THE 12TH ANNUAL CONFERENCE ON TEST SECURITY

ARIZONA STATE UNIVERSITY

OCTOBER 19-21, 2023 (IN-PERSON)

NOVEMBER 8-9, 2023 (VIRTUAL)

TEMPE, AZ

CONTENTS

AN EXTRA SPECIAL THANK YOU TO OUR HOSTS

Arizona State University

RIO SALADO COLLEGE

A MARICOPA COMMUNITY COLLEGE

A SPECIAL THANK YOU TO OUR CO-HOSTS

Pearson VUE

duolingo english test

caveon™ Test Security

MEASURE LEARNING

rosalyn

A SPECIAL THANK YOU TO OUR FRIENDS

Conference Agenda

Thursday, October 19th

ASU - Memorial Union 2nd Floor

3:00 PM	-	6:00 PM	Registration Packet Pickup & Information Engrained Cafe Entrance
4:30 PM	-	5:30 PM	Meet & Greet Reception Engrained Cafe

Friday, October 20th

ASU - Memorial Union 2nd Floor

7:30 AM	-	5:00 PM	Registration Packet Pickup & Information Arizona 221 Entrance
7:30 AM	-	8:15 AM	Breakfast & Announcements Arizona 221
8:15 AM	-	9:15 AM	Welcome & Opening Keynote Arizona 221
9:15 AM	-	9:30 AM	Break
9:30 AM	-	10:30 AM	Sessions Ventana 241 A, B, C Arizona 221 Gold 207

Conference Agenda

(Continued)

10:30 AM	- 10:45 AM	Break
10:45 AM	- 11:45 AM	Sessions Ventana 241 A, B, C Arizona 221 Gold 207
11:45 AM	- 12:45 PM	Lunch Arizona 221
1:00 PM	- 2:00 PM	Sessions Ventana 241 A, B, C Arizona 221
2:00 PM	- 2:15 PM	Break
2:15 PM	- 3:15 PM	Sessions Ventana 241 A, B, C Arizona 221
3:15 PM	- 3:30 PM	Break
3:30 PM		Sessions Ventana 241 A, B, C Arizona 221
5:00 PM	- 6:30 PM	Mocktails & Conversations Poster Presentations Arizona 221

Conference Agenda

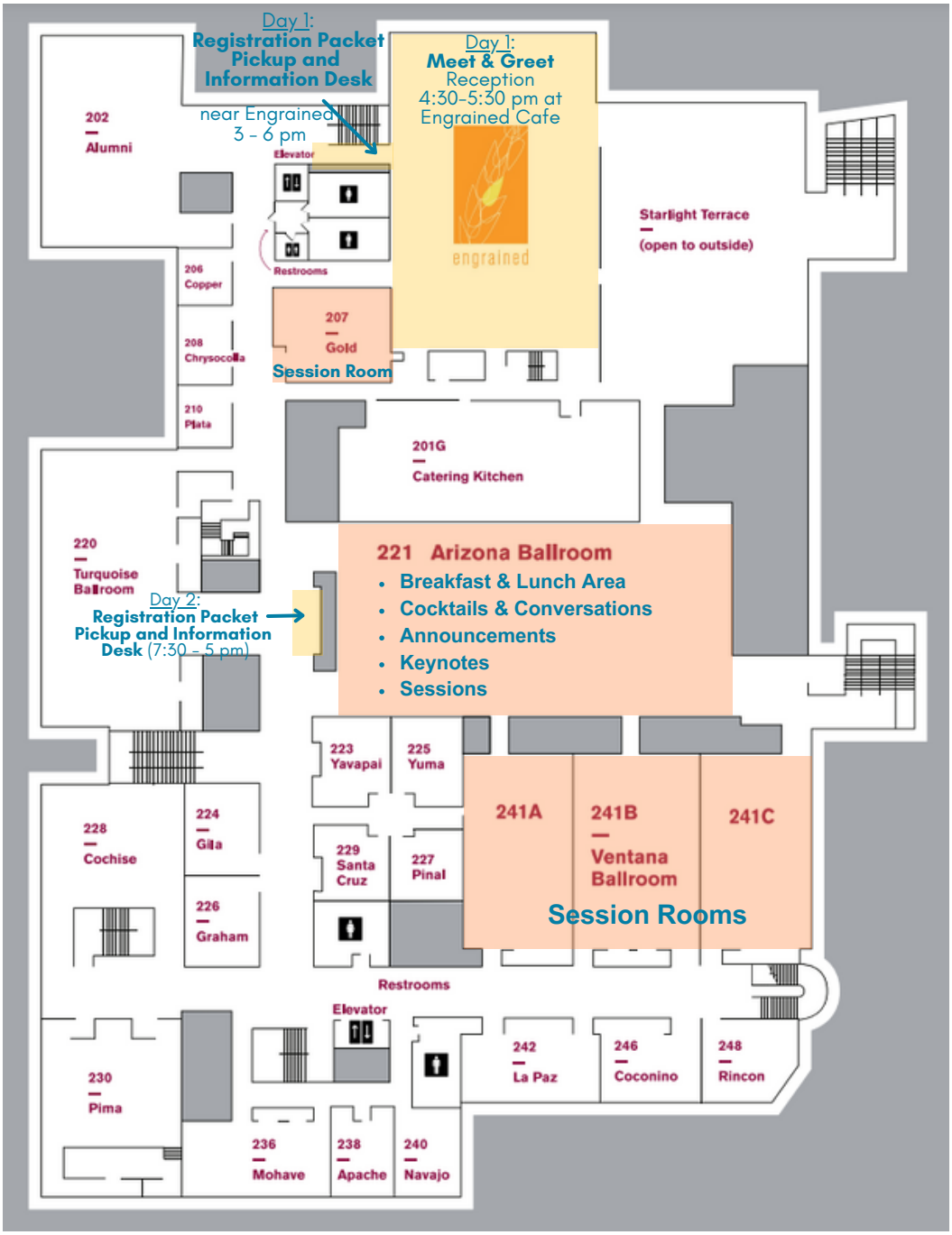
(Continued)

Saturday, October 21st

ASU - Memorial Union 2nd Floor

7:30 AM	-	8:15 AM	Breakfast & Announcements Arizona 221
8:15 AM	-	8:30 AM	Break
8:30 AM	-	9:30 AM	Sessions Ventana 241 A, B, C Arizona 221 Gold 207
9:30 AM	-	9:45 AM	Break
9:45 AM	-	10:45 AM	Sessions Ventana 241 A, B, C Arizona 221 Gold 207
11:00 AM	-	12:00 PM	Closing Keynote & Debates Arizona 221

Arizona State University
Memorial Union (MU) 2nd Floor Map



- The registration packet pickup desk will be located near the entrance of the Engrained Cafe on the 2rd floor of the MU. No on-site registration will be available for in-person attendance. In person attendance must be completed in advance.
- The Memorial Union Address: 301 E. Orange St., Tempe, AZ 85281. (Google Map, Building Map)
- Visitor parking is available at the Apache Boulevard Parking Structure. To view additional parking on campus and hourly rates, please visit Parking and Transit Services.

