THE CONFERENCE ON TEST SECURITY

2025

UMASS AMHERST

14th Annual

Welcome to COTS 2025

Welcome to the house that Ron built!

We are so pleased to welcome you to the University of Massachusetts Amherst, home of both the Research, Educational Measurement, and Psychometrics (REMP) doctoral program; and the UMass Center for Educational Assessment. On this hallowed ground you will join one of the largest and most productive educational measurement graduate programs in the world, featuring six faculty, three postdoctoral research professors, and 20 Ph.D. students. Welcome! We are thrilled to have you here!

We are grateful to the COTS Conference Team for bringing what has now become one of the most important conferences in the assessment industry to UMass. The in-person conference program is overflowing with cutting-edge research and practices that are featured as symposia, paper sessions, panel sessions, and posters. The Conference Team has made all of our lives very difficult because we will have to choose between excellent concurrent sessions throughout each day. Thus, we thank all of you for submitting outstanding proposals and sharing your research with us. It is research on educational and psychological testing, and the pursuit of more fair and valid use of test results, that binds us all together; and binds us to the history of REMP and our Center.

When I was a graduate student (and as I say, "that was more than 5 years ago!") there were two universities that to me seemed like the most important places in psychometrics. The first was University of Upsala in Sweden, because that is where the LISREL program (for covariance structure modeling—see I am dating myself by not calling it structural equation modeling) came from. Thanks to Jörskog and Sörbom I could use LISREL to test the structure of all of the assessments we came across in graduate school, and even test and quantitatively evaluate competing psychological theories. The second university of importance in my young psychometric mind was here—the University of Massachusetts Amherst. Not having the strong math background to quickly absorb Lord and Novick, I found myself reading every word from every book and article coming out of UMass, written by Ron Hambleton, Hariharan Swaminathan, and their graduate students. It was through their writings that I learned item response theory; and that knowledge set me on the path that led me here (and I have been here more than 5 years!).

One step on my path to UMass was my position as Psychometrician for the Certified Public Accountants Licensing Exam. The test was administered twice a year, and part of my job was conducting forensic analyses on the post-exam data, which was essentially looking at the similarities of incorrect response choices across test takers in the same testing center. Test security was important then—but oh how times have changed. Advances in technology have helped us to develop better and more efficient tests, but these advances have also helped those who want to invalidly increase their test score, impersonate test takers, or even steal test material. Over the years, thanks in large part to COTS researchers, we have suppressed test fraud to the greatest extent possible, and the field has expanded from detection to prevention. The topics to be discussed at COTS 2025 illustrate important new issues in test security and important developments in not only fraud detection and prevention, but in better test development, scoring, and administration, too.

Returning to our history, I will note there was a consultant when I worked for the CPA Exam who became my mentor—Ron Hambleton. I describe REMP as "the house that Ron built" because Ron built the REMP program as his extended family. His love of testing and psychometric research was contagious, and he had a long legacy of recruiting, supporting, and graduating literally generations of psychometricians. I hope while you are here you experience this feeling of family, because you are now part of it. As you participate in the conference program, know that Ron, and our alumni who have passed on (Craig Mills, Dan Eignor, Mary Lyn Borque, Frank Stetz, Bob Chulu) are smiling down on us and are as pleased as we are that some of the most important research in the assessment industry is being presented and discussed here in our home. It is notable that the only other conference we have ever hosted was the "Ronference" (a conference to honor the research of Ron Hambleton) in this same conference center in 2012. For those of you who were there, welcome back. For those who were not, you can get the book here: http://bit.ly/4oIBW3D

Although we are super proud to be hosting the 2025 COTS conference, we cannot take credit for its overall quality. The Executive Committee for the Conference (Jim Wollack, Steve Addicott, Rachel Schoenig, Steve Erickson, Kim Brunnert, Carol Eckerly, and Claire McCauley) have worked enormously hard to put the program together and ensure an enjoyable and productive experience for us all. We remain grateful to them, and we appreciate the trust they put in us and the UMass conference team. Please don't hesitate to let me or any of the UMass team (Craig Wells, April Zenisky, Lisa Keller, Jenn Lewis, Javier Suárez Álvarez, Rebecca Woodland, Maura Maxfield, or Scott Monroe) know if there is anything we can do to make your conference experience more enjoyable. And although we encourage you to interact with and get

to know our graduate students, please don't steal any of them until they have at least completed their dissertation proposal!

Welcome and enjoy the conference!

Stephen G. Sireci
Executive Director, UMass Center for Educational Assessment

UMassAmherst
College of Education
Center for Educational Assessment

A SPECIAL THANK YOU TO OUR CO-HOST SPONSORS

UMassAmherst
College of Education
Center for Educational Assessment

caveon®

Honorlock®

duolingo english test

Pearson VUE

Respondus®

Integrity Advocate

THANK YOU TO OUR FRIENDS & DEMONSTRATORS

TALVIEW

PROMETRIC

TABLE OF CONTENTS

AGENDA

MONDAY, NOVEMBER 3

10:30 AM - 5 PM **Registration & Information**
Concourse Area, Campus Center 1st Floor

12 - 4:50 PM **Sessions**
Campus Center 1st Floor

5 - 6:30 PM **Opening Reception**
Marriott Center, Campus Center 11th Floor

6:45 PM **Optional Group Dinners**
(See pg. 73 for more info)

TUESDAY, NOVEMBER 4

7:45 AM - 5 PM **Registration & Info**
Concourse Area, 1st Floor

7:45 - 8:30 AM **Breakfast**
Campus Center Auditorium

8:30 - 9:30 AM **Opening Keynote**
Ballroom B, 1st Floor

9:40 - 10:40 AM **Sessions**
Campus Center 1st Floor

10:40 - 11:20 AM **Coffee & Networking**
Concourse Area, 1st Floor

11:30 AM - 12:30 PM **Sessions & Sponsor Demonstrations**
Campus Center 1st Floor

12:30 - 1:40 PM **Lunch**
Campus Center Auditorium
Campus Walking Tour
(Optional--Weather Permitting. See pg. 73 for more info.)

1:45 - 3:55 PM **Sessions & Sponsor Demonstrations**
Campus Center 1st Floor

3:55 - 4:25 PM **Break**

4:35 - 5:35 PM **Sessions & Sponsor Demonstrations**
Campus Center 1st Floor

5:50 - 7 PM **Poster Presentations**
Amherst Room & Foyer, Campus Center 10th Floor

7:15 PM **Optional Group Dinners**
(See pg. 73 for more info)

WEDNESDAY, NOVEMBER 5

7:45 - 11 AM **Registration & Info**
Concourse Area, 1st Floor

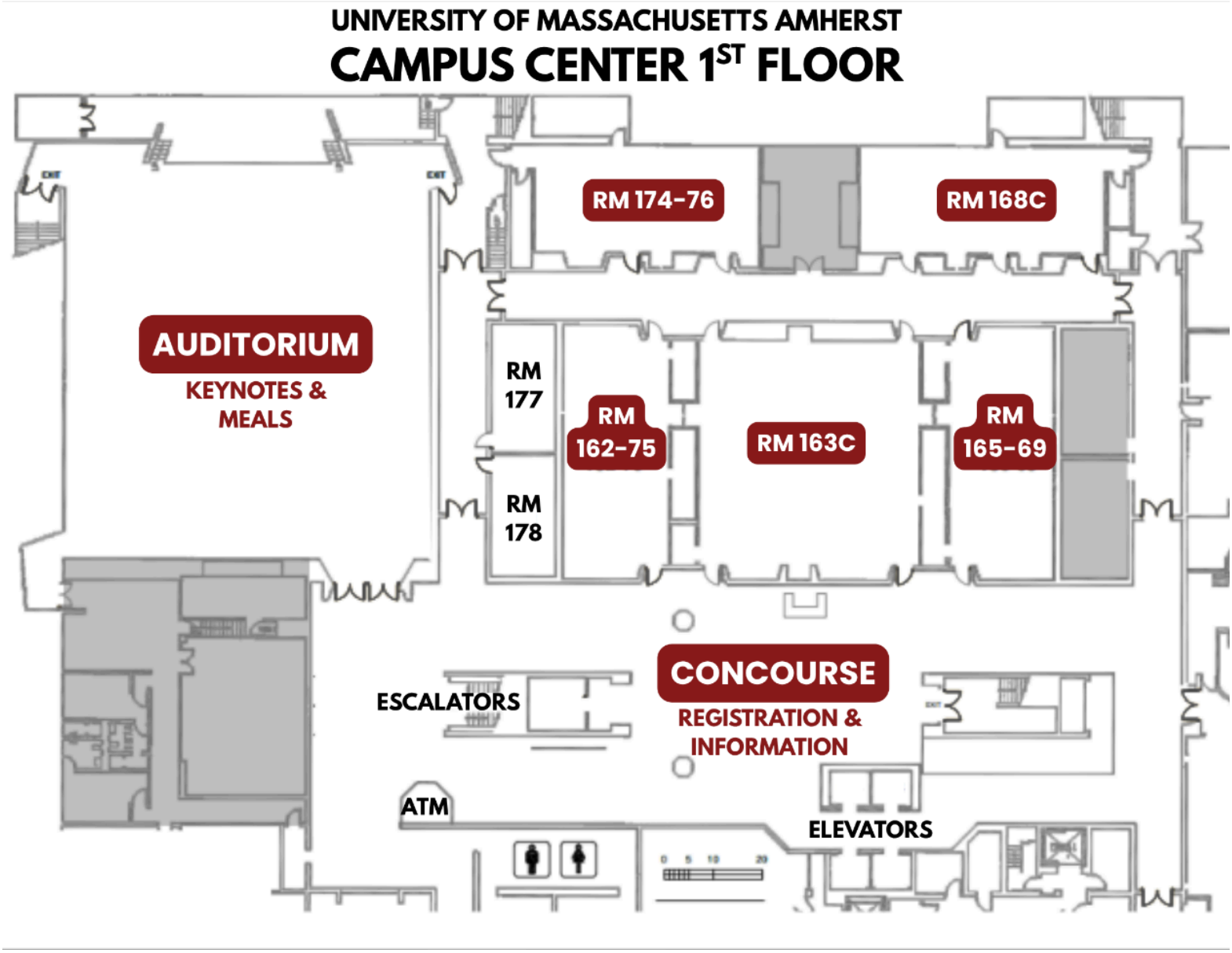
7:45 - 8:30 AM **Breakfast**
Campus Center Auditorium

8:30 - 11:50 AM **Sessions**
Campus Center 1st Floor

12 - 1 PM ***Closing Keynote & Debates***
Campus Center Auditorium

1 PM **Campus Walking Tour**
(Optional--Weather Permitting. See pg. 73 for more info.)

CAMPUS MAPS


UNIVERSITY OF MASSACHUSETTS AMHERST
CAMPUS CENTER 1ST FLOOR
RM 174-76
RM 168C
AUDITORIUM
KEYNOTES & MEALS
RM 177
RM 178
RM 162-75
RM 163C
RM 165-69
CONCOURSE
REGISTRATION & INFORMATION
ESCALATORS
ATM
ELEVATORS

