION WITH REAL DATA APPLICATIONS

Carol Eckerly, Educational Testing Service (ETS) | Russ Smith, Alpine Testing Solutions | Yi-Hsuan Lee, ETS
Demonstration

Testing organizations are becoming increasingly aware of threats to the validity of test scores due to examinees having unauthorized pre-knowledge of exam items. Item pre-knowledge may be obtained in a variety of ways, including but not limited to: examinees sharing what they remember from taking exam with others who have yet to sit for the exam; educators coordinating with examinees and creating banks of recalled items; or examinees purchasing full or partial copies of exam content from a brain dump website. In some cases, examinees may gain access not only to the item content, but also to an associated key (which may or may not be accurate). Given that item pre-knowledge can take many different forms, and thus will manifest itself in response behavior in differing ways, no one detection method will represent an optimal strategy in all circumstances. This demonstration presents several existing methods designed to address differing aspects of item pre-knowledge and applies these methods to real data from three IT certification programs that experienced a known compromise. The comparison of methods among the three data sets will illustrate the circumstances for which each method performs best and situations in which particular methods may not detect the aberrant behavior. Emphasis will be placed on interpretation of the results.

12:00 - 1:00 *UMBRELLA BAR*

LUNCH

THURSDAY, OCT 11
8 A.M. – 8 P.M.

1:00 - 2:30

ARROWHEAD

YOUR VENDOR AND YOU: COLLABORATING WITH VENDORS TO IMPROVE EXAM SECURITY

Bryan Friess, Pearson VUE | Rachel R. Watkins Schoenig, Cornerstone Strategies | Camille Thompson, The College Board | Michael Clifton, ACT
Panel

Working with assessment development and delivery vendors provides benefits to your programs while presenting different exam security risks. What can testing programs do to ensure the best possible relationship with their vendors? What are some common areas of concern that can be addressed up front to ensure a more effective working relationship? What tools can help ensure a solid vendor relationship that can weather exam security breaches? Join program and vendor professionals as they discuss vetting and due diligence practices, contract considerations, service level agreements, incident response plans, and more. You'll come away with practical tools for improving your vendor relationship and better protecting your program. This is a practical, hands-on session you won't want to miss!

WHITE PINE I

ENHANCING TEST SECURITY FOR STATE ASSESSMENT PROGRAMS – A STATE PANEL DISCUSSION

John Olson, Caveon Test Security | Jessica Fenby, Michigan Department of Education | Craig Walker, Oklahoma State Department of Education | David Ragsdale, Massachusetts Department of Elementary and Secondary Education | John Fremer, Caveon Test Security
Panel

Many states have taken various measures to enhance the security of their assessment programs and improve the integrity of testing procedures. Many factors have caused states to implement stronger and more comprehensive test security policies, procedures, and practices. Comprehensive state strategies include a variety of approaches that focus on four key aspects of test security--prevention, detection, follow-up investigations, and deterrence. This framework is useful to many states and is a good way of organizing security activities into an overall vision of testing integrity. Three states that have been successful in implementing a variety of enhancements to their test security will participate in a panel discussion. These states will discuss the various steps they've taken to implement a comprehensive strategy for a strong test security system that includes developing a strategic vision, creating a clear communications plan, standardizing plans for conducting investigations, implementing a secure CBT system, making site visits to check on test administrations, monitoring social media and the Internet, and using data forensics. The state panel will be facilitated by an expert on state assessments and current approaches to test security in K-12 education. An esteemed test security expert will serve as discussant and provide recommendations on the types of models that can best aid states in enhancing the security of their assessment programs.

THURSDAY, OCT 11
8 A.M. – 8 P.M.

1:00 - 2:30 *CONTINUED...* *WHITE PINE II*

STANDARD PRESENTATIONS

Memorizing Your ABCD's: Detection of Item Pre-Knowledge and Application for Future Item Development
Brooke Dresden & Nicole Tucker, PSI Services LLC

In the world of testing, fraudulent behavior is of constant concern. This is especially true of high-stakes licensure tests, which carry important consequences for both the individual test taker and the public with whom they interact. For the purposes of this study, item response and response time data from two six-month periods in a high-stakes state licensure test, as well as separate analyses on simulated data, were used in order to detect item pre-knowledge. Additional qualitative analyses were conducted on items identified as potentially compromised in order to better understand factors which may have made an item more "memorable". The results of this study inform future potential uses of item response and response time data for the identification of item pre-knowledge, as well as identifying situations where further investigation may be needed. Furthermore, the results of the qualitative analyses provide direction for future item writing.

A Study of Modern Detectors of Examinees with Pre-Knowledge Using Real, Marked Data
Sarah L. Toton, Caveon Test Security | Dmitry I. Belov, Law School Admission Council (LSAC)

This is a study comparing modern methods for detecting examinees with pre-knowledge using real data. An experiment was conducted to manipulate pre-knowledge of compromised items. In the control condition, participants simply took a test, but in the experimental conditions, some of the items (called compromised items) were given to the participants before the test. In one condition, only the items were given to participants and in another, the correct answers were also given. The resulting real data were used to determine the power of several methods for detecting examinees with pre-knowledge. First, the baseline case in which compromised items are assumed to be known was examined. Three detectors, based on lz, the posterior shift, and the Neyman-Pearson lemma, were compared. Then, the realistic case in which compromised items are unknown, but a subset of uncompromised items is known, was examined. A recently developed method, which injects a statistic (that measures a score gain from one item subset to another) into a specially organized Markov Chain Monte Carlo (MCMC) to detect examinees with pre-knowledge, was studied under a variety of conditions. Statistical power of all detectors was strongest when the data only contained the control condition and the condition where both the items and answers were compromised, as would be expected. The results show that when information about compromised items is missing or incomplete, the MCMC detector outperforms other detectors in real data, without the strict assumptions of simulations.