(The information on this page is clickable. Please download the PDF file to your device, so you can open the links.)

THURSDAY, OCT 19
3:00 PM - 6:00 PM

3:00 - 6:00 *Engrained Cafe Entrance*

REGISTRATION PACKET PICK UP AND INFORMATION

4:30 - 5:30 *Engrained Cafe*

MEET & GREET RECEPTION

Light hors d'oeuvres will be served

FRIDAY, OCT 20
7:30 AM – 6:30 PM

Conference On Test Security 2023

7:30 - 5:00 *Arizona 221 Entrance*

REGISTRATION PACKET PICK UP & INFORMATION

7:30 - 8:15 *Arizona 221*

BREAKFAST & ANNOUNCEMENTS

8:15 - 9:15 *Arizona 221*

Welcome to COTS 2023

OPENING KEYNOTE & DEBATES

Kick off COTS with the always-informative, always-entertaining COTS Debates! Using a fun and fast-paced format that includes opportunities for audience engagement, the annual COTS Debates will explore the impact of emerging technologies on testing and test security. Following the debate, debaters will participate in a short panel discussion to further explore some of the more nuanced considerations regarding these technologies. Join exam security experts as they debate top-of-mind issues facing our industry today.

FRIDAY, OCT 20
7:30 AM - 6:30 PM

9:30 - 10:30

Ventana 241A

A More Holistic Approach to Validity: Considering the Connections Between Validity, Inclusivity, and Test Security.

John Kleeman, Learnosity, Questionmark | Liberty Munson, Microsoft | Benjamin Hunter, Caveon
Panel Presentation

The strongest argument for test security is that failures in security negatively impact validity, with the risk that test-takers pass the test when they shouldn't. However, some well-intentioned test security measures can negatively impact parts of the test taker population (e.g., those with accessibility challenges, those who do not speak the language in which the test is published, etc.). An inclusivity failure in testing also negatively impacts validity, with the risk that test-takers fail the test when they shouldn't for reasons not related to the construct being tested.

Failures in inclusivity that negatively impact parts of the test taker population, negatively impact ALL stakeholders! This session explores how to balance security with inclusivity, covering:

1. What is inclusivity in testing and why it matters.
2. Real-world examples where security measures reduce inclusivity, including biometrics, too strict remote proctoring, and many more.
3. Real-world examples where inclusivity measures can pose a security risk, including text to speech allowing recording of question content.
4. A perspective from a leading IT company on how they balance security and inclusivity in their certification program.
5. Examples of how to get both security and inclusivity, including using observational assessment, performance testing and using large question banks to reduce the impact of content theft.
6. A suggested approach for balancing security and inclusivity, based on the goals of the testing program and balancing the risks of misclassification (incorrect passes or incorrect fails).

The session will also invite audience input into the challenge of balancing security and inclusivity.

Ventana 241C

AI & Test Security: Key Considerations in Leveraging AI Tools in Developing and Delivering Tests

Paul Muir, Surpass | Marc J. Weinstein, MJ Weinstein Law | David Yunger, Vaital | Isabelle Gonthier, PSI Services LLC
Panel Presentation

Artificial intelligence, or AI for short, is a hot topic these days, so hot that it generates polarized reactions across the testing and assessment industry. Some reactions are excitement and dreams of enhanced productivity and efficiency in how we develop and delivery tests. Other reactions are of concerns and warnings as to the risks associated with AI tools, especially as it relates to security, data, intellectual property, and cheating concerns. But are there not shades of grey in considering the opportunities and threats AI brings to our industry? In this panel discussion, we will outline and dissect concerns and opportunities related to the use of AI tools, including generative AI, both from a test development and delivery perspective. We will challenge our thinking, discuss pros and cons, and engage the audience in thought provoking discussions on security considerations related to the use of AI.

9:30 - 10:30 *CONTINUED...* *Arizona 221*

When Cheaters Prosper: Exploring the Business of Cheating

Rachel Schoenig, Cornerstone Strategies | Ray Nicosia, ETS | Camille Thompson, College Board | Faisel Alam, LSAC
Standard Presentation

According to the old adage, "cheaters never prosper." But is that true? Is there a business model for cheating that undermines the old adage? This session will address the practical realities of cheating, breaking it down to the dollars and cents motivations of many of the actors. Using real world case studies and examples, testing professionals will explore how individuals who make a business of cheating use targeted marketing strategies, develop unique pricing structures, and seek out funding to exploit worried test takers and undermine exam security. Together, we will break down the business of cheating and share various methods for addressing it, including use of new and emerging regulations and case law. It's a session you really can't afford to miss!

Gold 207

So Now What? Taking Action Once Exam Misconduct Investigation Has Been Completed.

Harry Samit, Bryan Friess, Jodi Reiser, Pearson VUE | Christa Jurasic, Kelsey Makkay, Association for Supply Chain Management
Standard Presentation

Misconduct by candidates and by people the candidates hire to assist them with cheating threatens the integrity of credentialing programs. Although only a small proportion of candidates engage in such activity, the impact is significant.

Various investigative techniques can be utilized to uncover such activity and to inform exam programs what happened and who is responsible.

But then what?

This presentation examines options available to exam programs to remediate misconduct once it has been detected and investigated. It focuses not just on traditional measures such as score revocations and bans, but also details more impactful tactics such as referral to law enforcement agencies, prosecutors, and regulatory bodies, as well as the use of civil litigation.

While more complex and resource-intensive than traditional measures, external referrals have the potential to be highly effective. This is because prosecution in many cases either deters criminal enterprises from continuing their illicit activities or stops them altogether by incarcerating those responsible and seizing their assets.

This presentation will examine a variety of misconduct remediation measures to give exam sponsors a slate of options once misconduct has been confirmed. The goal is to educate participants, so they can make the best decisions for their exam programs to neutralize exam security threats today and in the future.

FRIDAY, OCT 20
7:30 AM – 6:30 PM

10:30 - 10:45

BREAK

10:45 - 11:45

Ventana 241A

Test Security Masterclass: Small Changes That Will Make a Big Impact on Your Test Security

Cicek Svennson (Moderator), Steve Addicott, Caveon | Phil Dickison, NCSBN | Alana Chouman, Conference of State Bank Supervisors

Panel Presentation

In this masterclass, a panel of leading test security experts will draw on their years of hard-won experience to provide you a series of tangible, actionable test security recommendations that you and your team can easily implement. Our actionable recommendations will help you identify areas of weakness and develop strategies to mitigate risk effectively. Whether you are new to the test security game, or a testing program with years of experience, you will leave this session equipped with new ideas for boosting the security (and validity) of your exams. The moderator will introduce six proven, crucial test security best-practices. Our panelists, program leaders from financial and health care test programs, will share their experiences, both good and bad, in implementing them. They will share the benefits these best practices have forged in improving test security and validity.