CAMPUS MAPS


MAP KEY
0
500
1,000 Feet
Numbered Parking Lots
Metered/Public Parking
PVTA Bus Stops
N
E
S
W
A
B
C
D
E
1
2
3
4
5
6
To Agricultural Learning Center, Horse Barn
Wysocki
Kinney Renaissance Center
Amherst North Fire Station
To Tillson Farm, Intermediate Processing Facility (IPF)
Tillson House
UMass Police
Furcolo
Montague House
NORTH RESIDENTIAL AREA
North C
North A
North D
North B
McNamara
SYLVAN RESIDENTIAL AREA
Cashin
Brown
Totman
Nelson House North
To Mather
OLYMPIA DR.
Nelson House South
GOVERNORS DRIVE
Computer Science
Holocaust Center
EASTMAN LANE
Johnson
Leach
Dwight
Lewis
Hamlin
NORTHEAST RESIDENTIAL AREA
Thatcher
Auxilary Services Warehouse
PVTA Bus Garage
Engineering Lab
Astronomy
Duda
Gunness
Conte Polymer Research Center
Lederle Graduate Research Center
Arnold
Crabtree
Lyon
Engineering Lab II
Marston
Marcus
LGRC Tower
Knowlton
Observatory
Forest & Parks Buildings
Transit Facility
Paige
Knowles
Physical Sciences
East Experiment Station
West Experiment Station
Goessmann
Worcester DC
Holdsworth
Hatch
Parking Office
Bowditch
Agricultural Engineering
Stockbridge (Bowker Aud.)
Draper
Integrated Sciences Building
Life Science Laboratories
Dickinson
Grayson
ORCHARD HILL RESIDENTIAL AREA
Research Greenhouses
Chenoweth
Lincoln Campus Center (Hotel)
Hasbrouck
Hillside
Webster
Field
Art Barn
Flint
Cold Storage
Parking Garage
Skinner
Physical Plant
UWW
Student Union
Integrative Learning Center
NORTH PLEASANT STREET
Durfee Conservatory
University Health Services
Lorden Field
Practice Rink
Photo Lab
Machmer
IV
Morrill Science
Wilder
French
Central Heating Plant
Blaisdell House
Thompson
W.E.B. Du Bois Library
University Club
Brooks
Greenough
Mullins Center
Parks Marching Band
Grinnell
South College
CAMPUS POND
Shade Tree Lab
Franklin DC
Brett
CENTRAL RESIDENTIAL AREA
Baker
Van Meter
Chadbourne
William Smith Clark Memorial
Tennis Courts
Recreation Center
Dickinson Hall
Old Chapel
Morrill Greenhouses
Wheeler
New Africa House
CLARK HILL RD
Goodell
Sycamore
Elm
Memorial Hall
Fernald
Apiary
Butterfield
STRONG STREET
Waste Water Treatment
Kennedy Champions Center
COMMONWEALTH HONORS COLLEGE
Birch
Bartlett
Randolph W. Bromery Center for the Arts
Oliver Design Building
THATCHER ROAD
Gorman
Maple
Oak
Tobin
Curry Hicks
Business Innovation Hub
Studio Arts
Newman Center
MULLINS WAY
COMMONWEALTH AVENUE
Linden
Curry Hicks Cage
Herter
Isenberg
International Programs
Garber Field
Mahar Auditorium
472 No. Pleasant Street
Recreation Turf Fields
Boyden Gym
Munson
HAIGIS MALL
Alfond
Whitmore Administration
Middlesex
Berkshire
Army ROTC
Hampshire
MASSACHUSETTS AVENUE
Gordon
To Rte. 116, Rte. 9, I-91, Hadley Equestrian Farm, Telecommunications
Melville
Thoreau
James
Kennedy
Hampshire DC
Prince
Crotty Hall
NORTH PLEASANT STREET
New Construction
Hillel House
Emerson
Coolidge
Crampton
SUNSET AVE
LINCOLN AVENUE
Track & Field
Rudd Field
Sortino Softball Complex
SOUTHWEST RESIDENTIAL AREA
Hampden
MacKimmie
J.Q. Adams
Communication Disorders
Moore
Berkshire DC
J. Adams
FEARING STREET
Bowditch Lodge
Toddler House
Pierpont
Washington
Patterson
Farley Lodge
Cance
STADIUM DRIVE
UNIVERSITY DRIVE
TRIANGLE STREET
NORTH PLEASANT ST
Gladchuck Sports Complex
Jacobson Football Performance Center
McGuirk Alumni Stadium
To Development and Alumni Affairs at 380 Russell Street (Rte. 9)
To Amherst Center
YOU ARE HERE
LINCOLN CAMPUS CENTER (HOTEL)

Session Details

Monday, November 3

1 Behind the Mask: Social Engineering and the Hidden Threat to Test Security

Time:	12:00 PM – 1:00 PM
Location:	Room 168-74
Speakers:	Megan Rees & Heidi Green

We hear a lot about the impact technical exploits may have on the confidentiality and integrity of our examinations, but what about human manipulation and deceit? This session unpacks the tactics used to gain unauthorised access to test content, the impact of these breaches, and how one long-running, sophisticated case was uncovered. Attendees will learn how to work with law enforcement, prepare a prosecutable case, and recognise criminal operating models that may be targeting their examinations. Practical lessons and insights will help organisations build a resilient exam security framework.

2 From Pixels to Proctored: Relocating High-Risk Test-Takers to In-Person Exams through Scalable Online Modality Restriction

Time:	12:00 PM – 1:00 PM
Location:	Room 162-75
Speakers:	Jonathan MacKesson, Keith Becraft & Bryan Friess

This presentation examines an innovative approach to exam security within Amazon Web Services' (AWS) Certification Program, focusing on the strategic restriction of high-risk candidates to in-person testing environments. Rather than implementing complete program bans, AWS developed a targeted strategy of modality restriction for candidates with forensic flags or documented misconduct. Through strategic partnerships with Alpine and Pearson VUE, AWS automated this process to ensure scalable implementation while minimizing operational overhead. The presentation will share key metrics demonstrating the effectiveness of this approach, discuss critical operational challenges encountered during implementation, and explore the broader implications for the testing industry. We will also address the evolution of this security framework, including lessons learned,

necessary exceptions, and future developments in balancing test security with program accessibility.

3 Text Similarity Meets Test Security: Using Similarity Analyses to Identify Exposed Content

Time:	12:00 PM – 1:00 PM
Location:	Room 165-69
Speakers:	Russell Smith & Becky Ruedin

This presentation explores the use of multiple similarity statistics to compare exam content with material found on online "training" or braindump sites, as part of a comprehensive approach to protecting the integrity of high-stakes assessments. From a security management perspective, the session will highlight the efficiencies gained through automating the detection process, including the removal of subjectivity, streamlined identification of compromised content, and the ability to take swift and decisive action to mitigate risks. Additionally, the presentation will

delve into the psychometric and natural language processing (NLP) techniques used to analyze text similarity, such as cosine similarity and semantic similarity analyses, to quantify and understand overlaps between exam content and online material. By combining automation with advanced analytics, this approach provides a robust framework for safeguarding exams in an increasingly digital and interconnected world.

4 Anatomy of a Secure Test: Building Integrity from Start to Finish

Time:	12:00 PM – 1:00 PM
Location:	Room 163
Speakers:	Cicek Svensson, Isabelle, Gonthier, Jake Ritz, Susan Weaver, Ada Woo & Angela Bostwick

Test security is not a single step – it's a continuous process that must be embedded at every stage of an assessment's life cycle. This session offers a comprehensive walkthrough of the secure testing journey, from the initial stages of content development and item banking, through exam administration and monitoring, to post-test data analysis and score reporting. Attendees will gain practical insights into how to identify vulnerabilities at each phase, implement proactive safeguards, and build resilient processes that protect the integrity of their assessments.

Using real-world examples and case studies, we'll highlight best practices in secure item design, delivery platform controls, identity verification, data forensics, and incident response. Whether you are part of a certification body, educational institution, or test publisher, this session will provide a clear framework for integrating test security into your operations – no matter the scale or modality of your program.

5 Behind the Scenes – Test Security Experts Share Stories

Time:	1:10 PM – 2:10 PM
Location:	Room 168-74
Speakers:	Kim Brunnert, Danielle Bessette & Chris Glacken

Test security professionals face complex challenges daily: evaluating threats, identifying the root cause of security incidents, determining the threat level, and implementing effective countermeasures. This session includes real-world stories from ITS, an industry leader in test security, with one deep dive case study from Elsevier that will include how, why, what, and who. We will explore how we were alerted to something that sounded implausible, the investigation that ensued to reveal a proxy test taker, and what we needed to uncover the root cause. Beyond the technical detective work, we'll explore the broader implications: How do we balance transparency with security? What assumptions should we question? How do we turn incidents into opportunities for system strengthening? Attendees will gain practical insights into threat detection methodologies, investigation frameworks, and response strategies. Whether you're new to test security or a seasoned professional, you'll leave with actionable approaches to enhance your own security programs and incident response capabilities.

6 Solving the Commercial Cheating Mystery

Time:	1:10 PM – 2:10 PM
Location:	Room 163
Speakers:	Rachel Schoenig, Bryan Friess, Alana Chamoun & Kim Snyder

Commercial cheating services are actively working to undermine exam security and academic integrity. These services are well organized, technically savvy, and willing to break the rules to make a profit. And they would get away with it, too, if it wasn't for the industry taking a stand!

Join testing experts as they describe how these services are organized and how they target learners and test takers for profit. Together, we will explore the additional risks commercial cheating services pose to test takers, how they impact workforce readiness, competence and competitiveness, and the growing threat to national security. We will explore various means for identifying these services as well as policy, contractual, and legal remedies for addressing these entities. It's a fun, informative mystery that we'll solve together - and that's enough to make you want to say, "Jinkys!"

7 Evaluating the Effectiveness of Proctoring and Improving Test Security Through the Observation of Test Administrations

Time:	1:10 PM – 2:10 PM
Location:	Room 165-69
Speakers:	Brooke Westerlund & Marc Weinstein

Proctoring, whether conducted by humans, technology, or a combination of both, is broadly considered essential for maintaining exam security, preventing test content theft and cheating, and identifying violations that could compromise test validity. Observations made by trained individuals, either in real time or by reviewing recorded sessions, offer valuable insights into the effectiveness of proctoring practices. These observations may involve overt monitoring, where proctors know they are being observed, or covert methods, where individuals pose as examinees and attempt to subvert test protocols. For remote exams, full-session video recordings allow trained reviewers to assess proctoring quality and detect rule violations. Despite their value, findings from such observations are rarely published. However, this data offers a rich source of evidence to evaluate and enhance proctoring methods and determine whether additional security solutions are needed. It also enables comparisons between in-person and remote online proctoring, highlighting strengths and weaknesses of each. This session will present examples of observation data from both formats and share key takeaways to inform better proctoring practices and improve overall test security.

8 Latest Research and Development in Using Generative AI for State Assessment Programs: Cutting Edge R&D and Applications for K-12 Testing

Time:	1:10 PM – 2:10 PM
Location:	Room 162-75
Speakers:	John Olson, Stephen Sireci, Sergio Araneda & Walt Drane

In the past year, the use of Generative Artificial Intelligence (Gen AI) for educational assessments has exploded. Vendors, universities, and a variety of testing programs, including state assessments, are now using it. In 2025, most major test developers were already using it or at least experimenting with Gen AI and/or Large-Language Models (LLMs). In this session, information will be shared from the latest research and development at the Center for Educational Assessment at UMass and at Caveon Test Security, with examples on how Gen AI can be applied to state assessment programs. Many things have been learned at U-Mass and Caveon concerning the use of AI. Details of their R&D and findings will be provided. Among the many uses for state assessments are: better automated scoring, new tools for content (item and test) development, conducting field testing without students,

removing the need for test “forms”, tools for alignment studies, innovative ways of conducting psychometric analyses, use for improved test security. Attendees will gain valuable insights into recent R&D efforts of Gen AI for K-12 testing, its potential impact on future item/test development methodologies, and recommendations on how to apply AI for state assessments.

Workshop 1	**Foundations of Test Security in the Age of AI**
Time:	2:40 PM – 4:50 PM
Location:	Room 168-74
Speakers:	David Foster

This workshop will introduce and/or reinforce the fundamental concepts, and the effective application of those concepts, to test security, covering such topics as recognizing threats, risk management, test security planning, dealing with threats and breaches, communication, protection schemes, test design and test security, recent interesting scandals, and, of course, AI's impact on test security,

The session will be highly interactive with hands-on exercises, lively discussion, entertainment, demonstrations of technology, and so much more. Attendees will leave the session more prepared to succeed within their particular test security circumstances.

If you are wondering if you would benefit by attending, then the answer is Yes. While the topics are fundamental, they are not simple. They require repeated exposure and use to become proficient.

Workshop 2	**Identifying Kryptonite and Activating your Super Powers**
Time:	2:40 PM – 4:50 PM
Location:	Room 163
Speakers:	Rachel Schoenig, Camille Thompson, Ray Nicosia, Paul Muir, Jarret Dyer, Claire McCauley, Bryan Friess, Daniel Clough, & Kelly Robinson

This hands-on workshop will explore the state-of-the-art risks plaguing the testing industry today – the kryptonite that can weaken our testing programs. Together, we will discuss and demonstrate methods for undermining exam security and engaging in exam misconduct during the test event. From AI and AI agents to virtual proxies, deep fakes, and new technology aids, presenters will share how these tools are being used, distributed, and accessed by students, trainees, and test takers. In addition, we will discuss the risks for different modalities of delivery and monitoring services. Attendees will then work together to identify the exam security superpowers that we can activate to help guard against these risks and the new tools we need to develop to take security to the next level. (Attendees should bring their laptops for this session if possible).

9 Detecting Compromise Using Response Time

Time:	2:40 PM – 3:40 PM
Location:	Room 165-69
Speakers:	Nicholas Trout, Kylie Gorney, Danielle Lee, Kylie Gorney, Cassie Chen, Jason Underwood, & Greg Hurtz

Enhancing the Use of Response Times in Answer Similarity Analysis
(Nicholas Trout & Kylie Gorney)
Answer similarity statistics such as the ω statistic (Romero et al., 2015; see also Wollack, 1997) are frequently used as tools for detecting test collusion, but may offer room for improvement. For example, recent research has shown that an extended version of the ω statistic, known as the weighted ω statistic, tends to perform better than the original (unweighted) version (Trout & Gorney, 2025). In addition, the use of the unweighted ω statistic with response times has been shown to perform better than the unweighted ω statistic without response times (Gorney & Wollack, 2024). Taken together, it seems reasonable to believe that detection rates could be improved even further if response times were used with the weighted ω statistic. The purpose of this paper is to investigate this research question. In this study, we compare the performance of several new and existing versions of the ω statistic using detailed simulations with several manipulated factors. Preliminary results suggest that the new versions of the ω statistic control the Type I error rate and are more powerful, on average, than existing versions.