Comparison Study of Item Pre-knowledge Detection Statistics in Multi-Stage Testing Environments
Xinhui Xiong, AICPA | Dmitry I. Belov, Law School Admission Council (LSAC)

Item pre-knowledge describes a situation in which a group of examinees have had access to some items (called compromised items) from an administered test prior to the exam. Item pre-knowledge negatively affects both the corresponding testing program and its users (e.g., universities, companies, government organizations) because scores for examinees with pre-knowledge are invalid and bias is introduced into item parameter calibration from the aberrant responses. In general, item pre-knowledge is hard to detect due to multiple unknowns: unknown groups of examinees accessing unknown subsets of items that were exposed in prior test administrations. Belov (2016) demonstrated that even when compromised items are not known precisely multiple statistics are able to detect examinees with pre-knowledge. This research is extending Belov's study to multi-stage testing (MST). Several modern statistics will be adapted to MST format and studied using real and simulated data.

The Modified Signed Likelihood Ratio Test and its Application to Detect Cheating
Sandip Sinharay, Educational Testing Service (ETS) | Jens L. Jensen, Aarhus University

Among the six categories of statistical methods to detect cheating on tests mentioned by Wollack and Schoenig (2018), one is "score differencing"—this category of methods essentially involves a test of the hypothesis of equal ability of an examinee (or a group of examinees) over two sets of items, typically against a one-sided alternative hypothesis. Depending on the application, the two sets of items could represent items administered at two time points, compromised items and non-compromised items, items with erasures and items with no erasures etc. Barndorff-Nielsen (1986) suggested the modified signed likelihood ratio test (MSLRT) for testing the equality of two parameters. The MSLRT has since been applied in several areas and has been found as powerful as or more powerful than other alternatives; however, the MSLRT has lacked application in educational measurement. Because score differencing involves the equality of two (ability) parameters, it is possible to apply the MSLRT in score differencing. This paper will discuss how one can perform the MSLRT for testing the equality of two ability parameters in the context of IRT models and prove that the asymptotic null distribution of the test statistic is standard normal. A simulation study will be performed to demonstrate that the Type I error rate of the MSLRT is close to the nominal level and the power is larger than other alternatives. The MSLRT will be applied to a real data set.

THURSDAY, OCT 11
8 A.M. – 8 P.M.

1:00 - 2:30

CONTINUED... *KOKOPELLI PARLOR I*

TEST SECURITY: WILL IT MATTER IN 2023?

Brian Adams, Alpine Testing Solutions | David Foster, Caveon Test Security
Facilitated Roundtable

You notice a smile as the passenger next to you glances at a vibrating watch.

"Good news?" you ask.

"Yes," the passenger says, "I just received on update on my progress towards a credential. It is good news, but I smiled because I was reminded of a conversation with my parents. I was submitting my registration for the credentialing program which, among other things, required me to authorize the program to collect, process, and store all sorts of past, current, and future data relevant to measuring me against the requirements to earn and maintain the credential. I remember my parents laughed and said to one another, remember when measurement required us to take tests?" The passenger continued, "I can't imagine how inconvenient that must have been, not to mention the challenge of really representing that which the test was trying to measure."

Now it is your turn to smile as you think back to the days in which you, in your role in the assessment industry, struggled with the challenges of building, validating the use of, and maintaining tests. The ability to measure without testing certainly changed the assessment industry and resulted in numerous benefits ranging from convenience, to public knowledge and acceptance of measurement, to significant increases in the level of confidence we place in the interpretation and use of measurement data. You also think of the security challenges that simply ceased to exist with the move to measurement without testing. But then the smile leaves your face as you wonder which of today's security risks might have benefited the passenger sitting next to you, and how you and your colleagues in the measurement industry will mitigate those risks.

This session picks up where the story leaves off. A round table will explore the idea of measurement without testing with a focus on related security concerns. Specific questions put to the panel will include:

- How might we measure without testing?
- Which of today's security risks would cease to concern us?
- Which of today's security risks would continue to concern us, and how might we address them?
- What new security risks would emerge, and how might we address them?

2:30 - 3:00

EXTENDED BREAK

l.ead.me/cots18

THURSDAY, OCT 11
8 A.M. – 8 P.M.

3:00 - 4:00

KOKOPELLI PARLOR I

A PANEL ON SMARTITEMS AND THEIR SURPRISING EFFECTS ON TEST SECURITY

David Foster, CEO; Sarah L. Toton; Alison Foster Green; and Tara Williams – Caveon Test Security
Panel

This session provides several papers/presentations on the SmartItem, a testing technology recently introduced by Caveon. SmartItems are designed and built to solve security and other practical testing problems by covering the entire breadth of a skill or objective of interest, and by generating a massive number of item variations, presented individually and uniquely to test takers. This SmartItem symposium will be interactive. After introducing the audience to the SmartItem, speakers will give demonstrations, share key findings from their research, and more! There will be opportunities to ask questions and engage in discussion on this revolutionary and useful new testing technology.