Ventana 241B

Combining the Strengths of Humans and AI to Win the Test Security Arms Race

Basim Baig, Chenhao Niu, Kim Snyder, Rose Hastings, Vanessa Villarreal, Will Belzak, Yong-Siang Shih, Duolingo

Coordinated Symposium

The Duolingo English Test (DET) security and proctoring teams present on four topics related to their human-in-the-loop approach to test security, followed by a Q&A section. The presentation topics are as follows.

Combining the Strengths of Humans and AI: Rose Hastings and Basim Baig introduce DET's human-in-the-loop approach to test security – namely, using AI to enhance human decision-making, and using humans to build fairer, more effective AI – with the primary goal of staying one step ahead of sophisticated actors in the test security arms race.

Building and Maintaining a Remote Proctoring Team: Vanessa Villarreal and Kim Snyder describe work on building and maintaining a high-quality, international team of remote proctors, including hiring, training, metrics tracking, and engagement, all of which are critical to jumpstarting and facilitating human-in-the-loop test security.

(Continued on next page...)

10:45 - 11:45 *CONTINUED...* *Ventana 241B*

Combining the Strengths of Humans and AI to Win the Test Security Arms Race

Basim Baig, Chenhao Niu, Kim Snyder, Rose Hastings, Vanessa Villarreal, Will Belzak, Yong-Siang Shih, Duolingo
Coordinated Symposium

(Continued...)

Using AI to Improve Human Decision-Making: Basim Baig and Yong-Siang Shih describe new AI security features built to enhance decision-making, including plagiarism detection, keystroke and mouse-tracking analysis, and cheating ring detection, while also highlighting some of the greatest AI-based security threats facing the industry.

Using Human Decisions to Improve AI: Will Belzak and Chenhao Niu focus on work related to measuring and improving proctor decision-making through automated administration of recycled test sessions, as well as using AI prediction models to make fairer, more consistent decisions about which test takers are cheating.

Ventana 241C

Test Security and Data Privacy: Guidelines for Technology-Based Assessment

John Weiner (chair), Lifelong Learner Holdings | David Foster, Caveon | John Kleeman, Learnosity, Questionmark
Panel Presentation

The proliferation of technology-based design and online delivery of assessments has heightened concerns about test security, as new threats and risks continue to emerge which may undermine validity and fairness, as well as the cost-effectiveness of testing programs. At the same time, these technology advances have increased the digital footprints of test takers and testing programs, making it possible to store and analyze expansive interconnected datasets in new and innovative ways, enabling advanced assessment and security methods. The use of individual data is restricted by data privacy laws and regulations that testing organizations must consider in operating their programs.

This session examines test security and data privacy considerations and best practices in the context of new Guidelines for Technology-Based Assessment (TBA), recently published by the ITC and ATP. The panelists are senior leaders and experts in high stakes assessment, technology, security and privacy issues and methods, who were involved in creating the Guidelines.

The session will:

1. Introduce the Guidelines to those who are not familiar with them, explaining how the Guidelines can help contribute to an effective test security program.
2. Suggest how the Guidelines can be used to align vocabulary and concepts with stakeholders to help communicate how to make test security and data privacy more effective.
3. Discuss technology-enabled test security methods, data privacy laws and regulations, and the application of best practices outlined in the TBA Guidelines.

Standard Presentations

Transitioning Our Exam and Learning Programs to a Virtual Model

Alex Tran, Tabasom Eftekari, Adrian Frisina, Touchstone Institute

Touchstone Institute is an organization that delivers high stakes assessments and orientation and learning programs for internationally educated health professionals (IEHPs). Prior to the pandemic, Touchstone Institute's primary mode of delivery was an in-person administration model. This presentation focuses on Touchstone Institute's transition to virtual program delivery, the challenges and experiences with online systems and how we ensured that our exams and orientation programs abided by security conditions remotely.

Due to the pandemic, there were restrictions to our OSCE exams and learning programs that required a setting in which in-person client encounters were important to assessing an IEHP's skills and knowledge. While other similar organizations decided to postpone and delay their programs until in-person gatherings were allowed, Touchstone Institute forged forward with developing new innovative modalities to replicate in-person assessments. Transitioning to an alternative delivery model highlighted the many different facets of security requirements. Within our exams and learning programs, security conditions differ and as an organization we needed to address the appropriate processes and policies were implemented to fit the model.

Several iterations of exams and learning programs were offered virtually in which psychometric analysis determined that examinee and learner results were not influenced by the mode of program delivery.

Security Innovations – A Necessity for Any Test Sponsor

Dan Sheets, Meazure Learning

Meazure Learning has taken significant steps to improve the security posture and methods for clients. In the age of international test delivery, exam threats are real and significant. Join Dan to understand what tools, programs, tactics, and even physical centers have been developed to improve security. These security innovations discussed will be thought-provoking and give test sponsors ideas they may be able to leverage for their program. Unfortunately, when there is something of value there are bad actors that will try to take it. Bad actors are innovating, test sponsors must also innovate.

Innovations come in all shapes and sizes. The most important part is that you aren't sitting idle.

Mitigating the Effects of Preknowledge in Multistage Testing

Merve Sarac (presenter & author), James Wollack (co-author), University of Wisconsin-Madison

Mitigation strategies for correcting scores obtained by fraudulent behavior are critical for increasing test security following detecting such behavior in real time. We propose a new preknowledge-adjusted weighting method that downweights observations on potentially compromised items. The weighted ability estimation is an intervention to mitigate the bias in interim ability estimates in multistage testing (MST). This approach is expected to route examinees to more accurate modules of appropriate difficulty adaptive to their purified ability estimates. Most importantly, more accurate routing of examinees based on weighted ability estimates improves the accuracy of the unweighted ability estimates after completing the adaptive test.

FRIDAY, OCT 20
7:30 AM – 6:30 PM

10:45 - 11:45 *CONTINUED...* *Gold 207*

Exam Security 101: An Overview of Exam Security for Newbies

Nicole Miller, NBME | Rachel Watkins-Schoenig, Cornerstone Strategies | Camille Thompson, College Board
Standard Presentation

If you are new (or new-ish) to the testing industry or test security, this session is for you! This session is designed to help individuals new to test security professionals or individuals new to testing better understand the basics of test security. We will review the testing lifecycle and common risks and security practices at each stage of the life cycle. Test security professionals at different stages of their career will discuss how to get up to speed on exam security and review some of the key principles regarding exam integrity and risk management. We'll discuss emerging trends and technologies and how they impact exam security and administration of standardized exams. Finally, we will share helpful resources and lessons learned as well as provide insights around communicating with the rest of the organization, and answer questions you may have.