Using Item Scores and Response Times to Detect Item Compromise in Computerized Adaptive Testing
(Danielle Lee, Kylie Gorney, & Cassie Chen)
Sequential procedures have been shown to be effective methods for real-time detection of compromised items in computerized adaptive testing. In this study, we propose three item response theory-based sequential procedures that involve the use of item scores and response times (RTs) and compare them with classical test theory-based sequential procedures. The first procedure requires that either the score-based statistic or the RT-based statistic be statistically significant, the second procedure requires that both the score-based statistic and the RT-based statistic be statistically significant, and the third procedure requires that a combined score and RT-based statistic be statistically significant. Results suggest that the third procedure is the most promising, providing a reasonable balance between the false positive rate and the true positive rate across a wide range of simulation conditions.

Analysis of Similarly-Aberrant Item Response Time Patterns between Pairs of Test Takers

(Jason Underwood & Greg Hurtz)

Excessive similarity in item responses is a longstanding method for detecting aberrant testing behaviors such as collusion, shared pre-knowledge, and proxy texting. Such aberrant behavior can also result in atypical item response time patterns that are detectable as misfit to the lognormal response time model (e.g., Hurtz & Mucino, 2024; Marianti et al., 2014; Sinharay, 2018). When two test takers share preknowledge of the same set of items, they may exhibit similar patterns of misfit in their response times, meaning similar items on which their response times deviate from model predictions. Because these residuals are expected to be independent and random under normal conditions, a strong correlation between two test takers' residuals calls their behavior into question (van der Linden, 2009). We explore this expected relationship in real-world datasets and explore associations with response similarity indices.

10 Science Under Scrutiny: Tips for Testifying on Test Security

Time:	3:50 PM – 4:15 PM
Location:	Room 165-69
Speakers:	Sarah Toton & Marc Weinstein

Testifying on test security can be a daunting challenge, especially when your credibility, expertise, analytical methods, and statistical results are scrutinized in court. This session provides essential tips for experts called to testify in test security cases, covering how to present and defend your analysis and results with clarity and confidence.

This presentation offers a practical framework for navigating the complex landscape of expert testimony. Participants will learn how to successfully establish their expertise, prepare for legal proceedings, communicate complex technical information with clarity, and anticipate common challenges from opposing counsel. Practical tips will be given on courtroom demeanor and maintaining credibility through precise language and ensuring your interpretations and conclusions are firmly within the scope of the analysis and your domain of expertise. The session will also address the critical skill of acknowledging

uncertainty without undermining your analysis, breaking statements down into components and using precise language to enhance credibility rather than weaken it.

Attendees will leave with a clear, actionable framework for communicating data forensics findings clearly, credibly, and confidently in legal settings.

11 Pennsylvania State Keystone Exams-PSSA On-Site Monitoring Program (OMP) - The Gold Standard

Time:	4:25 PM – 4:50 PM
Location:	Room 165-69
Speakers:	Craig A. Weller

The Pennsylvania Department of Education's (PDE) Keystone Exams-PSSA On-Site Monitoring Program (OMP) is considered the "Gold Standard" of large-scale high-stakes standardized test security monitoring programs in the United States.

We have successfully completed approximately 2,100 on-site KE-PSSA monitoring visits over the past eleven (11) school years. The feedback that we provide to Local Education Agencies (LEAs)/schools from these visits has been instrumental in ensuring test administrations are completed in the most secure and appropriate manner. This safeguards that valid and reliable data is being pushed forward for district and school-level accountability. This 25-minute session will detail a quick history of the Pennsylvania KE-PSSA OMP and share ideas for ensuring other states have the opportunity to expand their monitoring programs.

Tuesday, November 4

Opening Keynote	**Panel: Let's Talk Tech!**
Time:	8:30 AM – 9:30 AM
Location:	Campus Center Auditorium
Panelists:	Claire McCauley, Basim Baig, Daniel Clough, Bryan Friess, & Jim Wollack
Moderator:	Rachel Schoenig

Technology is neither good nor bad, but it's not neutral, either. In the hands of individuals intent on engaging in test fraud or theft, technology can be a massive threat to testing programs. On the other hand, in the hands of testing programs, there are many technological advances that can be leveraged to benefit test takers. This session will talk all things tech and provide an update on the latest in tech advances – and threats!

12 The ABC's of the Legal Foundation for Exam Security: A Primer

Time:	9:40 AM - 10:40 AM
Location:	Room 163
Speakers:	Camille Thompson, Mike Clifton, Jennifer Semko & Rachel Schoenig

Curious about the latest copyright office protections for secure exams and the impact of AI? Interested in the latest impact of privacy or AI laws on exam security? Wanting to position your program to take action against individuals who breach your contract or engage in ethics violations? This session will discuss the ABC's of the legal underpinnings of standardized testing, content protection, and exam security. We will use hypotheticals and case studies to explore the legal foundations protecting your testing program and how you can leverage those to enhance exam security. Attendees will leave feeling more confident and informed when it comes to the legal aspects of their exam security framework.

13 The Evolution of a Teacher Licensure Testing Program – Changing the Tires While Driving the Car

Time:	9:40 AM - 10:40 AM
Location:	Room 168-74
Speakers:	Alex Lapointe & Amy Schmidt

Customized state-run teacher licensure programs face a wide range of challenges, from responding to urgent legislative priorities in a timely manner all the way to determining how to set cut scores on tests with small sample sizes. The Office of Postsecondary Assessment in Florida has been dealing with these challenges for many years in a proactive way, turning them into opportunities and ensuring the integrity of their examinations while meeting the need to certify qualified professionals to serve the people of Florida. During this session, two members of the Florida Office of Postsecondary Assessment and two representatives from their testing vendor, Evaluation Systems group of Pearson, will discuss how the program has evolved and flourished over the past two decades and how it has embraced technological improvements in item development, test construction, scoring, and equating, and test administration, all while facing new security challenges and new legislative and societal imperatives.

14 Preknowledge and Fairness Issues in Test Security

Time:	9:40 AM - 10:40 AM
Location:	Room 165-69
Speakers:	Stuart Barnum, Min Liang, Peter Tran, Craig Wells, Michael Fauss, & Ikkyu Choi

Functional Network Analysis in Identification of Preknowledge in Speeded Assessments

(Stuart Barnum & Min Liang)

We investigate application of functional network models for detection of preknowledge in moderately speeded medical assessments, using simulated and real data. Because of rapid guessing in the assessments, together with research indicating that response times for correct answers tend to be lower for examinees with preknowledge than for other examinees answering the same items, links between persons in the networks are weighted according to both the similarity of

measures of decreases in responses times from baseline values deemed unaffected by preknowledge, and the magnitudes of the measures of decrease. Such a relationship between persons may be expressed as the vector dot product between the sequences of the measures for the persons.

Communities within networks may correspond to sets of compromised items shared by persons who correctly answered the compromised items more quickly. Community detection methods employed include Louvain, Leiden, and Girvan-Newman, together with visualizations of the networks, with the network-analysis results compared to other methods including k means. Simulated data is from mixtures of persons with preknowledge for some items, persons rapidly guessing for some items, and persons fitting a hierarchical model with response accuracies following an IRT model and response times following a lognormal model.

An Examination of Randomly Parallel Test Forms
(Peter Tran & Craig Wells)
Cheating introduces construct irrelevant variance that negatively influences test scores leading to compromised interpretation. One way to mitigate the effects of cheating is to use randomly parallel tests (RPTs) where each examinee receives a unique test form. The RPT forms are based on randomly sampling items from a unique test form. The RPT forms are based on randomly sampling items from a very large bank so that there is very little to no overlap between examinees and the issues of pre-knowledge is eliminated. However, there are concerns with RPT forms – mainly that some examinees may receive a more difficult or easier form due to chance. In this study, we examine via empirical and simulated data the strengths and challenges in implementing RPTs forms in practice, especially with respect to the psychometric properties of RPTs. Our analyses will focus on comparing RPTs with traditionally equated forms and will emphasize the practical consequences attached to score interpretations.

Bayesian Anomaly Flagging with Fairness Constraints
(Michael Fauss & Ikkyu Choi)
This study investigates the problem of automatically flagging test takers who exhibit atypical responses or behaviors for further review by human experts. The objective is to develop a flagging policy that optimizes the identification of test

takers who warrant additional scrutiny while satisfying constraints that promote fairness across given but possibly overlapping groups of test takers. Specifically, we consider upper bounds on the groupwise false positive rate (FPR) and the groupwise false discovery rate (FDR) as fairness metrics. The fair-flagging objective is formalized in a Bayesian setting, and the corresponding optimal policy is derived. Since calculating this policy and the underlying posterior distributions is computationally infeasible in practice, a variational approximation and a heuristic policy based on a Bayesian variant of oversampling are proposed. This approach modifies the effective group sizes in the inference step to control the fairness metrics in the flagging step. The performance of the proposed policy is assessed via numerical experiments using both synthetic and real data and compared against alternative, off-the-shelf methods.

15 Exploring Preknowledge Mitigation Techniques with Real Data

Time:	9:40 AM - 10:40 AM
Location:	Room 162-75
Speakers:	James Wollack, Merve Sarac, Angelica Mulfinger, Russell W. Smith & Carol Eckerly

When it comes to test security, we often talk about taking a comprehensive approach to combatting test fraud that involves three stages: prevention, detection, and enforcement. Typically, prevention refers to actions taken prior to test administration to thwart fraud, detection refers to identifying potential fraud post-administration, and enforcement is action taken based on discoveries made in the detection stage. A fourth stage of increasing interest is mitigation. Mitigation falls between prevention and detection, allowing programs to identify and act in real time to reduce the harmful impact of test fraud on program integrity. The aberrant behavior of interest in this presentation is examinee preknowledge. The researchers have set out to explore the application of three different methods for detecting potential preknowledge in real time and applied them to the same real exam dataset with an eye toward real-world applicability of mitigation strategies. We will evaluate these different detection approaches with respect to various exam structures and their potential to mitigate the preknowledge effects with real time decisions.

Sponsor Demonstrations	
Time:	11:30 AM - 12:30 PM
Location:	Room 168-74
Speakers:	Brandon A. Smith, John Dight, & Jarret Dyer

Integrity Without Intrusion: Human-Centric Approaches to Continuous Test Security

Integrity Advocate (Brandon A. Smith)

11:30 AM - 11:55 AM

Traditional test security often relies on front-end identity checks or invasive monitoring technologies. But maintaining exam validity and protecting program reputation requires a broader, more ethical approach.

In this session, Brandon Smith, CEO of Integrity Advocate, will share lessons learned from working with global organizations to implement continuous yet non-intrusive test security practices. These strategies combine identity assurance, real-time monitoring, and contextual risk detection — all designed to respect participant privacy while upholding the highest standards of fairness and integrity.

Attendees will

- Examine how continuous assurance models reduce risks like proxy testing, collusion, and unauthorized resource use.
- Explore the balance between security and test-taker experience, ensuring accessibility and compliance with privacy expectations.
- Learn practical steps for building a holistic test security program that goes beyond compliance to reinforce trust, validity, and brand reputation.
- Understand how transparent communication and ethical safeguards strengthen acceptance of security measures by all stakeholders.

By moving beyond single checkpoints and invasive methods, programs can create a sustainable culture of integrity — protecting results while safeguarding the dignity of test-takers.

Turbocharge item production without jeopardizing accreditation, with Surpass Copilot

Surpass (Jarret Dyer & John Dight)

12:05 PM - 12:30 PM

Discover how to revolutionize your item development process using AI. In this engaging session, we'll present a live demonstration of Surpass Copilot, the

powerful AI-driven item generation feature-set seamlessly integrated into the Surpass Platform.

Learn practical strategies and best practices for leveraging AI to efficiently generate high-quality assessment content while ensuring compliance with accreditation standards. We'll guide you through the critical steps to document your AI processes effectively, helping you stay ahead of regulatory frameworks like the EU AI Act and avoid common pitfalls such as content hallucinations and copyright concerns.

By the end of this session, you'll gain a clear understanding of how to significantly expand the depth and breadth of your item pool with ease - without compromising on quality, integrity, or compliance.

Need more items to complete your test security strategy? Turbocharge item production without jeopardizing accreditation, with Surpass Copilot.