ARROWHEAD

NETWORK ANALYSIS FOR TEST SECURITY INVESTIGATIONS

Joe Grochowalski, The College Board
Demonstration

In this presentation, we use features of network analysis to investigate suspected test taker collaboration in a large-scale assessment. We introduce network analysis with basic analytic concepts and illustrate how collaboration can be identified and tested. The methods discussed include how to form and interpret networks, how to test for associations, and how to limit the false positive detection rate. These methods are illustrated using the iGraph package in R, and we demonstrate the interpretation of output along with various methods for visualizing networks using iGraph. We demonstrate the methods and statistical programming with simulated large-scale assessment data, which illustrate how suspected networks appear in an applied setting.

WHITE PINE I

CREATING A RESOURCE FOR STATE ASSESSMENT TEST SECURITY POLICIES AND PROCEDURES -- THREE EXAMPLES FROM STATES

John Fremer, Caveon Test Security | Jessica Fenby, Michigan Department of Education | Chris Seay, South Carolina Department of Education | Patsy Kenner, Kentucky Department of Education | John Olson, Caveon Test Security
Panel

Many states know that a Test Security Handbook is a valuable tool that serves as a resource for test security purposes. Handbooks can be over-arching documents containing all DOE security-related procedures, processes, and regulations, including the escalation path to be followed in the event of a test security breach. A well-designed, comprehensive state Test Security Handbook will provide:

- An electronic document based on results of a comprehensive Test Security Audit, review of DOE documents, and interviews.
- A single, comprehensive source defining all state procedures, processes, and regulations.
- A flexible means to document the most critical processes related to DOE testing activities.
- A historical summary of testing irregularities, if any, and the follow-ups done to remedy them.

Continued on next page...

THURSDAY, OCT 11
8 A.M. – 8 P.M.

3:00 - 4:00 *CONTINUED...* *WHITE PINE I*

CREATING A RESOURCE FOR STATE ASSESSMENT TEST SECURITY POLICIES AND PROCEDURES -- THREE EXAMPLES FROM STATES

John Fremer, Caveon Test Security | Jessica Fenby, Michigan Department of Education | Chris Seay, South Carolina Department of Education | Patsy Kenner, Kentucky Department of Education | John Olson, Caveon Test Security
Panel

Continued... In this session, representatives from three states that have developed, or are in the process of developing, test security handbooks (each in a slightly different way) will describe the models they used in creating their handbooks and how they are being used in their state. The state presenters will provide examples of how they are using the contents of the handbook to implement specific security practices in the state, e.g., monitoring schools and making site visits to check on test administrations, implementing training procedures for conducting investigations, and/or monitoring the Internet/websites for compromised test items.

A renowned expert on test security will serve as the Chair. The Discussant, an expert on state assessments, will provide feedback on the types of handbooks states are using and how to improve them.

WHITE PINE II

STANDARD PRESENTATIONS

The Prevention and Detection of Test Fraud: Two Empirical Studies on Educational Data Forensics
Sebastiaan de Klerk & Kees Boonman, Xquiry

In this session, we present the results of two research studies on the prevention and detection of test fraud. The first study is a quantitative empirical study on the reliability of two educational data forensics parameters: the Guttman error person-fit parameter (Meijer, 1994), and a lognormal response time model parameter (Van der Linden, 2006). We analyzed and compared aberrant response patterns for these parameters for five experimental conditions of participants instructed to cheat on a test (i.e., collusion, proctor assistance, cheat sheet, smartphone use, and pre-knowledge) and a control group of honest test takers. Results show that combining the person-fit parameter and the response time parameter leads to the most optimal judgment. Although the detection ratio was still rather low - 38% of instructed cheaters were detected -, the reliability was high - the true positive ratio was 97%. This means that detected test takers are very likely to have cheated. The second study is a qualitative study an educational data forensics protocol. The protocol is a self-developed appraisal system consisting of evidence-based guidelines on the prevention and detection of test fraud. A prototype of the protocol was subject of eight semi-structured subject matter expert interviews. Furthermore, the protocol has been applied in a Dutch testing organization. Based on the interviews and the test case, the prototype has been adjusted into its final form. The results will be presented.

Integrating Multiple Sources of Evidence in Test Security Analyses: Using Bayesian Inference to Weight the Strength of Evidence and Make Robust Decisions
William Skorupski, ACT | James Wollack, University of Wisconsin – Madison | Sonya Sedivy, University of Wisconsin – Madison

In test security analyses, we are often faced with incomplete evidence from various sources. Some results may be contradictory, while other sources of evidence may complement one another. Test security evidence may also come in different forms, such as statistical analyses, proctor reports, eyewitness testimony, circumstantial evidence, etc. These sources of evidence are not consistently reliable and may focus on different, unrelated aspects of security (e.g., a statistical erasure analysis will not necessarily be related to answer copying behavior or aberrant growth analyses). The absence of scientific methods to aggregate evidence from various sources frequently leaves testing entities to make arbitrary decisions with regard to whether a testing anomaly occurred. The following paper proposes a fully Bayesian Inference Network for combining multiple sources of test security evidence of various kinds, assigning prior weights for the strength of these evidences, and combining all information into a posterior probability. The posterior probability and Bayesian odds ratios may then be used to make robust inferences with regard to the probability of a genuine test security violation, given the available evidence.

THURSDAY, OCT 11
8 A.M. – 8 P.M.