11:45 - 12:45 *Arizona 221*

LUNCH

1:00 - 2:00 *Ventana 241A*

How to Effectively Conduct On-Site Monitoring of Statewide Assessments

Walt Drane, Caveon | Jennifer Baird, LA DOE | Karen Tohinaka (virtual), HI DOE | Joe Kamell, Caveon
Coordinated Symposium

State Education Agencies (SEAs) are required by the US Department of Education (USED) to conduct on-site monitoring of not only assessments for general education students, but also for students with significant cognitive disabilities. The bottom-line is SEAs must get into classrooms to watch the test administration process to ensure it is being administered appropriately to students according to policy in a fair and consistent manner. As the old saying goes, "Inspect what you expect". Sounds easy, but for many SEAs implementing a quality on-site monitoring program is difficult. This session will allow for the discussion of best practices to effectively conduct on-site monitoring of statewide assessments.

1:00 - 2:00 *CONTINUED...* *Ventana 241B*

Prevent, Deter, Detect, React – Best Practices for Pre, During and Post-exam Security Processes

Paul Muir, Surpass Assessment | Ben Hunter, Caveon
Panel Presentation

Exam security is a vital aspect of any assessment programme in the battle to create and sustain academic integrity and fairness. Best-in-class exam security involves protecting exam materials pre-exam, preventing and detecting cheating during an exam, and ensuring that exam results are valid and reliable.

1. Pre-exam, we will look at the latest approaches for ID/person verification, web-patrol and test forms in the design stage – from randomisation of question-and-answer order, to Linear On The Fly testing.
2. During the exam, we will present the people, process and technology steps you can take to secure the exam environment.
3. Post-exam, we will evaluate the value of data forensics in protecting against exam fraud, and the latest tools available.

This session will provide practical tips for maintaining exam security and ensuring the integrity of the exam process. Attendees will leave with a deeper understanding of best practices for exam security and how to implement these in their own work.

Ventana 241C

Finding the Right Balance: Accommodations and Test Security

Ashley Norris, Meazure Learning | Faisel Alam, Brittany Thornton, LSAC
Panel Presentation

Many testing programs have moved to hybrid delivery models to support candidates and test-takers with accessibility and reach. This transition has provided unique challenges and opportunities when it comes to supporting accommodated test-takers. For instance, how are testing programs balancing critical issues like test security, test-taker experience, and variability during the administration? How do you define accommodations? In addition to ADA accommodations, how do testing programs and vendors support religious accommodations and other accommodated practices? What are the benefits and challenges in supporting accommodations in each modality? What is the best way to support all test-takers?

This presentation will explore unique and frequent use-cases that address these questions and more when it comes to the right "balance" when supporting accommodations and test security.

1:00 - 2:00 CONTINUED... Arizona 221

Standard Presentations

Operational Security Considerations for Professional and Language Testing Programs

Kirk Becker, Bill Bonk, Pearson

This session examines two very different test security environments. For professional licensure and certification testing, the characteristics of numerous testing programs are used to better understand risk. Descriptive statistics and stepwise regression quantify the relationship between unusual behavior and industry, pre-requirements, value, and other characteristics. In language testing, where automated scoring is used for constructed response items in both writing and speaking, test taker strategies used to inflate scores will be identified.

'May I Go to the Bathroom, Please?' Analyzing the Impact of Unscheduled Breaks on Changes in Item Responses

Cecilia Alves (presenter), Qi Guo (co-author), Andrea Gotzmann (co-author), Medical Council of Canada

Medical high-stakes exams require stringent exam security to ensure that only individuals possessing the requisite knowledge and skills are granted licenses or certifications. Unscheduled breaks may provide candidates the opportunity to access unauthorized study materials. Exam vendors provide information in the examination interaction log that documents the interactions and activities of a candidate during an examination such as buttons clicked and respective time stamps, but they do not consistently document unscheduled breaks. This absence of labels for unscheduled breaks presents a significant challenge when attempting to scrutinize candidates' behaviors post these unscheduled breaks. Therefore, the primary objective of this study is to detect potential unscheduled breaks within examination interaction logs. More specifically, the research questions are:

1) What percentage of candidates took unscheduled breaks?
2) What percentage of candidates revisited and changed previously answered questions after unscheduled breaks?
3) What percentage of changes led to an increase/decrease in test scores?

We used Python to implement the proxy definition of unscheduled breaks, count the number of revisits and changes, and compute the final statistics. This information can illuminate the extent to which unscheduled breaks can compromise examination security. Furthermore, these insights could inform and enhance the design of exam delivery, and develop processes and procedures around unscheduled breaks, thereby mitigating associated risks.

(Standard Presentations continued on next page...)

1:00 - 2:00

CONTINUED... *Arizona 221*

Standard Presentations *(Continued...)*

The Proctor Calibration Tool: Improving Proctor Decision-Making Through Automated Measurement and Feedback

William Belzak, Duolingo

Remote proctors monitor test takers through video/audio recording software and the internet. Proctors make judgments about test-taker behaviors such as "the test taker appears to be reading something off-screen", and decide whether test takers have violated testing rules or cheated. Although proctors have training and experience to ensure their judgments are made as objectively as possible, other factors like personality, selective attention, and cognitive bias can lead to more subjective decision-making processes. For example, a "strict" proctor may believe a test taker is reading something off-screen, whereas a "lenient" proctor may believe the same test taker is thinking while looking off-screen. In this talk, we describe the proctor calibration tool (PCT), which is designed to make proctors more reliable in their decision-making. This tool is surreptitiously integrated into proctors' work-flow through the administration of recycled test sessions, and immediate feedback is automatically generated via peer decisions. We present results from the PCT that show that, as expected, proctors show high variability in decision-making but appear amenable to intervention.

2:00 - 2:15

BREAK

FRIDAY, OCT 20
7:30 AM - 6:30 PM

2:15 - 3:15

Ventana 241A

Documenting Your Way to Better Test Security

David Ragsdale, Massachusetts Department of Elementary and Secondary Education | Camille Thompson, The College Board
Panel Presentation

Documentation may not seem exciting. But the simple act of writing things down -- procedures, precedents, goals, risks, policies, results -- is a powerful tool for keeping a testing program on track and enhancing its security. Some of the benefits of good documentation include:

- articulating the goals of your testing program
- defining the risks your test security program is intended to address
- ensuring that your policies are clearly articulated and well-thought-out
- keeping your responses to situations consistent with past practice
- making it less likely you will deviate from your procedures, and ensuring that changes are deliberate and intentional
- reducing ad-hoc decision-making
- keeping track of how successful your test security efforts are
- putting your testers and other stakeholders on notice for what rules and standards they will be held to
- establishing expectations for staff and any contracted vendors

In this session, several test security experts will explain why documentation is so important, and will share numerous examples of the types of documents, document review, and other practices that keep your test security program running smoothly and effectively.