16 The Sorting Hat Speaks: Placing AI on the Spectrum from Test Security Hero to Villain

Time:	11:30 AM - 12:30 PM
Location:	Room 163
Speakers:	Paul Muir, Isabelle Gonthier, Mike Sobczak, Rachel Schoenig & Marc Weinstein (*ATP Test Security Committee*)

AI has arrived at the gates of test security, but is it a Gryffindor-worthy protector or a Slytherin saboteur? Our engaging panel explores AI's dual nature: a powerful ally capable of detecting subtle cheating patterns or an accomplice that facilitates deceitful tactics such as sophisticated deepfakes and algorithmic cheating strategies.

Beyond security, we'll sort through ethical questions (When does AI surveillance cross the line?), moral responsibilities (How transparent must AI be?), and intricate legal considerations (Who owns accountability when AI misbehaves?).

In this truly interactive session, join our panel of experts, test security professors, legal counsellors, and ethics wizards, to bravely tackle these issues and determine the ultimate Hogwarts house for AI in testing.

17 Exam Security Investigations – Recommended Best Practices In Light of Recent Threats

Time:	11:30 AM - 12:30 PM
Location:	Room 165-69
Speakers:	Brent Morris, Brent Hill, Harry Samit & Bryan Friess

The last five years have seen many changes in the certification industry. The expansion of online testing, the emerging yet increasingly crucial role of artificial intelligence (AI), and the use of ever more creative item types are just a few of these.

Unfortunately, while industry leaders have pushed the boundaries of what is possible, so too have organized criminal theft rings bent on stealing and selling valuable intellectual property or orchestrating ever-more sophisticated proxy tests.

These threats call for a robust exam security framework to protect the integrity of exam programs. In addition to vigilant and innovative security measures, securing an exam program also requires having a strong investigative capability able to flag misconduct and then determine exactly what happened. The evidence and intelligence resulting from investigations must flow to stakeholders to allow exam sponsors to make certification decisions.

New innovated techniques: AI Photo Comparisons, Machine ID and Network Name matches, Candidate interviews, Human Source Operations, Undercover Operations, Law Enforcement Referrals

18 Innovative Monitoring Solutions

Time:	11:30 AM - 12:30 PM
Location:	Room 162-75
Speakers:	Sanjoe Jose, Steve Addicott, & Aleia Kim

Threats to Assessment Security in the Age of AI and How to Prevent It
(Sanjoe Jose)
As AI technology advances, it is transforming the way we assess skills and knowledge—but it is also giving rise to new and sophisticated threats to exam

security. From AI-generated responses to deepfake impersonations and remote access tools, traditional proctoring measures are no longer enough to ensure exam integrity. Organizations must adopt a proactive, AI-driven approach to combat these evolving risks.

In this session, Sanjoe Tom Jose, CEO of Talview, will share insights from his experience leading AI-powered assessment solutions for some of the world's top certification bodies, enterprises, and educational institutions. He will discuss how AI can be both a threat and a solution in the realm of high-stakes assessments and introduce best practices that organizations can implement today.

Disrupting Test Security- A Smarter, Data-Driven Approach to Test Monitoring (Steve Addicott & Aleia Kim)
In an era where trust in test results is eroding and candidate experience is under scrutiny, exam security is at a crossroads. For decades, proctoring—whether in-person or remote—has been treated as the gold standard. But research, including secret shopper studies, data forensics, and internet monitoring, reveals that it often fails to prevent cheating, protect content, or ensure valid outcomes.

Examinees use hidden cameras, AI tools, proxies, and smart devices to circumvent detection. Secret shopper engagements show that proctors miss violations in over 90% of cases—ranging from unauthorized breaks to AI-assisted cheating. Even when misconduct is detected, consequences are inconsistent, while test content is stolen and reused.

Caveon Observer offers an alternative. Rather than relying on visual observation, Observer monitors test session data and collects behavioral and response data in real time. It then calculates dynamic risk scores for every test session, notifying administrators only when a session crosses a predetermined risk threshold. This targeted approach minimizes intrusion, reduces cost, and drastically improves the ability to detect and respond to threats as they occur.

This session will share evidence on the limitations of current security models and explore how a data-driven approach can improve fairness, accuracy, and trust in test results.

Sponsor Demonstrations	
Time:	1:45 PM - 2:45 PM
Location:	Room 168-74
Speakers:	Chris Glacken, Heather Klesch, Bryan Friess, & Harry Samit

Lurking in the Shadows: Test Security Monsters
Internet Testing Systems (Chris Glacken)
1:45 PM - 2:10 PM
Test security threats often hide in plain sight—are you prepared to confront them? This engaging session explores advanced strategies, best practices, and emerging trends in test security. Attendees will gain insights into how to optimize secure browser deployment, tackle complex and evolving security challenges, and stay ahead of potential "monsters" in the testing landscape. Whether you're

strengthening your current defenses or preparing for what's next, this session offers practical takeaways and a fresh perspective on safeguarding assessment integrity.

Securing State Teacher Certification Programs
Pearson VUE (Heather Klesch, Bryan Friess, Harry Samit)
2:20 PM - 2:45 PM
When bad actors threaten the integrity of exams used for the licensure of teachers, it puts students at risk. These cases present some unique challenges and opportunities and usually make for exciting investigations. In this presentation the team will discuss threats to educator licensure programs and will offer some examples of cases which resulted in prosecution at the state and federal levels for those responsible.

19 Using Technology to Manage and Combat Security Issues

Time:	1:45 PM - 2:45 PM
Location:	Room 163
Speakers:	Whitney Chandler, Amy Mann, & Paul Morales

Case Closed: How Salesforce Transformed Our Test Security Operations
(Whitney Chandler & Amy Mann)
Test security operations often involve a wide variety of cases, each with distinct requirements and workflows. At the College Board, we needed a flexible, efficient system to manage these cases, from initial intake and investigation through to resolution while fostering seamless collaboration across teams.

In this session, we'll share how we leverage Salesforce to manage the complete lifecycle of test security cases, from intake and triage to investigation, communication, and closure. Our approach centers on flexibility, centralized tracking, and cross-functional collaboration, all within a single, integrated platform.

We'll highlight key system features, lessons learned, and real-world examples of how this setup has enhanced visibility, standardized processes, and enabled us to adapt over time. Whether your team is just getting started or refining existing tools, we hope you'll leave with practical insights to strengthen your own test security efforts.

Enhancing Test Security Amidst Technological Evolution

(Paul Morales)
Test security professionals must stay ahead as cheating methods evolve with technological advancements. This session delves into the latest technological tools and methodologies designed to detect and prevent sophisticated forms of exam misconduct. Attendees will gain insights into practical applications of these technologies, supported by real-world demonstrations.

20 Novel Detection Strategies

Time:	1:45 PM - 2:45 PM
Location:	Room 165-69
Speakers:	David Budd, Chenhao Niu, Irina Grabovsky, & Rich Feinberg

Catching Imposters using Voice Recognition
(David Budd)
Test fraud seriously undermines the integrity of exams and erodes the confidence stakeholders have in our products. The risk of malpractice is particularly high for tests taken at home, and remote proctored invigilation is limited in what it can detect. A popular ruse is for candidates to employ proxy testers, or imposters, to take tests on their behalf, using various tactics to avoid detection. Join us to discover how we implemented voice recognition technology to flag multiple test takers with the same voice, track serial imposters and make decisions about who to disqualify. We'll cut through the technical jargon to show you our game changing tool at work and demonstrate its success. Expect an engaging and informative session that will get you thinking about your own test security toolkit.

Detecting LLM-Assisted Cheating in Open-Ended Writing
(Chenhao Niu)
The increasing capabilities of large language models (LLMs) introduce new challenges to the security of open-ended tasks in high-stakes online assessments. This presentation introduces a practical solution developed by the Duolingo English Test (DET) to help detect LLM-assisted cheating, particularly when test takers manually type AI-generated responses due to restrictions on copy-pasting. Unlike earlier approaches that depend heavily on synthetic data, our framework is designed around real-world behaviors, accounting for common human errors introduced during manual transcription.

Our approach enhances text classification with contrastive learning and self-training. These techniques help the system detect AI-generated content even after it has been modified. The model achieves improved performance while maintaining a low false positive rate, which is critical for use in a high-stakes setting.

This talk will focus on the real-world applications of this framework in test security and its integration with human review workflows. Attendees will gain insight into how this method supports existing proctoring strategies, contributes to test validity, and reflects DET's ongoing efforts to address evolving AI-related threats through research-informed innovation.

Identifying Clusters of Test Collusion with Random Forest Method
(Irina Grabovsky & Rich Feinberg)
This research aims to assess the sensitivity and utility of Random Forest as a forensic tool for detecting group-level patterns in test collusion.

21 Strengthening Test Security Through Policy and Manual Revisions

Time:	1:45 PM - 2:45 PM
Location:	Room 162-75
Speakers:	Timothy Butcher, Walt Drane, Jessica Fenby & John Olson

In an era of rapidly evolving assessment technologies and increasing threats to test integrity, robust and adaptive policies are essential to safeguarding the validity of test results. This presentation explores the critical role of policy and procedural manual revisions in strengthening test security across educational testing environment. Drawing on recent case studies and best practices, we will examine how outdated or ambiguous policies can create vulnerabilities, and how systematic revisions can close security gaps, clarify responsibilities, and enhance compliance.
The session will outline a structured approach to policy review, including stakeholder engagement, risk assessment, and alignment ethical standards.

Attendees will gain insights into effective strategies for revising test administration manuals to reflect updated protocols, incorporate digital testing considerations, and support training and enforcement efforts. Emphasis will be placed on creating

clear, actionable, and adaptable documentation that supports both prevention and response.

Participants will leave with practical tools and a framework for initiating or refining their own policy and manual revision processes, ensuring their organizations are better equipped to maintain the integrity of their assessments in an increasingly complex testing landscape.

Sponsor Demonstrations	
Time:	2:55 PM - 3:55 PM
Location:	Room 168-74
Speakers:	Steve Addicott, David Foster, Jordan Adair, & Patrick Devlin

Beyond Test Security Theater: Building Security That Works
Caveon (Steve Addicott & David Foster)
2:55 PM- 3:20 PM
Too often, test security has focused on visible but ineffective measures that make exams more expensive, intrusive, and frustrating for test takers. Proctoring and other "catch-the-cheater" approaches fail to prevent misconduct and rarely deliver the promised protection.

This session will present an alternative: test security built directly into the exam itself, paired with monitoring that only intervenes when it is truly needed. We will introduce Observer PLUS, a system that combines secure test design with AI-powered monitoring to prevent theft and cheating before they happen while reducing costs and improving the test-taker experience.

Participants will see a brief demonstration, hear insights from real-world use, and consider what it means for the future of fair and secure testing.

Proactive Assessment Strategies: Reimagining Test Security and Integrity
Honorlock (Jordan Adair & Patrick Devlin)
3:30 PM - 3:55 PM
While it is often suggested that online exams may lead to increased dishonest behavior, is this idea accurate? According to the Wiley survey titled "New Insights into Academic Integrity," reinforced the belief that cheating is more common during online exams. Interestingly, however, the same study showed that 70% of test takers stated they would be less likely to cheat if online exam proctoring were implemented. In fact, another survey, "Cheating on Unproctored Online Exams,"

revealed that only 15% of test takers attempted to cheat when proctoring was involved.

Join Honorlock experts for an engaging discussion based on data-driven insights and trends into exam dishonesty. They will delve into prevalent exam violations, share effective strategies for fostering integrity in assessments, and demonstrate how technology like proctoring can reinforce exam security in both online and traditional testing environments.

22 The Costs of Labor vs. Technology in Remote Assessment: When the Pendulum Swings Too Far

Time:	2:55 PM - 3:55 PM
Location:	Room 163
Speakers:	Angela Bostwick & Sarah Toton

The increasing integration of artificial intelligence (AI) and automation in remote assessment has sparked a debate regarding the balance between labor and technology in proctoring and psychometric evaluation. While AI offers benefits such as scalability, efficiency, and potential cost savings, an over-reliance on technology can result in unintended consequences, including higher costs, compromised security, and reduced fairness for test-takers. This session examines the hidden costs associated with AI-driven proctoring and automated psychometric analysis, such as false flagging, appeals, and legal challenges. It also highlights the value of live human oversight, showing how experienced proctors and psychometricians enhance both the integrity and fairness of assessments, offering more cost-effective and scalable alternatives. The session will advocate for a balanced approach, emphasizing the importance of combining human expertise with technology to avoid the pitfalls of over-automation and inflated software expenses. Attendees will gain practical insights on optimizing assessment strategies to ensure security, fairness, and cost-effectiveness.

23 Effective Whack-A-Moling: Strategies and Benefits of Web & Social Media Monitoring and Intervention

Time:	2:55 PM - 3:55 PM
Location:	Room 165-69
Speakers:	Becky Ruedin, Frank Aceves & Jake Ritz

Social media platforms and the Web are used to target learners and testers to influence them, dupe them, and profit from them by promoting academic dishonesty.