4:00 - 4:15

BREAK

4:15 - 5:15

ARROWHEAD

STORIES FROM THE FRONT LINES: A PANEL DISCUSSION ON IMPLEMENTING ITEM PROTECTION STRATEGIES IN IT CERTIFICATION PROGRAMS

Janet Lehr, Hewlett Packard Enterprise | Beverly Bone, IBM | John Sowles, Ericsson | Scott Thayn, Certification Management Services | Jamie Mulkey, Caveon Test Security
Panel

In its most elemental form, the basic principle of securing a program's exams comes down to protection. What happens when the item theft and cheating are so prevalent, it makes it nearly impossible to protect your most prized possessions? You turn to use items that are virtually cheat-proof; items and test delivery strategies that make cheating a thing of the past.

Join us for this panel discussion as three different IT (Information Technology) certification programs discuss their journeys in designing, creating, and implementing protective item strategies for high stakes assessments. This panel will provide an engaging, interactive discussion of considerations for addressing the problems of item exposure and fraudulent test taking using innovative item protection methodologies.

WHITE PINE I

TESTWISENESS AND CHEATING: TWO RELATED SCOURGES AFFECTING THE VALID USE OF TEST SCORES

David Foster, CEO, Caveon Test Security

Both testwiseness and cheating are test taker behaviors that negatively affect our ability to use test scores properly. Both add significant amounts of unrelated variance to test scores, and both can be significantly reduced using new selected response item types. Scientific research results will be presented that demonstrate the ability to significantly reduce or completely remove these sources of variance, allowing test scores to used with less qualification and ambiguity. This presentation describes these two effects, shows the relationship between them, and demonstrates methods of dealing with them. Attendees will learn how relatively simple test design modifications can remove these sources of deception and confusion forever.

THURSDAY, OCT 11
8 A.M. – 8 P.M.

4:15 - 5:15 *CONTINUED...* *KOKOPELLI PARLOR I*

HISTORY OF CHEATING ON TESTS

Ray Nicosia, Educational Testing Service (ETS) | Rachel R. Watkins Schoenig, Cornerstone Strategies | John Fremer, Caveon Test Security
Panel

There is a great deal of evidence that cheating on tests goes back for many centuries, millennia even. What are the kinds of cheating that have been observed? What efforts have been made to prevent or deter cheating? How have cheaters been treated when caught? What has changed over the centuries and what remains the same? How have standards and expectations evolved in the US and elsewhere in the world? Are things getting better or worse and in what ways? Do severe penalties deter cheating? What strategies have been tried at different times and places for particular types of cheating? Noteworthy examples will be provided.

The history of cheating, whether on tests or related domains, provides us an opportunity to learn what has worked, what hasn't, and what has stood the test of time. Together, we'll take a fun and informative walk through history and apply what we learn from the past to improve our programs of today!

WHITE PINE II

STANDARD PRESENTATION

Do People with Item Pre-knowledge Really Respond Faster to Items They Had Prior Access? An Empirical Investigation.
Murat Kasli & Cengiz Zopluoglu, University of Miami

As the test security becomes a hot topic in recent years, one particular area of research is to develop methods for identifying individuals with potential item pre-knowledge. The methods that propose the use of response time information in identifying individuals with item pre-knowledge have an implicit assumption that the individuals with item pre-knowledge differ in their response time patterns compared to other individuals with no item pre-knowledge, and therefore response time brings an additional valuable information beyond the raw item responses into the analysis. However, we are not aware of any empirical research based on real data to investigate whether or not this implicit assumption is necessarily true. This is also important to inform future simulation studies. For many reasons, researchers do not have access to many real datasets to test their methods that utilize response time and simulation is a de-facto approach for researchers to test these methods. On the other hand, researchers have to have realistic assumptions while simulating their data for honest test-takers and dishonest test-takers. Do people with prior access to test items respond faster than the people with no access to the same test items? Do people with prior access to test items respond faster to items they have seen before compare to the items they haven't seen before? If so, to what degree? The purpose of current study is to investigate whether or not individuals with item pre-knowledge have different response time distribution on items they had prior access using a publicly available dataset.

Detecting Examinees with Item Pre-knowledge via Posterior Shift using Responses and Response Times
Stephen Cubbellotti, American Board of Internal Medicine (ABIM) | Dmitry I. Belov, Law School Admission Council (LSAC)

The purpose of this study is to assess whether the time spent on items can be used as an indicator function for identifying compromised items. In particular, for each examinee, the response vector (including scored responses and response times) is partitioned into two disjoint sub-vectors: responses where responses times (RTs) were normal and responses where RTs were unusually short. The proposed statistic measures a difference in performance (in terms of score and speed) between these sub-vectors. For each examinee, the difference in performance is computed as a weighted sum of the posterior shift between corresponding posteriors of ability and posterior shift between corresponding posteriors of speed. In other words, examinees that had short response times and gain scores on these items compared to other items may be flagged by the new statistic. The performance of the new statistic on simulated and real responses to a high-stakes computer-based test will be compared with other popular statistics.

Standard Presentations Continued on Next Page...

THURSDAY, OCT 11
8 A.M. – 8 P.M.