Ventana 241B

Using ChatGPT for the Automatic Creation of Smartitems™ for the Purposes of Test Security and Culturally Responsive Assessment.

Sergio Araneda (presenter), David Foster (author), Caveon
Demonstration

In this demonstration, we will show how to leverage the existing capacities of Scorpion, a technology developed by Caveon Test Security, with the new capabilities of Large Language Models (LLM) to enhance item development. We will show how to write SmartItems™ using ChatGPT's new capacities to improve test security while tuning the questions to the cultural backgrounds of the examinees at the same time. New LLM offers a new way to create SmartItems™ via the automatic writing of javascript code compatible with Scorpion, facilitating the use of randomly parallel tests as a strategy for test security, accelerating the process of item development, and making the final product culturally responsive.

2:15 - 3:15 *CONTINUED...* *Ventana 241C*

Study Sites or Cheating Sites? A Discussion on How to Treat Web Sites Selling Study Aid to Students.

Jake Ritz, Christine Erickson, Ascend Learning | Dan Sheets, Meazure Learning
Facilitated Roundtable

In today's connected world there are more ways to collaborate and share information than ever before. People are more likely to turn to the internet to find the answer to a question or for help solving a problem than to someone they know. In the education industry there are now hundreds of sites and forums where learners can buy and sell class notes, practice tests/quizzes, papers, or even hire others to do their homework and take their tests for them. With little regulation or processes in place to ensure the content is what they say it is, the buyer beware warning is very relevant. And the pervasiveness of these easy to obtain short-cuts can jeopardize the ability for companies and institutions to deliver well-prepared and competent workers to the workforce. There is no easy solution and one-size does not fit all, but what options are there and how do you decide what is best? Please join this roundtable to hear from others across the industry about the challenges these sites present and ways to influence the impact to your business.

Arizona 221

Standard Presentations

Strengthening Evidence of Preknowledge by Analyzing Differences in Response Times Over Item Subsets

Marcus Scott, Caveon

Score differencing techniques can detect examinees who likely had preknowledge of some, but not all items. However, score difference results are typically not strong enough to be actionable evidence. It has been observed that preknowledge of test content can lead to reduced item response times. Therefore, showing that an examinee had significantly better performance on a known set of disclosed items and significantly faster response times on those items can strengthen the evidence that the examinee had access to the disclosed content.

This presentation discusses the development of a time differencing statistic. For each item, a model for estimating an examinee's response time is created. Using the item response time models, standardized residuals between the estimated response times and the observed response times are computed. The Mann-Whitney U Statistic is used to test the difference between the distributions of standardized residuals for the disclosed items and the nondisclosed items.

To demonstrate the statistic, it was applied to data from a testing program that used new and old items in its administrations. One particular training program had a history of having better performance on the old items that were being reused than on new items. This led the testing program to suspect that examinees in this training program were gaining preknowledge of the old items by harvesting items from each administration.

(Standard Presentations continued on next page...)

2:15 - 3:15 CONTINUED... Arizona 221

Standard Presentations (Continued...)

Proceed With Caution When Using Caution Indices for Data Forensics

Steven Svoboda (presenter & author), PSI Services LLC | Greg Hurtz (author), California State University, Sacramento and PSI Services LLC

Several person-fit measures and caution indices have been reviewed and evaluated in the literature on detecting fraudulent testing behavior, and some have been deemed effective at detecting data patterns mimicking inappropriate testing behavior. We will discuss several such indices in terms of what they do and do not measure relative to other types of indices (e.g., response similarity), and factors we have found to increase false alarm rates. In so doing, we will explain why we suggest to "proceed with caution" when using caution indices and other similar person-fit measures as part of a data forensics analysis. We will use examples from our own analyses on real data, as well as simulations we have run. Our goal is to share our insights and help increase understanding and awareness of the behavior of the statistics we review, and also invite the audience to share any of their own experiences and insights with these types of measures.

A Score-Based Method to Detect Both Item Compromise and Preknowledge in Computerized Adaptive Testing

Kylie Gorney (presenter for virtual conference & author), Michigan State University | Danielle Lee, (presenter for in-person conference & co-author) American Institute of Internal Medicine | Jianshen Chen, College Board

In recent years, several methods have been proposed to simultaneously detect compromised items (CI) and examinees with preknowledge (EWP) (e.g., Chen et al., 2022; Gorney et al., 2023; Pan et al., 2022). However, the majority of these methods are limited in one of two ways: (a) the method was designed for non-adaptive tests, and may therefore not be robust to the assumptions of adaptive tests, or (b) the method draws on information from item response times, which are not always available. In this paper, we address both limitations by proposing an iterative method that is designed for computerized adaptive tests and only requires information from the item scores. We examine the performance of our method using detailed simulations where the following factors are manipulated: test length, the proportion of CI, the proportion of EWP, and EWP ability distribution. Results show that across most conditions, the new method is able to produce small false positive rates and large true positive rates for both items and examinees.

3:15 - 3:30

BREAK

FRIDAY, OCT 20
7:30 AM – 6:30 PM

3:30 - 5:00
90 MINUTES

Ventana 241A

Exam Security: Protecting Everything Everywhere All the Time

Rachel Schoenig, Cornerstone Strategies | Camille Thompson, College Board | Ray Nicosia, ETS | Faisel Alam, LSAC

Exam security is not for the faint of heart. Protecting a program's testing assets, score validity and brand reputation requires consideration of threats posed by a multitude of potential threat actors – from testing candidates and fraud rings to rogue insiders and illicit test prep providers. At the same time, it also requires consideration of how best to spend limited resources and time. How can testing programs best use their resources to protect everything, everywhere, all the time? This interactive session will highlight common and emerging threats to testing. It will also address methods for combating risks, including some cost-saving alternatives. Join experienced testing professionals for a fun, engaging and potentially award-winning session that will better position your program to protect its testing assets, score validity and brand reputation from everything, everywhere, all the time.

3:30 - 4:30

Ventana 241B

Taming the Beast! Turning Generative AI From Threat to Your Best Test Security Ally

David Foster, Caveon
Standard Presentation

There is concern that ChatGPT, or other tools of generative AI or large language models, is the biggest test security threat we have ever faced. The AI models have been able to pass tests for medical licensing, becoming a lawyer, college admissions, AP courses, and even sommelier exams. The logic is solid that if AI can pass these tests, that helping people cheat would be child's play. What should we do in the testing industry to prevent or blunt this new type of fraud? Or is the demise of the testing industry truly inevitable? The answers to these questions are covered in this session, and (spoiler alert!) will hopefully ease your concerns a bit.