The basic approach for identifying compromised content and trying to get it removed from the web can be challenging, slow, and frustrating. When one URL is removed, it is often just replaced with a new one. But, it doesn't have to only be an exhausting game of whack-a-mole. In combination with other test security methods (e.g., content refresh, effective proctoring), an active Web & Social Media intervention program may be an important layer of defense.

In this session we will explore how bad-actors market their services, how to identify and disrupt them, how they change tactics, and how to realize real benefits for your testing program.

24 Improving the Integrity of Innovative Balanced Assessment Systems for States: Steps Being Taken by SEAs to Ensure Secure Implementation

Time:	2:55 PM - 3:55 PM
Location:	Room 162-75
Speakers:	John Olson, Shannon Jordan, Timothy Butcher & Garron Gianopulos

In 2025, one of the more promising, yet challenging issues in states is their increasing adoption of innovative Balanced Assessment Systems (BAS) that may include a combination of summative assessments, Interim/Benchmark Assessments (I/BM), and/or Thru-Year Assessments (TYA). This more balanced approach to testing has gotten the attention of many Governors, State Education Agencies (SEAs), and State Legislators, and in recent years, increasing numbers of states have moved to implementing a BAS instead of an I/BM or TYA. Many states are realizing that it is essential to implement all of the assessments used in a BAS with fidelity and integrity, so they yield valid and reliable outcomes. To gain a fair and accurate understanding of student performance on I/BMs, TYAs, etc., test

security needs to be as robust as that typically seen in statewide summative assessments used for accountability. In this session, presenters will discuss steps being taken to implement new BAS along with approaches added for maintaining test integrity and security. State assessment directors and other experts will present details on what is being done to improve security for all parts of state assessment programs so the entire BAS is well protected.

Sponsor Demonstrations	
Time:	4:35 PM - 5:35 PM
Location:	Room 168-74
Speakers:	Sanjoe Tom Jose & William Belzak

Agentic AI in Action: Reimagining Test Integrity with Alvy
Talview (Sanjoe Tom Jose)
4:35 PM - 5:00 PM
Join Sanjoe Tom Jose, CEO and Co-Inventor at Talview, for an inside look at Alvy, the world's first patented Agentic AI Proctoring system. Designed to detect and respond to AI-assisted cheating techniques, including tools like Cluely, misuse of ChatGPT, and deepfakes, Alvy operates autonomously to preserve integrity in high-stakes assessments.

This session will showcase how agentic AI actively perceives, interprets, and adapts to candidate behaviour, offering not only unmatched security but also real-time support and guidance to ensure a fair, stress-minimised experience for test takers.

Quantifying Proctor Judgment in High-Stakes Testing
Duolingo (William Belzak)
5:10 PM - 5:35 PM
As high-stakes exams move online, ensuring their integrity poses both technical and human challenges. In this demo, we will showcase the Professional Calibration Tool (PCT), a psychometric system designed to quantify and improve the quality of human judgment in proctoring- a critical but often overlooked aspect of test security. Like other professionals, proctors make high-stakes decisions under uncertainty, with outcomes shaped by attention lapses, inconsistent training, and individual biases. The PCT addresses these issues by embedding stealth assessments (pre-adjudicated cases) into live workflows and delivering immediate feedback based on expert and peer consensus. The demo will highlight how the tool works in practice and its potential to enhance decision accuracy and consistency in remote proctoring.

25 Exploring the Impact of AI-Based Item Pool Expansion

Time:	4:35 PM - 5:35 PM
Location:	Room 163
Speakers:	David Foster, Anna Rubin, Isabelle Gonthier, & Rachel Schoenig

Randomly Parallel Tests for Cisco Systems Certifications: A Modern Solution to Better Testing and Test Security
(David Foster & Anna Rubin)
Randomly Parallel Tests (RPTs), first proposed by Fred Lord in 1955, involve generating unique test forms for each examinee by randomly sampling items from a large common pool. Though once impractical, modern technology makes RPTs easy to implement. They offer key advantages: scores that generalize to the full domain, greatly reduced test theft and cheating, and virtually unlimited test forms for reuse. Despite these benefits, concerns persist about fairness, reliability, and validity—since traditional test design relies on carefully assembled, equated forms. To evaluate RPTs in practice, Cisco Systems partnered with Caveon to administer one RPT and ten traditional equated forms. Data collected through May 2025 enabled a detailed psychometric comparison across all formats.

This presentation shares findings from that real-world study, highlighting how RPTs stack up against traditional approaches on metrics like score reliability and pass/fail consistency. A follow-up simulation further assessed how reliably RPTs support high-stakes decisions. The results offer compelling insights into how RPTs can meet professional testing standards while solving long-standing problems of test security and scalability.

More items, more risk? Securing AI-driven test development pipelines for scalable content security
(Isabelle Gonthier & Rachel Schoenig)
AI-assisted content generation is accelerating the way test items are developed, enabling credentialing programs to scale faster, reduce item exposure, and strengthen test security through expanded item banks. But the promise of “more items equals more security” only holds if the AI development process itself is secure, transparent, and psychometrically sound.

This session explores the intersection of AI-enabled test development and content security, with a focus on securing the pipelines as well as the outputs. Attendees will examine the tools, workflows, and guardrails required to maintain trust in

AI-generated items at scale. Topics include model transparency, source data integrity, SME collaboration, and human-in-the-loop oversight. Technical risks such as prompt leakage, uncontrolled content drift, and psychometric compromise will also be addressed. We'll also address technical risks such as sensitive information being unintentionally revealed by AI, AI-generated content changing unpredictably over time, and test items that fail to meet quality or fairness standards.

Through case-informed discussion, the session will offer applied strategies for testing organizations to evaluate and strengthen their AI content generation practices. Ensuring that speed and scalability don't come at the cost of validity or control.

26 Comprehensive Literature Reviews of Data Forensics Methods

Time:	4:35 PM - 5:35 PM
Location:	Room 165-69
Speakers:	Daihui Xiao, Kylie Gorney, & Kevin O'Rourke

A Systematic Review of Machine Learning-Based Approaches for Detecting Test Fraud
(Daihui Xiao & Kylie Gorney)
Test fraud presents an ongoing threat to test fairness and the validity of test score interpretations. In response, researchers have increasingly explored machine learning (ML) as an innovative tool for identifying and detecting patterns associated with test fraud. This systematic review synthesizes studies from peer-reviewed journals, research reports, dissertations and theses, and conference proceedings that apply ML techniques—including supervised, unsupervised, and ensemble methods—to detect test fraud. The methods in these studies have been used to detect suspicious examinees, items, or both. By identifying strengths, limitations, and research gaps, we hope to provide an overview of the methodological landscape and help inform the development of data-driven approaches to maintaining test security.

100 Years of Detection Methods: A Quick Historical Tour of Copying Indices, Similarity Measures, and Other Test Security Detection Methods
(Kevin O'Rourke)

From the late 1920's through the present day, detection of test security violations due to preknowledge/etc has taken many forms, and has had many contributors. Presented here is a lit review and historical summary which attempts to trace, condense, and contextualize the nearly 100 years of methods and contributors framed by the conditions and technology that has evolved alongside the methods.

27 Panel of Nationally-recognized Experts Address Critical Issues on Recent Changes to Testing Policies and Security for National and State Assessments

Time:	4:35 PM - 5:35 PM
Location:	Room 162-75
Speakers:	John Olson, Walt Drane, Michol Stapel, Maria D'Brot & Juan D'Brot)

This session brings together the perspectives of a nationally-recognized group of testing experts to discuss the latest strategies, policies, and practices for national and state assessments. The panel will address critical issues related to recent changes and challenges to K-12 testing in the U.S., along with actionable strategies for strengthening the integrity of assessments while balancing the need for flexibility in innovative assessment approaches. These critical issues are more important than ever given the ever-changing political landscape the recent presidential election has brought to national/state assessment programs this year, and the major revisions being made to the U.S. Department of Education.

Panelists will share insights into specific areas of test integrity, test design, and testing purposes, focusing on emerging issues and major changes being made to national/state assessment programs. The panel will provide a roadmap for states seeking to implement flexible yet secure testing protocols, enabling them to meet today's challenges while preparing for those coming up on the horizon. Participants will learn how to balance rigorous procedures with flexibility needed for innovation and ensuring each part of an assessment system is implemented securely according to its stakes.

Poster Reception & Cocktails

Time:	5:50 PM - 7:00 PM
Location:	Marriott Center, Campus Center 11th Floor

1. Detecting Fraud Rings in Higher Education Applications Using AI, NLP, and Graph Analytics (Emrah Anayurt & Rishi Bhartiya)
2. Identifying Aberrant Behavior in CAT: A Hybrid Clustering (DBSCAN) and XGBoost Classifier Approach (Murat Kasli & Joe Betts)
3. Response Revision Analysis for Test Security: A Zero-Inflated Negative Binomial Approach (Zhuoran Wang, William Muntean & Shonai Someshwar)
4. A Unified Latent Space Item Response Model for Test Security Diagnostics (William Muntean, Zhuoran Wang & Joe Betts)
5. Detecting and Visualizing Aberrant Test-Taking Behaviors in Sparse Data via Joint Response Time/Accuracy Modeling (Jake Cho)
6. Use Hidden Message on DET Certificate to Identify Cheaters (Larry Chen)
7. How Do We Build User-Friendly Room Scans in the Duolingo English Test? (Yi He)
8. A Case Study in Braindump Evolution and Proliferation (Marcus Scott)
9. Visualizing Path Usage in Multistage Testing using Alluvial Diagrams (Fernando Mena Serrano & April Zenisky)
10. Localized Fade-Away Method for Controlling Item Exposure across Performance Levels (Ye Yuan & Kyung (Chris) T. Han)
11. From Time Zones to Time Series: Combining Classical and Learning-Based Methods to Spot Anomalies in Real-World Testing Data (Mengyau Zhang)
12. Securing Item Banks with AI-Driven Assessment Engineering (Anna Nasyrova)

(Please see Poster Abstracts on page 53)

Wednesday, November 5

28 Creating a Holistic Exam Security Structure for Your Exam Suite

Time:	8:30 AM - 9:30 AM
Location:	Room 163
Speakers:	Rachel Schoenig, Mike Clifton, Cathy Koenig & Ada Woo

When it comes to exam security, one size does not fit all. Programs with two or more exam offerings are often faced with the challenge of juggling the exam security rules, roles, and responses to exam security incidents in ways that make sense for each program and stretch already limited exam security resources. This can leave internal stakeholders frustrated and external stakeholders confused. This session will discuss methods for evaluating exam security needs by program and resources available to help guide programs in more effectively creating holistic exam security structures that address your suite of exams. Together, we will discuss the pitfalls that can arise from a one-size fits all or a "drive through menu" approach to test security. With experienced practitioners across the testing spectrum, from higher ed/admissions to licensing and certification, we will share lessons learned and effective strategies to create a holistic structure that is manageable across a suite of exams.

29 Responding to Today's Test Security Landscape

Time:	8:30 AM - 9:30 AM
Location:	Room 168-74
Speakers:	John Dight, Jarret Dyer, & Wally Dalrymple

Test Security 360: Peel, Protect, Prevent - The Onion Approach to Test Security (John Dight & Jarret Dyer)
Test security is often tackled in silos - focusing separately on cheating prevention, data integrity, cybersecurity, or physical security. But what if we viewed it holistically, as a unified, interconnected ecosystem? This engaging session explores the full spectrum of test security, revealing how these seemingly separate elements work together to form stronger, layered defenses. Through interactive discussions, storytelling, and a touch of humor, attendees will learn about the critical layers of prevention, detection, and response, along with their

relevance, implementation effort, and associated costs. Whether you're new to test security or a seasoned professional, this session will offer valuable insights and perhaps uncover gaps or opportunities you haven't yet considered.

Evolving defense strategies in the face of deepfakes, synthetic IDs, and the AI arms race
(Wally Dalrymple)
The scale and sophistication of AI-assisted test fraud is accelerating, with deepfakes, digital document forgeries, and synthetic identities now occurring at unprecedented rates. The 2025 Identify Fraud Report from the Entrust Cybersecurity Institute reveals that in 2024 deepfake attacks accounted for 40% of biometric fraud, and digital forgeries surged 244% year-on-year. These trends highlight the urgent need for a new defense model.

This session, led by Wallace Dalrymple (Joint Chief Security Officer at PSI and ETS), explores the evolving AI threat landscape through a practitioner lens. The session will include examples of deepfake detection in action and provide applied insights into advanced biometric defenses. This includes multi-layer ID verification (face, voice, keystroke), liveness checks, and behavioral forensics.