4:15 - 5:15

WHITE PINE II

STANDARD PRESENTATIONS - CONTINUED

Detecting Examinees with Item Pre-knowledge using Extreme Gradient Boosting (XGBOOST)
Cengiz Zopluoglu, University of Miami

Machine-learning methods are becoming more important and widely used in many fields. In the area of test security, there has been a relatively small number of studies that utilized these methods in identifying potential testing fraud (Sotaridona, 2001; Kim, Wooo, & Dickison, 2017; Man, Sinharay, Yao, & Harring, 2018). In the very first attempt, Sotaridona (2001) explored the performance of Backpropagation, a supervised learning algorithm, in detecting examinees with item pre-knowledge in a small simulation study as a part of his dissertation. Later, Kim et al. (2017) used another supervised learning algorithm Market Basket Analysis in identifying aberrant response patterns. More recently, Man et al. compared the performance of several data mining algorithms in detecting test fraud.

In this study, we investigate the utility of a recently developed state-of-the-art algorithm, Extreme Gradient Boosting (XGBOOST; Chen & Guestrin, 2016) in detecting examinees with potential item pre-knowledge using an empirical dataset, and also discuss the challenges that need to be addressed in using such methods. Tree boosting is a highly effective and widely used machine learning method. Chen and Guestrin (2016) further developed the traditional gradient tree boosting algorithm by introducing XGBOOST. Since its development and introduction in 2014, XGBOOST became a de-facto method for many data science competitions and won a majority of competitions. For instance, out of 29 winning solutions posted on Kaggle during 2015, 17 used XGBOOST. The current study will explore the effectiveness of XGBOOST, a machine-learning algorithm, in identifying examinees with item pre-knowledge.

6:00-8:00

SUNDIAL PAVILION

COCKTAILS & CONVERSATIONS

POSTER EVENT

Light hors d 'oeuvres will be served

You won't want to miss it!

#COTS2018

l.ead.me/cots18

FRIDAY, OCT 12
8:00 A.M. – 2:15 P.M.

7:00 - 8:00

BALLROOM LOBBY

CONTINENTAL BREAKFAST

8:00 - 9:30

WHITE PINE I

TEST SECURITY IN HIGH STAKES TESTING PROGRAMS: WHERE DO WE STAND NOW IN SOME KEY AREAS?

John Fremer, Caveon Test Security | Jennifer Semko, Baker McKenzie | David Foster, Caveon Test Security | James A. Wollack, University of Wisconsin – Madison
Panel

This session asks some experienced test security professionals to "take stock" of where the field stands at this point and where we seem to be heading. Each area will be addressed by a speaker who is not only very connected to developments in the test security field and profession, but who is personally participating in some of the major developments. The topic areas that will be addressed are the following ones:

- Test Security Program Policies and Procedures
- Test Development and Administration
- Psychometrics and Research
- Legal Considerations

Each presenter will give their perspective on the status of developments in the domains that they are addressing. In addition to observations specific to the area that they are covering, the following common framework will be used:

- What issues and approaches seem relatively settled at this point?
- Where is there uncertainty about what to do and how to do it?
- What should a program dealing with each area be aware of as major challenges and decision points that they will encounter?
- What additional research, policy development, or operational work is needed?
- What are the "hot spots" that may result in pressure being brought to bear from within the program or by test takers, the media or other sources?

Each presenter will provide their view as to most important "next steps" for professionals working in the area addressed or dependent on the results from that area.

FRIDAY, OCT 12
8:00 A.M. – 2:15 P.M.

8:00 - 9:30

CONTINUED... *ARROWHEAD*

EXAM SECURITY: THE NEXT FRONTIER

Rachel R. Watkins Schoenig, Cornerstone Strategies | Ray Nicosia, Educational Testing Service (ETS) | Camille Thompson, The College Board | Ardeshir Geranpayeh, Cambridge Assessment English
Panel

So maybe the "Men in Black" memory eraser isn't ready for market yet, but there are a range of other tech-enabled capabilities that could have a big impact on exam security in the near future. Join a panel of experts as they explore the potential impact of Artificial Intelligence, thermal imaging, new biometrics, crowd sourcing, and other technologies on cheating detection and prevention. There are a host of new and emerging technologies that will help testing programs more effectively secure their exam. Want to learn about the next frontier in exam security? Make sure you attend this session!

WHITE PINE II

STANDARD PRESENTATIONS

Application of a Sequential Procedure for Detecting Compromised Items to a CAT Licensure Exam
Chansoon (Danielle) Lee, & Hong Qian, National Council of State Boards of Nursing (NCSBN)

In Computerized Adaptive Testing (CAT), test items are reused over time and under the risk of being compromised despite efforts to control item exposure rate and test overlap rate. The increasing number of compromised items in test can threaten the reliability and validity of the test. To detect compromised items in a real CAT item pool, this study used a sequential procedures based on Classical Test Theory (CTT; Zhang, 2014) and based on Item Response Theory (IRT; Zhang & Li, 2016). The sequential procedure, as a real-time monitoring procedure in CAT, examines changes in the individual item response function using a series of statistical hypothesis tests. The previous simulation studies showed that the procedure maintained a lower Type II error while controlling Type I error. The item pool
for the current study had 1,464 items, which were administered to 65,753 examinees using CAT in a licesure testing organization. This research explored (1) whether the sequential procedure is applicable for the real item pool administered in CAT and (2) whether the sequential procedure identifies any compromised items in the item pool. The results of this study could help practitioners choose an appropriate data forensic method while providing useful information about compromised items for test developers and stakeholders.