But don't forget the other side of the coin, the other obvious question: How can the same AI tools make testing safer, more secure? What do they have to offer that makes cheating more difficult, that makes the theft of test content less of a problem? This session will present several benefits of AI to us who are tasked with protecting the value of test scores in aiding important life decisions? Wouldn't it be amazing if a tool with such vast capability could become an ally rather than a threat?

This session basically ignores "the sky is falling" over-reaction to AI "possibilities", and presents unique and realistic perspectives to help each of us navigate this new world of test security. Evidence from data simulations and experimental research will be presented to support these perspectives where applicable.

3:30 - 4:30 *CONTINUED...* *Ventana 241C*

Designing Comprehensive Test Security Procedures for Innovative State Balanced Assessment Systems and Through-Year Assessments: Recent Implementations by States

John Olson (chair), OEMAS, Caveon | Jennifer Baird, LA DOE | David Ragsdale, MA DOE | Walt Drane (discussant), Caveon
Coordinated Symposium

An important, and new, development in the state assessment world is the implementation of Balanced Assessment Systems (BAS) or Through-Year Assessments (TYA). A state BAS requires a system of high-quality assessments that serves multiple purposes for a wide range of audiences. Several states are moving toward implementation of an innovative BAS that may include interim, benchmark, formative, or other assessments along with the summative. In addition, many SEAs are busy implementing Through-Year Assessment (TYA) designs, which may be part of their BAS or be an approach to reduce the emphasis on end-of-year summative assessments. However, all of these assessments must be implemented with fidelity to yield valid and reliable outcomes. Thus, the need for comprehensive test security policies/procedures for all parts of the assessment system. In this session, three state presenters will discuss the implementation of a BAS or TYA in their state and how a comprehensive approach to security is being integrated into the procedures used for test administrations. Two national test security experts will share information on what states are doing to improve security for all the different tests that make up their BAS or TYA, and make recommendations for improvements to their security procedures.

Arizona 221

High Stakes, Higher Security: Safeguarding Your Certification Program From Common Pitfalls

Cicek Svensson, Susan Weaver, Caveon | Jake Ritz, Ascend Learning | Isabelle Gonthier, PSI Services
Standard Presentation

High-stakes testing is a critical component of educational and professional advancement, with significant consequences for test-takers. As such, ensuring the validity, reliability, and fairness of high-stakes tests is paramount. This discussion will focus on the importance of test security in creating a new high-stakes certification program and highlight common pitfalls to avoid. This session will cover topics such as test design, test administration and post-test analysis, and will address common oversights that lead to mistrust in the certification process, compromised security, and invalid exam results, including inadequate test design, administration, and lacking post-test analysis. The discussion will also address key considerations for test administration, such as proctoring and test environment, to prevent cheating and irregularities. Throughout the discussion, common hurdles that can lead to mistrust in the certification process, compromised security, and invalid exam results will be highlighted as well as lack of transparency and communication with stakeholders. Attendees will gain an appreciation for the importance of test security throughout the certification process and learn practical strategies to avoid common mistakes associated with creating new programs. By focusing on test security, certification program administrators and educators can ensure the validity and fairness of high-stakes assessments, and maintain public trust in the certification process.

FRIDAY, OCT 20
7:30 AM - 6:30 PM

5:00 - 6:30

Arizona 221

MOCKTAILS & CONVERSATIONS
WITH POSTER SESSION EVENT

Light hors d'oeuvres will be served

You won't want to miss it!

Poster Presentations

A Machine Learning Case Study of Relationships Among Answer-Copying and Similarity Statistics for Multiple-Choice Tests

Mengyao Zhang (presenter & author), National Conference of Bar Examiners

Unusual response similarity among examinees is a type of irregularity that may be observed in test data and indicate potential cheating. There are currently a large number of answer-copying and similarity statistics proposed for multiple-choice tests, with different model assumptions and features. This study will examine relationships among these statistics, based on operational real-world cases over ten years of administration of a high-stakes licensing examination. Machine learning approaches will be applied to exploring the potential interrelationship among the different statistics and how consistently these statistics, or combination of some statistics, flag or predict a case that is worth further investigation.

Creating a Benchmark to Detect Item Response Trajectory Similarity

Hongling Wang (presenter), Chi-Yu Huang (author), Xin Li (author), Bob Schwartz (author), ACT Inc.

Using item response timestamps and orders of item responses, Wang & Huang (2022) defined the concept of item response trajectory and an index of item response trajectory similarity. The index is used to flag suspected examinee pairs from item response time point of view. However, the criteria used to flag suspected examinee pairs in that study are subjective. This study will improve the way how the index is defined. As a result, the new index makes it possible to create an objective benchmark which is then used to evaluate the magnitude of similarity of item response trajectories of a pair of examinees and flag examinee pairs with significant item response trajectory similarity.

(Poster Presentations continued on next page...)

MOCKTAILS & CONVERSATIONS
WITH POSTER SESSION EVENT

Light hors d'oeuvres will be served

You won't want to miss it!

Poster Presentations *(Continued...)*

Detect Aberrant Testing Behavior Applying a Hierarchical Model

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

Investigating fraudulent testing behavior, especially for high-stakes assessments, has been a common practice for maintaining test score validity. Item response and item response time, either separately or combined, have been utilized to evaluate examinees' performance, such as the hierarchical model (van der Linden, 2007) which incorporates both item response and item response time. Tests with very short response time and poor scores could be a sign of item harvesting or pre-knowledge. The current study implemented van de Linden's hierarchical model to identify extremely rapid responders using both real and simulation data. The effectiveness of this hierarchical model in flagging faster responders is discussed based on real and simulated data.

Trauma-Informed Testing Guidelines for Enhanced Test Security

John Jones (author), Kelly Dages (author), FifthTheory, LLC

This poster will share guidelines and standards to help test sponsors become trauma-informed. The term "trauma" is increasingly used in the post-pandemic economy and encompasses a wide range of experiences that can have profound and lasting effects on an individual's mental, emotional, and physical well-being. When the test-taking process and environment is sensitive to traumatized test-takers, test-takers are less likely to trigger emotional reactivity and less likely to be tempted to cheat. Test anxiety is a common concern for many students and test-takers, but it can be particularly debilitating for those who have experienced trauma. The heightened emotional reactivity and difficulty with emotional regulation associated with trauma can exacerbate feelings of anxiety and stress during high-stakes tests, leading to a negative impact on test performance or to a decision to commit test fraud. Test security is essential for maintaining the credibility of high-stakes tests; however, some security measures may exacerbate test anxiety and stress or create additional barriers for those that need accommodation. To create a more trauma-informed testing environment, test sponsors can adopt a range of practical strategies that align with the principles of trauma-informed care. Therefore, the authors have worked with GPT 4.0 to generate 45 trauma-informed testing guidelines that can be grouped into five categories: Understanding Trauma and Its Impact, Inclusive Test Design and Administration, Support for Test-Takers, Continuous Improvement and Self-Assessment, Advocacy and Policy Change.