Participants will gain an inside look at how layered AI security architectures can be deployed to enhance test security, by combining real-time anomaly detection, backend pattern analysis, and human-in-the-loop oversight. With both defensive and offensive tactics advancing rapidly, this session offers a grounded, strategic exploration of how to outpace fraud without compromising fairness or accessibility.

30 Operational Test Security Issues

Time:	8:30 AM - 9:30 AM
Location:	Room 165-69
Speakers:	Kimberly Snyder, Steven Svoboda, Greg Hurtz, Nicole Tucker, & Larry Chen

Small Budget, Big Impact: Low-Cost Cheating Ring Investigations
(Kimberly Snyder)

As assessments have scaled globally, organized cheating has evolved from isolated incidents to complex, coordinated operations, often operating much like criminal enterprises. These fraud networks may now mirror the structure and sophistication of legitimate organizations—complete with marketing, pricing tiers, client onboarding, and technical support.

As a result, today's security approaches may involve ongoing OSINT operations, data-sharing networks, and close collaboration between product teams, developers, assessment scientists, legal advisors, and test center operations as bad actors operate at a larger and more complex scale while the rapid advancement of technology makes keeping up with these tactics evermore challenging. Some larger test security companies may even have the legal resources and global presence to involve local police and take legal action against far-reaching collusion rings.

But not every organization has resources to invest in enterprise security platforms or dedicated fraud teams. This proposal is aimed at showing how innovation and resourcefulness, not just budget, can drive results.
The session will empower smaller test providers and academic institutions with limited teams, legal and/or technological resources to monitor, investigate, and deter organized cheating. It promotes a mindset of frugal innovation and grassroots intelligence that can significantly improve security outcomes regardless of scale.

The Need for Speed: Demonstration of a Data Forensics Tool Getting Close to "Real Time"
(Steven Svoboda, Greg Hurtz, & Nicole Tucker)
The holy grail of psychometric data forensics is a system that flags anomalous behavior in "real time" during the administration of a test. As an initial step towards finding that holy grail, we have developed methodologies enabling near real time, "next day" forensics analyses. These methods allow clients to act quickly and enhance the integrity of their testing programs. We believe that a flagging system based on next day forensic analytics has the potential to reshape the landscape of test security.

In this presentation we will describe a Power BI dashboard we have created to identify anomalous behaviors the day after testing. We will present forensic indices and techniques that can be implemented for detecting anomalous behavior,

followed by a demonstration of our “next day” dashboard based on real world datasets. Finally, we will discuss our planned enhancements aimed at enabling real time forensic capabilities.

Test Security Data Coverage and Usage
(Larry Chen)
This presentation highlights the data coverage that powers test security for the Duolingo English Test (DET). Across every stage of the test lifecycle, we collect highly detailed data to monitor behavior, detect anomalies, and protect the integrity of our assessment.

Our data is organized into distinct phases aligned with the test journey: onboarding, in-test activity, offboarding, post-data processing, proctoring review, post investigations and decertifications. During onboarding, we log setup conditions and verify the user environment. In-test activity includes real-time tracking, system-level monitoring, and user interaction data. Offboarding captures session closeout data and feedback. After the test, we process scoring, grading data, and video recordings with secure test-taking. Proctoring data from AI and human reviewers further supports integrity decisions. Finally, we have offline investigations to drive enforcement actions like decertification.

To make this data actionable, the DET leverages a suite of internal tools and platforms. Arize is used to monitor AI models responsible for fraud detection and signal generation. BigQuery serves as our central warehouse for all structured test security data. Sigma Computing and Metrics dashboard enable no-code dashboarding and flexible metric exploration for managers and engineers. Our internal anomaly detection system continuously scans for unusual behavior patterns and infrastructure misuse.

31 Terrifying Testing Tales: Scary Stories from the Front Lines of Test Security for State Assessment Programs

Time:	8:30 AM - 9:30 AM
Location:	Room 162-75
Speakers:	John Olson, Walt Drane, Joe Blessing, David Ragsdale, Jessica Fenby, C. Alan Burrow, Mark Jackson & Timothy Butcher

Just days after Halloween, join a frightfully experienced panel of state test security specialists as they share true, spine-tingling stories of testing misconduct that will leave you wide-eyed and wary. From the illicit use of AI familiars, phantom phones, duplicate devices, and hidden cameras, to students using technology for nefarious purposes, this session exposes the dark arts of modern cheating—and how states are fighting back. With over 60 combined years of experience, state DOE panelists from GA, MA, MI, MS, TN, and WV will lift the coffin lid on emerging risks like Generative AI and social media mischief, best practices and protective spells (i.e., policies, tools, strategies) for securing assessments, and real-world case studies of students who tried to outwit the system—and got caught. This lively fireside chat, moderated by two national assessment security sleuths, promises candid insights, interactive discussion, and maybe even a few jump scares. Panelists will appear in Halloween costumes, and attendees are welcome to do so too. Yes, there will be photo ops and time for your burning questions. Enter if you dare!

32 Creating Effective Test Security Governance Structures

Time:	9:40 AM - 10:40 AM
Location:	Room 163
Speakers:	Rachel Schoenig, Camille Thompson, Ray Nicosia, Faisel Alam & Jake Ritz

Managing exam security across a testing program, board, internal stakeholders, and external stakeholders doesn't occur by magic. The most effective management occurs with clear governance structures and policies that guide the program in making changes to exam security and when responding to incidents. This session will address the fundamentals of creating effective test security governance structures that can help testing programs stay abreast of evolving threats and respond in a timely and effective manner to security incidents. Attendees will explore the various stakeholders to consider within their structure, key communication needs, and policy development and incident response structures that can help make managing test security more effective for the organization. Join security experts for a fun, gamified session that will leave attendees feeling more confident about next steps to advance their exam security framework.

33 Exam Security Investigations – Recommended Best Practices In Light of Recent Threats

Time:	9:40 AM - 10:40 AM
Location:	Room 168-74
Speakers:	Brent Morris, Harry Samit & Bryan Friess

The last five years have seen many changes in the certification industry. The expansion of online testing, the emerging yet increasingly crucial role of artificial intelligence (AI), and the use of ever more creative item types are just a few of these.

Unfortunately, while industry leaders have pushed the boundaries of what is possible, so too have organized criminal theft rings bent on stealing and selling valuable intellectual property or orchestrating ever-more sophisticated proxy tests.

These threats call for a robust exam security framework to protect the integrity of exam programs. In addition to vigilant and innovative security measures, securing an exam program also requires having a strong investigative capability able to flag misconduct and then determine exactly what happened. The evidence and intelligence resulting from investigations must flow to stakeholders to allow exam sponsors to make certification decisions.

New innovated techniques: AI Photo Comparisons, Machine ID and Network Name matches, Candidate interviews, Human Source Operations, Undercover Operations, Law Enforcement Referrals

34 Innovating Fairly: How to Experiment on Remote Proctoring Without Risking Test Equity

Time:	9:40 AM - 10:40 AM
Location:	Room 165-69
Speakers:	Will Belzak, Angel Ortmann Lee, Brooke Westerlund, & Yong-Siang Shih

Experimentation is fundamental to understanding how changes affect outcomes. Yet in high-stakes testing environments, it is rarely used due to fairness concerns: any variation in proctoring procedures can risk invalidating scores and undermine equity for test takers. This presentation introduces a novel approach that enables experimental evaluation of proctoring practices using previously completed

(non-live) test sessions. By simulating the live proctoring experience and randomly assigning proctors to different instructional conditions, this method allows researchers and test security teams to rigorously assess the impact of changes without introducing risk or bias into operational assessments.

We will share results from experiments that examined modifications to proctoring speed, behavioral flagging criteria, and other decision processes. In one study, increasing video playback speed during session reviews reduced proctoring time by 10% with no loss in decision accuracy. In another, new guidance on flagging visual behaviors inadvertently reduced overall accuracy, highlighting the importance of empirical validation before deployment. Building on this foundation, we propose a multi-phase framework modeled after clinical trials to extend this methodology to live testing environments. This structured approach offers a scalable path for iterating on remote proctoring policies while safeguarding fairness, transparency, and the validity of test outcomes.

35 New AI-Driven Threats to Test Security: Understanding Risks and Building Defenses

Time:	9:40 AM - 10:40 AM
Location:	Room 162-75
Speakers:	Sergio Araneda, Sarah Toton, Christopher Foster, Andrew Marder & David Foster

As large language models (LLMs) become increasingly integrated into education, they introduce powerful capabilities, but also serious challenges for test security. This session presents three empirical studies that explore distinct dimensions of AI-related threats to assessment integrity. The first study investigates the behavior of LLMs when answering multiple-choice questions, focusing specifically on the patterns, accuracy, and consistency of their responses. By analyzing how models respond across content domains and temperature settings, researchers examine whether such behavior can be reliably detected through data forensics. The second study addresses the concept of “item pre-exposure,” a newly emerging risk in which test-takers may anticipate or recreate AI-generated items prior to their use. Through repeated item generation and human replication tasks, the study quantifies overlap between publicly generated content and test items, shedding light on vulnerabilities in current item development practices. The final study introduces Observer, a multimodal, AI-enhanced proctoring system designed for

detecting GPT-assisted cheating. Focusing on its underlying design and phase-one results, the study explores how risk-based scoring and explainable models can enable more efficient and defensible proctoring. Collectively, these studies offer practical, data-driven strategies for adapting assessment systems to the challenges of AI, while upholding fairness, security, and validity.

36 There is more than one way to skin a CAT

Time:	10:50 AM - 11:50 AM
Location:	Room 168-74
Speakers:	Stephen G. Sireci, April L. Zenisky, Peter Tran, Ketan, Jennifer Lewis, David Foster, & Sergio Areneda

Computerized adaptive testing (CAT) offers two major benefits: efficiency and security. By adjusting item difficulty in real time, CATs reduce test length by a surprising amount without compromising score reliability. This reduces both time spent and test-related stress, along with enhancing test security by generating unique test forms for each test taker from large item pools.

Despite these advantages, CAT adoption remains low. Traditional CAT implementations rely on item response theory (IRT), large manually built item pools, resource-intensive field testing, and complex statistical modeling - barriers that can deter smaller or resource-limited programs. Moreover, many vendor-based test delivery systems are not technically equipped to handle the dynamic nature and unique requirements of CATs.

Together, these four presentations aim to demystify CAT implementation, highlight its effectiveness, address specific issues with it, and explore innovative, more accessible approaches that could broaden its adoption across a wider range of testing programs.

37 Data Forensics – Stories from the Field!

Time:	10:50 AM - 11:50 AM
Location:	Room 162-75
Speakers:	Jarret Dyer, Sarah Toton, Regi Mucino, Anna Rubin, Jim Hussey & Steve Addicott

In today's high-stakes testing environment, maintaining the integrity of assessments is more critical than ever. This engaging panel discussion brings together leading experts in educational measurement, psychometrics, data science, and academic integrity to explore how data forensics is revolutionizing the way we detect and deter cheating in testing environments. From identifying suspicious response patterns to analyzing statistical anomalies, data forensics

provides powerful, evidence-based tools to uncover misconduct that traditional methods often miss.

Attendees will gain insights into real-world case studies, learn about cutting-edge forensic techniques, and discover how to implement these strategies ethically and effectively. Whether you're a test developer, administrator, educator, or policy maker, this session will equip you with the knowledge to protect the validity of your assessments and uphold fairness for all test-takers.

38 Technical Issues in Test Security

Time:	10:50 AM - 11:50 AM
Location:	Room 165-69
Speakers:	Sarah Quesen, Carol Eckerly, Kylie Gorney, & Kevin O'Rourke

When Too Many Flags Fly: Controlling False Positives in Classroom Test Forensics
(Sarah Quesen)
This study addresses a persistent challenge in statistical test security: controlling Type I error rates when screening multiple classrooms for potential testing irregularities. As the number of classrooms examined increases, the risk of false positives grows, potentially leading to unwarranted investigations. This presentation introduces an approach that combines bootstrap resampling with False Discovery Rate (FDR) control to reduce false accusations while preserving the ability to detect true instances of misconduct. Simulation studies across varying rates of actual cheating and classroom sizes demonstrate that the method significantly lowers the number of false flags compared to traditional threshold-based approaches, without sacrificing statistical power. Practical implementation guidance will be provided for psychometricians and test security professionals seeking to integrate these methods into operational protocols.

Efficient Answer Similarity Analysis using the M4 Statistic
(Carol Eckerly & Kylie Gorney)
The combination of an answer similarity index and cluster analysis is commonly used to detect groups of examinees with unusually similar responses. Groups of examinees with usually similar responses may occur due to common test security

threats such as item preknowledge (i.e., inappropriate prior access to live exam content) or collusion (i.e. working together to answer items). Answer similarity analysis is conducted at the examinee pair level, and the resulting pairwise statistics are used to group examinees into clusters.