Compromised Item Detection Using Item Response and Response Time
Chunyan Liu, Daniel Jurich, & Kimberly A. Swygert, National Board of Medical Examiners (NBME)

To maintain the validity of a continuously-administered exam, especially when there are high stakes for examinees, test developers should monitor items regularly to ensure they are not over-exposed or compromised. Examinees with pre-knowledge of compromised items tend to perform better and respond more quickly than those unexposed to the items previously. Responses from examinees with pre-knowledge likely do not reflect their true ability on the construct being assessed, so monitoring item performance to identify potentially compromised items provides crucial evidence to support the fairness and validity of the interpretations of the test scores.

Zhang (2014) demonstrated via a simulation study that a sequential procedure based on item percent correct (p-value) is efficient in flagging compromised items in the framework of computerized adaptive testing (CAT). However, the process produces false positives, so the flag alone does not mean an item is actually compromised. Additional validation must be gained to confirm whether the item should still be used in scoring. In addition, this method has not been tested on linear (fixed-form) computer-based tests (CBTs).

This study will evaluate the utility of Zhang's sequential procedure for a high-stakes linear CBT, where item response time (RT) will serve as an additional check to help assess whether the flagged items are actually compromised. Furthermore, this study will examine the results of the sequential procedure for different examinee subgroups, particularly those at different testing centers, to investigate whether there are unusual patterns that could indicate organized attempts at item compromise or pre-knowledge.

Standard Presentations Continued on Next Page...

FRIDAY, OCT 12
8 A.M. – 2:15 P.M.

8:00 - 9:30

CONTINUED... *WHITE PINE II*

STANDARD PRESENTATIONS - CONTINUED

Identifying "Miskeyed" Items on the Fly
Xin Li, PSI Services LLC

Examination questions are developed to be psychometrically sound and legally defensible so as to provide an accurate and unbiased measure of the respondents' knowledge and capability. Approaches based on either Classical Test Theory or Item Response Theory have been widely used to flag questions with poor statistics using well-established criteria. However, the underlying assumptions for both models require correct answers so that the data can be scored as correct or incorrect responses. Cultural Consensus Theory (CCT) was developed to model the commonly shared knowledge among respondents assuming no prior information about the correct responses (Batchelder & Romney, 1988). This study adapts the General Condorcet Model (GCM) based on the CCT framework to estimate the answer keys as a set of parameters together with ability, difficulty, and guessing parameters using MCMC algorithm with raw response data (Karabatsos & Batchelder, 2003). The accuracy of recovering keys and flagging miskeys will be evaluated using both simulated and operational data. The benefit is to detect potential miskeyed questions at an earlier stage with a small sample of respondents. Thus, a routine analysis using this method can help maintain the measurement accuracy and enhance test validity.

9:30 - 9:45

BREAK

FRIDAY, OCT 12

CLOSING KEYNOTE

9:45 – 11:00 AM | KOKOPELLI PARLORS II & III

COTS KEYNOTE DEBATES

Moderator: Rachel R. Watkins Schoenig, Cornerstone Strategies

Without appropriate exam security, the trust in exam results and reputation of our programs – and our industry – is quickly eroded. While we may recognize the need for exam security, the debate concerning what constitutes "appropriate" exam security remains on-going.

Back for another year, the COTS keynote debates feature experienced professionals presenting differing views on controversial exam security topics. At turns lively, humorous, serious, and always well-informed, the debates summarize the key industry insights on a given topic.

In addition to engaging debates from testing professionals, audience members will be asked for their opinions on each topic. By the end of this keynote, you will have heard the positions of seasoned professionals and learned the collective wisdom of the crowd to help YOU ultimately decide what "appropriate" exam security means for your program.

A SPECIAL THANK YOU TO OUR CO-HOSTS

caveon™ Test Security

ascend LEARNING

CISCO

FRIDAY, OCT 12
8:00 A.M. – 2:15 P.M.

11:00 - 11:15

BREAK

11:15 - 12:15

WHITE PINE I

LEVERAGING A BIG DATA INFRASTRUCTURE TO IMPROVE TEST SECURITY ANALYSIS

Michael Clifton, John Greene, & Gavin Henderson, ACT
Symposium

As testing organizations implement infrastructures capable of analyzing streaming event data, leveraging clusters to harness computing power for complex algorithms to run at high speed, and parsing and combining high variety data, opportunities abound to improve the speed and reach of test security algorithms. This coordinated symposium will share the journey of one testing organization as it learns to leverage a big data infrastructure to improve test security analysis. The first presentation will describe the process of discovering opportunities as the test security team articulates and prioritizes business needs to infrastructure, data science, and analytic teams. The second presentation will illustrate options for big data architecture solutions to service test security needs. The third presentation will showcase forensic analyses made possible or faster via the use of the big data architecture.

ARROWHEAD

USING NUDGES TO ENCOURAGE RULE-FOLLOWING BY OUR TEST-TAKERS

Cynthia G. Parshall, Touchstone Consulting | John Fremer, Caveon Test Security
Panel

Nudges are small or subtle encouragements to help people make choices in their own best interest. A substantial body of research shows that well designed nudges can be surprisingly effective. In assessment, we can use these nudge tactics to help examinees be honest in their test preparation and test-taking.

Behavioral nudges are grounded in evidence regarding human nature. For honesty, the critical elements are two natural human desires, at war with one another. Most people want to see themselves as "a good person." However, people also want to gain all their legitimate advantages. For example, most Americans regard it as morally wrong to cheat on their taxes, but they also want to gain all the deductions and tax breaks which they are legally allowed. The internal conflict between these two desires can be thought of as each person's individual "fudge factor."