SATURDAY, OCT 21
7:30 AM – 12:00 PM

7:30 - 8:15

Arizona 221

BREAKFAST & ANNOUNCEMENTS

8:30 - 9:30

Ventana 241A

A Top Secret Strategy to Eliminate Cheating on Tests

Ronald A. Berk, Johns Hopkins University | David F. Foster, Caveon
Standard Presentation

You might be thinking: "It can't be done." Actually, it can and the technology is finally in place. Cheating is snuffed out when you create tests by sampling items from a gigantic pool of items. A test is just a sample of items that measures a defined domain of content. In yesteryear, we assumed items to be a sample; today, we can actually draw random samples that are statistically equivalent. They are like multiple test forms. Instead of just one or two, we can generate an unending truckload of them, with each test taker getting a different unique form. This sampling process will be demonstrated in the session. How do you cheat on those tests?

Technically, we are moving from the long-held assumption of randomly parallel tests to the reality of those tests. More than 70 years ago, Frederic Lord proposed that assumption. Now it is possible to convert that assumption into reality and solve major test security problems as well. With test takers having less ability to cheat, many typical test security strategies, such as proctoring and finding tests on TikTok, will need re-thinking.

This session will address 15 reasons you believe randomly equivalent or parallel tests will not work and a list of myths about the practical constraints to implement such a testing program. The evidence from test theory, statistics, simulations, and empirical research will be presented for this old-but-new approach to the design and administration of tests. You do not want to skip this session.

Ventana 241B

~~Exam Cheating – Tools and Techniques to Use Online Information, Item Data, and Technical Delivery~~

~~*Brent Hill, Brent Morris, Cisco | Bryan Friess, Harry Samit, Pearson VUE*~~
~~*Panel Presentation*~~

This session has been canceled.

8:30 - 9:30 *CONTINUED...* *Ventana 241C*

Will You Ever Stay Ahead of the Nefarious? No, but You Can Try!

Susan Weaver, Caveon | Saundra Foderick, HPE

Panel Presentation

As long as there are exams, there will be those that chose to gain their credentials without really studying or knowing the subject matter. Many do this by gaining pre-knowledge, so how can we make that tougher for them? Using Web Patrol, Data Forensics, Audits and Investigation you can determine how and maybe even who is not on the straight and narrow path. But then what?

Larger item pools are the key. Easier said than done. A member of the HPE exam development team will discuss how they are attempting to thwart the nefarious and get ahead of them. Using Cloning, AIG, and SmartItems – HPE is doing just that.

Arizona 221

Standard Presentations

Item Exposure vs Generative Content

Christopher Foster, Caveon

In the field of education, test questions play a critical role in assessing students' knowledge and skills. However, the reuse of test items can compromise the integrity of the assessment process and lead to cheating. To address this issue, one solution is to generate large numbers of test questions using generative AI models, such as Chat GPT, so that exposed items can be replaced.

In this study, we present a framework for using generative AI to generate test questions, which includes the following steps: (1) identifying the learning objectives and topics to be assessed, (2) generating a large pool of test questions using the trained model, and (3) reviewing and selecting the questions that meet the desired quality criteria.

We discuss the potential benefits and challenges of using generative AI for test question generation, including the ability to rapidly generate large numbers of items and the need for careful quality control. We also present preliminary results from a pilot study that demonstrate the feasibility and effectiveness of our framework as well as the quality of generative AI questions.

Overall, our study highlights the potential of generative AI for addressing the issue of item exposure and providing a sustainable solution for test question generation.

Keywords: generative AI, item generation, item replacement, education, Chat GPT

(Standard Presentations continued on next page...)

8:30 - 9:30 *CONTINUED...* *Arizona 221*

Standard Presentations *(continued)*

A Mark-Recapture Approach to Approximating Item Pool Compromise

Rich Feinberg (presenter), Ian Micir (author), National Board of Medical Examiners

Testing organizations routinely investigate if secure test material has been compromised and is consequently invalid for scoring and inclusion on future assessments. Beyond identifying individual compromised items, it would be worthwhile to know the extent to which a form is compromised to inform discussions for when it can no longer be deployed or when an item pool is sufficiently compromised such that serious action needs to be taken. Previous research on estimating the aggregate level of item compromise is sparse, however, this is a more generally long-studied problem in ecological research. In this study, we apply a mark-recapture technique, specifically an unbiased estimator of the Peterson method (Seber, 1982), to estimate the population of compromised items for a high-stakes examination. A resampling simulation will also be employed to investigate the effect on the estimated population of compromised items as the number of captured and recaptured items vary. The discussion will include considerations for practice and how a mark-recapture approach could be incorporated into a testing organization's routine monitoring procedures.

How an Assessment Organization Conducted a Mock Exam Security Audit

Tabasom Eftekari, Alex Tran, Touchstone Institute

Concerns about cheating and exam score integrity are heightened in the context of licensure or registration exams where the exam serves as a gatekeeper to practicing in a profession, for which incompetent practice presents a risk to the public health and safety. Preventing and detecting exam security incidents and reducing vulnerability to test security threats are therefore important for any organization involved in the business of assessment. Test security audits are generally costly and are usually conducted by external parties with expertise in this type of reviews. When the need to ensure that security protocols comply with best practices was identified as a strategic goal, and given the scarce resources, Touchstone Institute engaged a group comprised of internal staff to conduct a mock exam security audit to review the current status of its examinations security, identify potential security concerns, and make recommendations for improvement in policies and procedures to address the identified concerns.

Gold 207

Test Fraud: How Secure Is Your Strategy?

John Kleeman, Learnosity | Steve Addicott, Caveon | Dave Clements, Accenture
Standard Presentation

"Learn from the mistakes of others. You can never live long enough to make them all yourself." ~Groucho Marx

In a world where data breaches and identify theft are daily occurrences, come discover what you can do to keep your assessments from becoming headline news. This presentation is a primer on the elements of test fraud, covering the threats and risks of cheating. The panel is comprised of three testing experts who will discuss real-world examples and cover the ins and outs of prevention, deterrence, detection, and what to do when a breach is confirmed.

This session is for those who are new to test security or those who want to explore test security a little further. Attend this to cover the basics of test security, learn from the mistakes and successes of the community, and discover ideas on how to improve the validity of your test program.