Because the relationship between the number of examinees and the number of examinee pairs is quadratic, the computation time for conducting similarity analysis among all pairs of examinees may be prohibitive for large testing programs. This presentation introduces an efficient computational method for calculating the M4 similarity statistic among all pairs of examinees using lookup tables and matrix multiplication. The method utilizes the Rasch model with noninformative distractors as the IRT model underlying the M4 calculations. The method is evaluated with a simulation study.

Demonstrating the Impact of Compromised Items and Examinees with Preknowledge on Parameter Recovery and Measurement Precision
(Kevin O'Rourke)
Few studies have quantified how test security breaches—specifically compromised items (CIs) and examinees with preknowledge (EWPs)—distort psychometric outcomes. This simulation study evaluated the impact of CIs and EWPs on item parameter recovery and measurement precision. Using 3PL-generated data for 5000 examinees under five CI/EWP conditions, both 1PL and 2PL models were fit across 100 replications. Compromised items showed substantially greater bias and RMSE in difficulty (b) and discrimination (a) estimates, with distortions most pronounced under the 2PL model. In the most severe conditions, parameter errors more than doubled, and item information decreased by up to 40% in key theta ranges (–1 to 1). These findings demonstrate that test security violations systematically degrade the accuracy and utility of test scores. Beyond fairness concerns, item compromise undermines the psychometric foundations of high-stakes assessments and should be treated as a measurement threat requiring proactive detection and response.

39	**Crisis to Confidence: Rebuilding an Exposed Exam with AI Support**
Time:	10:50 AM - 11:50 AM
Location:	Room 163
Speakers:	Diane Long & Chris Foster

In this session, we'll simulate a scenario every testing program dreads: the exam is exposed.

Now what? Using AI exam development tools, our panel of test development and machine learning specialists will walk through how an exposed or compromised test form can be rapidly updated—new items generated, flagged, reviewed, and slotted into a refreshed blueprint—all within the space of an hour.

This session showcases how AI can help teams respond quickly to security breaches, while still maintaining quality, fairness, and defensibility in the revised exam content. It's a real-world use case, with real-time pressure—and a human-led process every step of the way.

Closing Keynote	**COTS Debates**
Time:	12:00 PM - 1:00 PM
Location:	Campus Center Auditorium
Debaters::	Kim Brunnert, Carol Eckerly, Sanjoe Jose, & Alana Chamoun
Moderator:	Rachel Schoenig

Join us for the closing COTS Debates! This always informative and definitely entertaining keynote session will feature industry luminaries as they debate topics of interest to testing professionals.

Poster Session Abstracts

Detecting Fraud Rings in Higher Education Applications Using AI, NLP, and Graph Analytics *(Emrah Anayurt & Rishi Bhartiya)*
Fraudulent applications pose a growing challenge to admissions integrity, often distorting metrics, wasting institutional resources, and undermining public trust. At Southern New Hampshire University, we developed an AI-powered fraud detection framework to proactively identify and mitigate such threats. By combining supervised machine learning, text mining, and network analysis, we flagged suspicious applications using features like IP address, email clustering, and behavioral patterns.

Our poster will present a real-world implementation of this system, focusing on how we generated fraud scores that are explainable and actionable for operational teams. We will share how lead-source anomalies, IP and email similarity metrics, and graph-connected components enabled us to detect fraud rings—including one with over 160 linked applications.

We also highlight how model explainability helped establish stakeholder trust and guided follow-up actions, while Spark-based infrastructure ensured scalable performance. Attendees will walk away with a practical example of how AI and network analysis can enhance test security, data integrity, and lead validation at scale.

Identifying Aberrant Behavior in CAT: A Hybrid Clustering (DBSCAN) and XGBoost Classifier Approach *(Murat Kasli & Joe Betts)*
Maintaining test security in computerized adaptive testing (CAT) environments presents unique challenges due to the personalized nature of item administration. This study proposes a novel hybrid approach combining density-based clustering (DBSCAN) with gradient boosting classification (XGBoost) to detect abnormal patterns without requiring prior knowledge of flagged items or examinees. Using a comprehensive dataset of over 10,000 examinees, including flagged cases of aberrant behavior—our methodology leverages examinee-level behavioral patterns rather than item-specific responses.

The first phase employs DBSCAN to identify natural clusters of examinees with similar behavior patterns without pre-specifying cluster numbers. These clustering results

then enhance feature sets for XGBoost classification, creating a powerful two-stage detection framework.

This hybrid approach is expected to identify both known aberrant patterns and potentially undiscovered behaviors. By focusing on behavioral metrics rather than specific item responses, this research addresses the fundamental challenge of security monitoring in adaptive testing environments where traditional fixed-form detection methods prove inadequate.

Response Revision Analysis for Test Security: A Zero-Inflated Negative Binomial Approach *(Zhuoran Wang, William Muntean & Shonai Someshwar)*
Item compromise poses a significant threat to the validity of high-stakes assessments. While many detection methods analyze final response patterns, this study introduces an innovative approach leveraging response process data -- specifically, response revisions captured via user interaction logs -- to identify potentially compromised items by detecting unusually low levels of examinee interaction. We hypothesized that examinees with pre-knowledge engage less in cognitive deliberation leading to revisions. An empirical study on a high-stakes licensure exam employed a zero-inflated negative binomial model, conceptualized as a novel Item Response Model, to predict revision counts based on intrinsic item characteristics (e.g., difficulty, type, word count). Results indicated that item characteristics significantly predicted revision frequency. Items that exhibited a statistically significant reduction in revisions than those predicted by the model, revealing under-interaction, were flagged as potentially compromised. This methodology offers a distinct, complementary tool for test security forensics, capable of identifying pre-knowledge exploitation that may evade traditional detection methods, thereby enhancing assessment integrity and fairness.

A Unified Latent Space Item Response Model for Test Security Diagnostics *(William Muntean, Zhuoran Wang & Joe Betts)*
Addressing the critical need for advanced test security, this research introduces a novel latent space item response model. Its unique contribution lies in unifying latent spaces from multi-modal, multi-variate data—item responses, response times, and computer-based exam interaction data—into a single, cohesive framework. This integrated approach is specifically designed to uncover hidden person-by-item interactions. Such dependencies, often undetectable by existing methods that analyze response data in isolation, serve as key indicators of item compromise, particularly item leakage and examinee pre-knowledge. The efficacy of this unified model, alongside its constituent submodels, is systematically evaluated against Yen's

Q3 statistic for identifying item dependencies within item sets. By providing a more nuanced understanding of complex response patterns, this research aims to enhance the detection of compromised items, thereby offering a more robust defense against threats to test integrity. Beyond detection, the proposed methodology demonstrates strong utility as a diagnostic tool for item inspection, aiding in the detailed analysis and strategic management of item security.

Detecting and Visualizing Aberrant Test-Taking Behaviors in Sparse Data via Joint Response Time/Accuracy Modeling *(Jake Cho)*
Linear-on-the-Fly Testing (LOFT) administrations often result in sparse response matrices, challenging conventional methods for detecting aberrant test-taking behaviors like item pre-knowledge. This study presents an approach using the LNIRT R package (Fox et al, 2023) to jointly model response accuracy and response times. This joint modeling offers several advantages, including a more holistic view of the test-taking process, potentially more accurate estimation of person ability and speed by leveraging information from both response types, and a better understanding of the speed-accuracy relationship, while also leveraging its strengths in handling data sparseness. We identify potentially aberrant examinees by combining statistical flags: a chi-squared test on standardized person ability and speed estimates (targeting high-ability, fast-response patterns indicative of pre-knowledge) and EAPCP values for response accuracy, response time, or joint aberrance. Suspects are further filtered by performance exceeding a defined cut-score. The core contribution is a comprehensive visualization method, generating individual profiles that intuitively display item-level response accuracy, personal vs. mean response times, and item p-values, differentiated by item type. This methodology provides an effective tool for enhancing test security by offering actionable, visual insights into suspect test-taking patterns in sparse data environments.

Use Hidden Message on DET Certificate to Identify Cheaters *(Larry Chen)*
This poster explores ways to catch test-takers who cheat on the Duolingo English Test (DET) by hiding secret messages in the test certificates.

Cheating agents often advertise certificates from real test-takers to attract more customers. To make the certificates look clean and legitimate, they usually hide important information while keeping the overall score visible. This helps them promote their services without revealing clues that the DET can use to trace them.

To fight back, DET certificates can include hidden patterns and messages—both visible and invisible—that link each certificate back to the original user. These security

features help us detect tampering, identify cheaters, and protect the integrity of the test.

How Do We Build User-Friendly Room Scans in the Duolingo English Test? *(Yi He)*
This poster introduces the development of a user-friendly and fraud-resistant 360° video room scan for the Duolingo English Test (DET). The goal is to detect device-assisted cheating while maintaining a comfortable experience for honest test-takers.

Cheating rings increasingly exploit low-tech methods, using hidden devices or manipulating their environment during testing. DET is introducing a video-based room scan during onboarding. This scan captures a panoramic view of the test-taker's environment, using mobile phone sensors and cameras.

A Case Study in Braindump Evolution and Proliferation *(Marcus Scott)*
This session presents a case study about how a braindump evolved over time as items were retired and created and how its use proliferated among the testing population. A testing program used three different test forms for an exam over an 18-month period, and each successive test form used some items from the previous form. At some point, the testing program discovered a braindump containing every item from the first form along with a very accurate suggested answer key. A large cluster of identical tests was detected for the first form. These tests had the exact same set of responses as the braindump answer key, which provided strong evidence of its use. Analysis of identical clusters detected for later test forms suggested that the braindump was being updated with new items. Data forensics analysis provided additional information about how long it took for the new items to be disclosed and added to the braindump, as well as how many examinees potentially used it.

Visualizing Path Usage in Multistage Testing using Alluvial Diagrams *(Fernando Mena Serrano & April Zenisky)*
In multistage testing (MST), examinees follow adaptive routing paths that can be analyzed to evaluate test design and enhance test security. Examining path frequencies helps identify unusual routing patterns, including low-frequency sequences that may signal aberrant responses or risks of overexposure. While tabular summaries offer useful information, visualizations can reveal patterns more intuitively. This study applies alluvial diagrams (Rosvall & Bergstrom, 2010) to depict routing in a six-stage MST. Examinees began in one of five difficulty levels and were routed by IRT-based proficiency estimates. The diagrams show colors for starting levels, flow widths for path frequencies, and columns for module exposure. This visualization

provides a clear, practical tool for monitoring MST routing and detecting anomalies and could strengthen both test validity and security practices.

Localized Fade-Away Method for Controlling Item Exposure across Performance Levels *(Ye Yuan & Kyung [Chris] T. Han)*
In computerized adaptive testing (CAT), items in an item bank are evaluated based on item selection criteria and selected given test taker's ability level. In this context, overexposure of items compromises security (e.g., item harvesting), while underutilization squanders item-development resources. For these reasons, controlling item exposure is important to maintain test security and ensure efficient use of the item bank. However, many existing exposure control methods are not designed to address conditional item exposure rate, which refers to the proportion of examinees seeing an item given to all the examinees at the ability level for a particular range. The purpose of this study is to introduce a new variation of the Fade-away (FA; Han, 2012) method to control conditional item exposure rate in CAT. The study presents FA and the adaptation of the FA, Local Fade-Away (LFA) methods, and compares existing exposure control techniques like Randomesque and Sympson & Hetter methods to evaluate their effectiveness. Results from the simulation study demonstrated the LFA method's capability in maintaining controlled and balanced exposure rates at each ability level, effectively reducing both item overexposure and underexposure.

From Time Zones to Time Series: Combining Classical and Learning-Based Methods to Spot Anomalies in Real-World Testing Data *(Mengyau Zhang)*
This study revisits approaches to detecting time zone cheating in testing data by expanding the focus to a broader time series framework. A combination of classical statistical methods and modern learning-based techniques is explored to uncover irregularities across time zones and testing sessions. Data from a large-scale licensure exam is used to assess the effectiveness and limitations of these methods, as well as the consistency of their detection results. Practical implications of the findings and challenges encountered in this context are also discussed.

Securing Item Banks with AI-Driven Assessment Engineering *(Anna Nasyrova)*
The presence of compromised items undermines the integrity of scores and the validity of score interpretations. In the context of high-stakes testing, the consequences can negatively impact individuals and erase the program's credibility. Although this issue can be mitigated by creating extensive item banks, the development and maintenance of such banks introduce additional costs associated with producing items at scale. With a growing need for high-quality items, the

demand for automating this process has intensified. Today, rapid advances in generative AI present novel opportunities for quicker and potentially more efficient production of nearly inexhaustible item banks.