The good news is that modest changes in the environment, using carefully implemented nudges, can reduce the fudge factor. Use of these well researched behavioral tools makes it less likely for a test-taker to rationalize misbehavior, which increases the likelihood of honest test-taking. For example, moral reminders immediately before a point of "temptation" can be highly effective at reducing dishonest behavior. And, social proof can nudge test-takers towards honesty by indicating that the majority of candidates follow the rules.

In this session, we will talk about strategies that test program managers can employ to encourage following test-taking rules.

FRIDAY, OCT 12
8:00 A.M. – 2:15 P.M.

11:15 - 12:15

CONTINUED... *WHITE PINE II*

STANDARD PRESENTATIONS

Graph Theory Approach to Detect Test Collusion Using Responses and Response Times
Dmitry I. Belov, Law School Admission Council (LSAC) | James Wollack, University of Wisconsin – Madison | Stephen Cubbellotti, American Board of Internal Medicine (ABIM)

Test collusion (TC) is a sharing of test materials or answer to test questions. TC is a broad topic which includes special cases such as: item preknowledge and aberrant answer changes (aberrant erasures). There are many potential sources of shared information including: teachers, test preparation entities, the Internet, or even examinees collaborating during the exam. Because of the potentially large advantages for examinees involved, TC poses a serious threat to the validity of score interpretations; hence, accurately identifying individuals involved in collusion is important.

TC is expected to produce response similarity on common items for a group of examinees. Recently, cluster analysis and factor analysis were applied to answer similarity data for detecting TC. The proposed approach operates similarly but applies graph theory methodology for the purpose of identifying groups. The performance of the graph theory approach on simulated and real responses to a high-stakes computer-based test will be compared with other TC detectors.

Evaluation of Different Clustering Approaches in Detection of Test Collusion
Sakine Gocer Sahin & James Wollack, University of Wisconsin – Madison

Test collusion, or large-scale sharing of test materials or answers, has come to the fore in the last few decades. Examples of test collusion include educators coaching students on or groups of students communicating about live test content, or the systematic changing of students' answers to test questions in an effort to improve their scores. In spite of the attention this type of cheating has received in the media, research on methods to detect test collusion remains relatively scarce. Wollack and Maynes (2017) developed a cluster analytic approach which appears promising; however the single linkage method they used has a tendency to combine clusters or to chain examinees together into one large cluster. This study extends Wollack and Maynes by investigating the utility of several hierarchical clustering methods, such as complete linkage, single linkage, average linkage, centroid method, and Ward's method. A simulation study was conducted in which test length, proportion of contaminated items, group size, and magnitude of contamination were all manipulated. For clustering purposes, van der Linden and Sotaridona's (2006) Generalized Binomial Test (GBT) statistic was used to quantify the extent to which two response vectors were similar. Type I error rate, power, and cluster integrity were computed to evaluate the results of each method.

On a New Method for Removing Noisy Data in Similarity Analysis
Zhongmin Cui, ACT

In test-security analyses, answer copying between two examinees can be detected by examining the similarity in item responses between them. The similarity, however, can possibly be inflated by noisy data resulted from item responses by unmotivated examinees. Low motivation usually happens when examinees have no interest in a test but must take the test because of mandatory requirements. These examinees may answer by repeating a pattern (e.g., ABCDABCDABCD or AAAAAAAA). It is important to remove this kind of noisy data before conducting any similarity analysis. One way to identify unmotivated examinees is to compute a person-fit statistic. However, person misfit can arise for reasons other than motivation (Meijer, 1996). As a result, removing data based on a person-fit index may lead to the removal of non-noisy data. In this study, a new data-cleaning method based on finding repeated patterns is proposed, and its performance was evaluated through a simulation with different conditions.

12:15 - 1:15

UMBRELLA BAR

LUNCH

FRIDAY, OCT 12
8:00 A.M. – 2:15 P.M.

1:15 - 2:15

WHITE PINE I

THE HIGHER STAKES OF ONLINE HIGHER ED: SHE SAID, HE SAID...

Cary Straw, Caveon Test Security | Christine Gee, Western Governors University (WGU)
Panel

Online higher education is a big deal and getting bigger all the time. The need for individuals to gain important knowledge and skills quickly, flexibly, and effectively continues to grow, particularly as forecasts of technology changes require individuals to tool up quickly to compete in a global economy.

Join our two speakers as they discuss the test security challenges. First, Christine Gee will discuss Western Governor's University's test security challenges and best practices. Then, Caveon's Cary Straw will discuss how other online higher education program address these same challenges. Christine will talk about the solutions WGU currently has in place and how their future test security needs are being anticipated and addressed. Cary will discuss how other programs do this as well. Participants in this session will be able to apply anecdotal information learned back to their own testing programs.

WHITE PINE II

DEVELOPMENT OF A DATA FORENSICS SYSTEM FOR SURVEILLANCE AND AD HOC INVESTIGATIONS OF INTERNATIONAL TESTING PROGRAMS

Greg Hurtz, John Weiner, & Zuru Du, PSI Services LLC
Demonstration

Our organization has had test security analytics in place for years to span the testing cycle, from the protection of content in test development, delivery, and administration, to a collection of data forensics tools to guide the analysis of candidate response and exam time data. In recent years we have carried out a considerable amount of research to advance our practices in statistical detection of anomalies in test results and other test security practices. As we continue to support an increasing number of global examination programs, the need has increased for a uniform and centralized approach to test security and data forensics as a quality assurance and control program for test administration sites, and for detection (through routine surveillance) or corroboration (through triggered investigations) of potential test fraud incidents. In this session we will discuss the development and implementation of our system which includes multiple measures and multiple levels of analysis to detect multiple patterns of potential test fraud. We will discuss some lessons learned and recommendations gleaned from our experience in conceptualizing and developing a system that is technically sound while attending to end-user needs and comprehensibility by multiple stake-holders.