SATURDAY, OCT 21
7:30 AM – 12:00 PM

Conference On Test Security 2023

9:30 - 9:45

BREAK

SATURDAY, OCT 21
7:30 AM - 12:00 PM

9:45 - 10:45

Ventana 241A

Vertically-Integrated Assessment: Lessons Learned by Keeping Test Security In-House

Andre Kenji Horie, Rose Hastings, Will Belzak, Kim Snyder, Mirza Basim Baig, Duolingo

Panel Presentation

While many providers employ third-party vendors, the Duolingo English Test (DET) adopts a in-house approach to ensure security—vertical integration. This methodology unifies key functions such as proctoring, item creation, and test security under one roof, providing a consistent and secure test administration. This panel, featuring members of the DET security and proctoring teams, will dive into the lessons learned from this integrated approach. Discussions will center around:

(1) designing tailored security measures for the assessment;
(2) adjusting test rules and test-taker consequences based on empirical evidence of cheating activity;
(3) fostering robust collaborations between engineers, operations managers, and researchers to swiftly tackle emerging security threats;
(4) utilizing test-taker response data in real-time to identify and curb cheating activity;
(5) achieving a delicate balance between stringent security measures and the test taker experience; and
(6) aligning the interests of the assessment provider with the security maintainer.

The panel will further elucidate the benefits and challenges of vertical integration. While this approach minimizes the risk of inconsistencies, potential security breaches, and information leaks, the shift to a fully integrated system may be resource-intensive and necessitate meticulous planning. Thus, potential hurdles and strategies to effectively navigate them will also be a point of discussion. This comprehensive session will encourage questions and comments from the audience. Attendees will gain actionable insights for their own test administration contexts, exploring how vertical integration can revolutionize the credibility and integrity of high-stakes English testing.

Ventana 241B

Test Security Requirements in USED Peer Review: Important Lessons Learned by State Assessment Programs in Recent Years

John Olson, (chair) OEMAS, Caveon | Timothy Butcher, WV DOE | Theresa Bennett, DE DOE | Jeff Holtz, MN DOE | Walt Drane (Discussant), Caveon

Panel Presentation

In recent years the USED included new requirements on test security for its peer reviews, with all states now required to show evidence that their assessment systems have integrity and are secure. Among the many issues described in both ESSA and peer review are critical ones like the validity and fairness of test scores and having high-quality processes in place to ensure test security. Peer Review Critical Elements 2.5 (test security) and 2.4 and 5.4 (monitoring assessment administrations) have been identified as problematic areas for some states. In this session, presenters will describe the types of evidence that best address the test security requirements and can be useful for peer review, such as use of State Test Security Handbooks that serve as comprehensive collections of all state policies, practices, and procedures. Three state presenters will describe the types of evidence they successfully submitted for peer review that address the test security requirements and the lessons they have learned from the feedback they received from USED on them. Their focus will be on best practices in the security of state assessment programs and methods for documenting it. Two test security experts will discuss the types of evidence found to be particularly supportive of high-quality and secure state assessment systems, and where evidence is sometimes found to be lacking.

9:45 - 10:45 *CONTINUED...* *Ventana 241C*

The Future of Test Security – Where Will We Be In 5 Years?

Paul Muir, Surpass Assessment | Rachel Schoenig, Cornerstone Strategies
Facilitated Roundtable

The Future of Test Security is a crucial topic for anyone involved in education, assessment, or certification. With the continued rise of online testing and remote proctoring and of course the emergence of Generative AI, such as ChatGPT, the challenge of ensuring test security has become even more complex. As we look ahead to the next five years, it is clear that test security will continue to be a pressing issue.

In this interactive session, we will explore the latest trends and innovations in test security, and discuss where we are likely to be in five years. We will examine the impact of emerging technologies such as artificial intelligence, biometrics, and blockchain, and discuss how they can be leveraged to enhance test security.

We will also discuss the challenges and ethical considerations associated with implementing these technologies, such as privacy concerns and potential biases. Additionally, we will examine the role of humans within an AI driven ecosystem and why humans will remain at the heart of any test security solution.

Finally, we will explore the implications of these developments for test takers, educators, and employers, and discuss strategies for ensuring that the benefits of enhanced test security are balanced with fairness and accessibility for all.

This session will provide valuable insights and practical advice for anyone involved in test security and assessment and will allow the entire group to share experiences to help us shape the next 5 years in test security.

Arizona 221

Great Figures in Testing – "History Doesn't Repeat Itself, but It Often Rhymes"

Jim Hussey, ACT

Across the test security profession we collectively bring a wide range of academic training and career expertise to our roles, often in educational measurement, law, teaching, and technology. One discipline that rarely makes the list but should: History.

Consider the challenges our forebears faced - testing at scale, the impact of testing on teaching, the emergence of the computer, the inclusion of the historically excluded, and choosing between psychometric evolution and revolution. These issues are top-of-mind in 2023, but most could have headlined the not-quite-there-yet COTS conferences of 1923, 1933, 1943, and 1953 (and every year since).

This presentation will not be a dry recitation of facts and figures, but instead a modest effort to introduce, understand, and appreciate a few of the storied individuals who preceded us, in hopes it might inspire all of us to make some history of our own.

SATURDAY, OCT 21
7:30 AM - 12:00 PM

9:45 - 10:45 *CONTINUED...* *Gold 207*

The Impact of Cultural and Linguistic Differences on Test Security and Fairness

Cicek Svensson, Caveon | Sue Orchard, Comms Multilingual | AlyssaRulf Fountain, Prometric
Standard Presentation

The increasing diversity of test-takers presents unique challenges to ensuring test security and fairness. Cultural and linguistic differences can impact test performance and may also affect test-taker behavior, including the likelihood of cheating. This session will explore the impact of cultural and linguistic differences on test security and fairness and highlight strategies for addressing these challenges. The session will begin by discussing the various ways in which cultural and linguistic differences can impact test performance, including differences in test-taking norms, language proficiency, and familiarity with test formats. The session will also explore the potential impact of stereotypes and bias on test-taker behavior, and how these factors may influence cheating and irregularities. The session will then shift to practical strategies for addressing these challenges. This may include the use of culturally responsive test design, such as incorporating culturally relevant content and using translated test materials. Furthermore, the presenters will explore the importance of providing accommodations for test-takers with language barriers, such as extended testing time, interface settings, content-related accommodations etc. Attendees will gain a deeper understanding of the challenges posed by cultural and linguistic differences in high-stakes testing and learn practical strategies for ensuring test security and fairness for all test-takers.

11:00 - 12:00 *Arizona 221*

CLOSING KEYNOTE

An Optimistic Look at the Future

A psychometrician, attorney, and exam security expert walk into a bar…. And so begins the Closing Keynote of the 2023 Conference on Test Security. Join us for an entertaining and informative look at the legal, practical, and psychometric changes on the horizon in assessment and exam security. We'll weave together the various sessions from the conference, discuss different perspectives on where technology and the law is headed, explore how these will impact our work, and highlight the bright side of being in the industry today. It's a closing session that will leave you energized and excited for the future!

NOTES

NOTES

NOTES

SAVE THE DATE

THE VIRTUAL CONFERENCE ON TEST SECURITY

November 8 & 9 2023