ARROWHEAD

A BRIEF HISTORY OF (RESPONSE) TIME

Marcus Scott, Sarah L. Toton, & Dennis Maynes, Caveon Test Security
Demonstration

The advent of computer-based testing has led to many testing advancements. One such advancement is the collection and use of response time data, particularly for detecting examinees who may have gained an unfair advantage, such as through item pre-knowledge or the use of a proxy test taker. This talk will cover the history of methodologies using item response times to detect examinees who may have had an unfair advantage. First, methods for modeling response times will be presented. This will include a discussion on the suitability of lognormal models for modeling response times, the effective response time developed by Meijer and Sotaridona, and van der Linden's model that incorporates item time intensity and examinee speed. Second, response time-based methods for detecting examinees who may have gained an unfair advantage will be examined. These methods include simple timing thresholds, aberrance statistics, and combinations of response time statistics with one another and other testing variables. We will discuss the strengths and weaknesses of each method and future directions for improving these methodologies.

CONFERENCE MAP

GRAND SUMMIT
A ROCKRESORT

PARK CITY MOUNTAIN

SHOULD YOU HAVE ANY QUESTIONS, PLEASE CALL 801.520.8758
OR THE GRAND SUMMIT HOTEL 435.615.8040

GENERAL INFORMATION

CONFERENCE URL

conferenceontestsecurity.org

COTS SOCIAL MEDIA

#COTS2018
LinkedIn: http://bit.ly/cotslinkedin
Twitter: http://bit.ly/cotstwitter

CONFERENCE APP

Once again, COTS is happy to provide electronic acess to the program through the conference app. Instructions for downloading the app may be found at the registration desk.

PRESENTER INFORMATION

If you are scheduled to present at COTS, please go to bit.ly/cotspresenter to view important updates and information.

FOOD ALLERGIES AND DIETARY RESTRICTIONS

All food allergies and dietary restrictions identified during the registration process have been communicated to the catering staff. If at any point you should have a question about the food selection, please speak with a member from The Grand Summit Hotel, or stop by the registration/help desk.

LOCAL CONTACTS

Because it isn't possible to list every incredible event, restaurant, and attraction that beautiful Park City has to offer, please look for attendees with the yellow "local contact" ribbon on their name badge. They are familiar with the area and would be happy to help. You can also stop by the registration/help desk for additional resources.

THINGS TO DO IN PARK CITY

Park City: bit.ly/2dxrkve
Visit Park City: bit.ly/2dwttpc
Utah.com: bit.ly/mn83mv

SHOULD YOU HAVE ANY QUESTIONS, PLEASE CALL 801.520.8758
OR THE GRAND SUMMIT HOTEL 435.615.8040

SESSION INFORMATION

YOU CAN LOOK FORWARD TO THESE PRESENTATION FORMATS DURING COTS 2018

Standard Presentation: 60 or 90-minute session with 3–5 presentations on related topics. Each presentation in the session will be 15–20 minutes long.

Coordinated Symposium: Three to five separate research presentations, all focused on a common theme. One of the presentations may consist of a discussion, analysis, and/or contextualization of another session or sessions. All symposia will occupy either a 60 or 90-minute time slot.

Panel Presentation: Two to five individuals discussing different aspects of a common theme with each other and the audience. All panel presentations will occupy either a 60 or 90-minute time slot.

Demonstration: 60 or 90-minute session in which presenters demonstrate a technique or method related to a core aspect of test security.

Workshop: A 1½-hour or 3½-hour deep dive into specific topics or key security resources, which provides attendees with an opportunity to gain hands-on experience and collaborate with presenters and other attendees.

Facilitated Roundtable: 60 or 90-minute session that promotes networking and invites audience engagement around an important test security topic. Session will begin with a short, informal presentation to frame a conversation, followed by a free-flowing dialogue among audience members.

Poster Presentation: 60-minute session to include multiple posters on various topics. Each presenter will prepare a poster to fit on a 4' x 6' poster board. Poster presentations are informal and involve one-on-one interactions with many attendees.

TO VIEW THE INCREDIBLE LINEUP OF PRESENTATIONS WE HAVE AT THIS YEAR'S COTS, GO TO BIT.LY/COTSPROGRAM

SHOULD YOU HAVE ANY QUESTIONS, PLEASE CALL 801.520.8758
OR THE GRAND SUMMIT HOTEL 435.615.8040

NOTES

NOTES

NOTES

NOTES

SAVE THE DATE

THE CONFERENCE ON TEST SECURITY

OCTOBER 2019

MIAMI, FL

HOSTED BY
THE UNIVERSITY OF MIAMI

UNIVERSITY OF MIAMI

THANK YOU TO OUR SPONSORS

CO-HOSTS

caveon™ Test Security

ascend LEARNING

cisco

FRIENDS

National College Testing Association NCTA

AICPA

ACT®

NCME national council on measurement in education

psi

Pearson VUE

Alpine Testing Solutions

CollegeBoard

proctorU

PROMETRIC