This poster proposes an operational framework for AI-assisted automated item production in adult reading assessments. First, we describe the process of item design, which involves creating item templates tailored to different levels and standards. Second, we outline the item production process in accordance with test specifications. Third, we provide suggestions for reviewing and evaluating item quality. We conclude with practical recommendations for future iterations of the proposed workflow.

Virtual COTS Program

Wednesday, November 12

All sessions are in Eastern Standard Time.

Welcome & Opening Keynote **A Case Study in Industry Strength: Tackling Commercial Cheating**	
Time:	11:00 AM - 12:00 PM
Speakers:	Noah Reandeau, Rachel Schoenig, & Jake Ritz
Moderator:	Mike Clifton
Commercial cheating services are targeting learners and test takers for their services, preying upon an individual's anxieties to gain a profit. Some of these commercial cheating services resort to ongoing cheating recruitment and extortion to expand their profits, all while undermining academic and testing integrity. This session will explore types of commercial cheating services, the risks to learners, test takers, institutions and testing programs, and how the industry has come together to tackle the issue. Together, we will explore the path taken to encourage lawmakers to take action against commercial cheating. Join a seasoned lobbyist and board member to share how, working together, the industry is changing the legal landscape and protecting learners, test takers, and testing programs. Attendees will learn more about what they can do to join this effort as well as how this case study in strength can be applied to other potential issues facing the industry.	

V1	**Detecting in the Now: Real-Time Anomaly Detection using Multimodal Inputs**
Time:	12:05 PM - 1:05 PM
Speakers:	Sarah Toton & Andrew Marder
Real-time detection allows flagging and intervention as anomalous behavior unfolds, rather than after the damage is done. This session is about designing and implementing a comprehensive real-time detection system, incorporating everything from webcam and screen feeds to keystroke logs, mouse movement patterns, log files, and real-time data forensics. Ultimately, this allows anomalous behavior or patterns exhibited by a test-taker to be assessed dynamically, moment by moment, as they take the test.	

A common misconception is that real-time data forensics indicators are simply post-hoc data forensics applied faster; but real-time data forensics are fundamentally different. Real-time data forensics relies on more immediate, but weaker information and patterns. It does not have the same robustness and contextual interpretation as post-hoc data forensics and demands a higher tolerance for uncertainty. However, when real-time data forensics indicators are combined with other streams of information, they can be powerful—triggering live actions such as shifts in content, proctor intervention, or session termination, rather than retrospective actions like score invalidation or investigation.

We'll discuss the trade-offs, challenges, and opportunities of real-time detection, as well as ethical considerations and practical constraints.

V2	**Practical Applications of Generative Artificial Intelligence in K-12 State Assessment Programs: Latest Updates on States' Use of AI in 2025**
Time:	12:05 PM - 1:05 PM
Speakers:	John Olson, Brian Reiter, Sarah Quesen, & Walt Drane

In 2024-25, use of Generative Artificial Intelligence (Gen AI) for many purposes related to testing continued to be implemented successfully. Gen AI has now been demonstrated by several testing companies and others to be viable for a variety of programs, including state assessments. However, there are also ongoing concerns about AI's use related to ethical principles, data management, and test security. These new AI approaches are transforming the testing industry and impacting state programs in areas like item development, forms assembly, and pilot testing. States are also busy sharing information with each other on what works/doesn't work with Gen AI and its ethical use. This collaborative-sharing is resulting in best practices for AI that leverage important technological innovations to help assessment programs improve, while reducing costs. In this session, an experienced state assessment director leading a cross-state effort on AI, a prominent researcher on AI applications for educational testing, and two national assessment/security experts will present information/recommendations on how Gen AI can greatly help state efforts. Attendees will increase their AI literacy by gaining valuable insights into practical implementations of AI, latest lessons learned from real-world applications, and recommendations on secure implementation of Generative AI for state assessments.

V3	**Advancing Test Security in the Age of Artificial Intelligence: An Integrated Framework for Research, Education, and Practice**
Time:	2:10 PM - 3:10 PM
Speakers:	Taiwo Feyijimi, Daniel Oyeniran, Olukayode Apata, Mubarak Mojoyinola, Ernest Amoateng, Justice Dadzie, John Ajamobe, Henry Makinde, & Ibukun Osunbunmi

Maintaining the integrity of assessments is crucial across educational and professional sectors, but sophisticated test fraud, significantly amplified by artificial intelligence (AI), poses unprecedented challenges to traditional security paradigms. This manuscript examines the evolving landscape of test security threats and countermeasures, encompassing statistical detection methods, AI-driven tools, content protection strategies, and ethical considerations. A central theme is the dynamic relationship where AI acts as both a significant threat and a powerful defensive tool. Furthermore, sophisticated security measures introduce tension with the test-taker experience, raising concerns about privacy, fairness, and anxiety. This paper argues for a paradigm shift towards a holistic and adaptive approach. This study proposes the Secure-Ethical-Educational (SEE) framework, an integrated, multi-layered model designed to advance test security. The SEE framework integrates advanced threat detection and analytics, robust prevention and deterrence strategies, ethical governance with human oversight, and a foundational culture of integrity coupled with continuous learning. This framework is novel due to its explicit embedding of educational dimensions, including test security education for professionals, pedagogical strategies for fostering academic integrity, and a research-informed adaptive cycle, as integral components of a comprehensive security posture. The implications of SEE framework for enhancing research, professional education, and teaching practices are discussed.

V4	**Proxy Testing – The Whys and Wherefores of a Proper Investigation**
Time:	2:10 PM - 3:10 PM
Speakers:	Mike Clifton, Joe White, & Jeff Marsh

Proxy testing is a sophisticated fraud that requires deliberate methods to protect your program while avoiding unpleasant repercussions from dishonest testers. In this interactive case study, attendees will hear about current trends in proxy testing as the backdrop for exploring best practices involved in identifying proxy-testing violations, gathering evidence, drawing appropriate conclusions, and taking actions

against the testers that are most-defensible and easily applied to other types of testing violations.

V5	Novel Psychometric Approaches to Detecting Test Fraud
Time:	3:15 PM - 4:15 PM
Speakers:	Regi Mucino, Greg Hurtz, Edmund Jones, Chris Bell, Tom Benton, Stephen Cromie, William Muntean, Zhuoran Wang, & Shonai Someshwar

Item-Focused Forensics: Searching for the Items that Might be Compromised (Regi Mucino & Greg Hurtz)
Data Forensics has spent much focus on identifying individuals or groups with statistically aberrant patterns of test taking to ensure the validity of an examination. Another important route to explore is the items. In this presentation, we explore methods for identifying items that may have been compromised. We do this by describing various methodological approaches to investigating changes in item performance and response times, such as changes over time, disparate patterns of item results among candidates suspected of aberrant testing behavior, and systematically higher rates of item response time model residuals. Each of these methods is also accompanied by real-world examples.

Using Response-Times in a Test Where Candidates Can Go Back and Redo Items (Edmund Jones, Chris Bell, Tom Benton & Stephen Cromie)
Response-times can be used to detect aberrant behaviour in computer-based tests. Most research on this topic has been for adaptive tests or other tests where candidates have to answer each item in order. However, some computer-based tests allow candidates to skip items, do them in any order, and go back and change their responses.

This presentation will be about a high-stakes computer-based test of English reading proficiency that works in this way. The test has eight parts; most parts contain one passage of text and several items. We developed a method for screening candidates for possible cheating, using their response-times for their first attempts at each part.

We use a separate log-normal distribution for each part. To avoid flagging candidates who merely did not make a serious effort in their first attempt at a part, we only include those who would have scored at least 80% on it (based on their first

attempt). We also use a non-standard method to fit the model, using the quartiles of the normal distribution, so that the log-normal model fits the non-aberrant response-times even better. Candidates are flagged if they have response-times that are very short and unlikely under the model.

The Semantic Testlet Model: Integrating Natural Language Processing into Item Response Theory
(William Muntean, Zhuoran Wang, & Shonai Someshwar)
Traditional item response theory (IRT) models often fail to capture the complex influences of shared semantic content across test items, resulting in local item dependencies and unexpected response patterns. This research introduces the Semantic Testlet Model (STM), which integrates quantitative representations of item meaning derived from natural language processing into psychometric modeling. Unlike traditional testlet approaches with discrete item assignments, STM employs content-driven semantic clusters and models the influence of shared features through fuzzy group membership. By explicitly accounting for semantic relationships, the STM can identify anomalous dependencies in item responses, potentially revealing content-specific preknowledge or systematic item compromise. Comparative analyses with standard Rasch, traditional testlet, and multidimensional IRT models demonstrate that STM offers a unique approach to modeling item dependencies.

V6	**When the Lawyers Come Knocking**
Time:	3:15 PM - 4:15 PM
Speakers:	Camille Thompson, Jennifer Semko, & Rachel Schoenig

Taking action following an exam security incident carries the risk of conflict and, potentially, litigation. This session will address how exam security practitioners can position their program to limit that likelihood of dispute and, in the event a dispute arises, how to respond in an objective and defensible manner. Legal experts with decades of experience will address the need to navigate both the court of law and the court of public opinion when incidents occur. Together, we will explore how to address issues related to board communications, evidence collection, and stakeholder management and help you avoid some of the pitfalls that can arise. Attendees will leave feeling more confident in how to respond when the lawyers come knocking!

Thursday, November 13

V7	Practical Considerations to Improve Test Security
Time:	11:00 AM - 12:00 PM
Speakers:	Sergio Araneda, David Foster, Daniel Gualtieri, Olukayode Emmanuel Apata, Segun Timothy Ajose, & John Oluwaseun Ajamobe

A Phenomenological Framework for Stakes Definition in Educational Testing
(Sergio Araneda & David Foster)
The classification of tests as "low-" or "high-stakes" is often based on institutional assumptions rather than systematic criteria. This paper proposes a model for defining test stakes grounded in the lived experiences of test-takers, drawing on an experiential framework informed by Deweyan principles of continuity, interaction, and sense-making. By focusing on the pre-experience phase—when individuals anticipate possible test outcomes—this approach captures how people internalize stakes in cognitive, emotional, and behavioral terms. Through structured reflection on potential outcomes and personal goals, test-takers generate quantifiable, context-sensitive profiles of what is at stake for them. This individualized, phenomenological method challenges binary classifications and recognizes that any test may carry high stakes for someone, depending on their situation. The framework also provides a foundation for identifying populations with greater incentives to engage in misconduct, enabling future extensions into risk/reward models for predicting cheating behavior. By reframing how stakes are conceptualized and measured, this study offers a new perspective on fairness, validity, and security in educational testing.

1% Better: Improving Exam Security Every Day
(Daniel Gualtieri)
Exam security for organizations can always improve. However, there is often an expectation for organizations to go from a fixed-form, multiple choice exam utilizing a small item pool to a LOFT exam with a long list of options and a huge item pool. That leap cannot happen overnight, nor should it. The presentation will focus on taking the next logic step for test security, meeting organizations where they are and helping their program get 1% better every day.

Academic Shortcuts and AI Disclosure: Emerging Threats to Academic Integrity in Literature-Based Assessments
(Olukayode Emmanuel Apata, Segun Timothy Ajose, & John Oluwaseun Ajamobe)

The Consensus App, a generative AI-powered academic search engine, is reshaping how students engage with literature-based assessments, especially in open-book and unproctored testing contexts. Although designed to enhance academic efficiency by retrieving and synthesizing peer-reviewed literature, the tool introduces new challenges for test integrity. This presentation draws on a recent rapid review examining academic reporting and ethical considerations surrounding the Consensus App. We included five peer-reviewed articles and five editorial commentaries to identify the patterns of use and disclosure among researchers and students. Findings reveal a significant underreporting of Consensus App use in published articles, which raises concerns about its unacknowledged application in higher education. The session explores how students may rely on the app to shortcut reading-intensive tasks during research-based tests, blurring the lines between legitimate assistance and academic dishonesty. The lack of institutional policy on AI disclosure in such contexts contributes to underreported violations, resulting in lapses in detection within test security systems. Attendees will consider real-world scenarios and receive practical recommendations for designing AI-aware assessments. This discussion is especially relevant now that many institutions are shifting to online and AI-integrated assessments, which makes it more challenging to ensure fairness and maintain the integrity of testing.

V8	**Engaging Your Ecosystem to Enhance Security**
Time:	11:00 AM - 12:00 PM
Speakers:	Jarret Dyer, Isabelle Gonthier, Jim Hussey, Rachel Schoenig, & Ray Nicosia

Testing programs are just one part of a much broader ecosystem, one that includes educators, trainers, proctors, vendors, SMEs, credential holders, test takers, score users, and often the public. Trying to implement exam security practices without engaging the broader ecosystem is like trying to tie your shoelaces with your tongue – you can try, but there are much easier (and safer) ways to proceed. Just like using your fingers to tie your shoes is much more effective, so is engaging your ecosystem to partner with you on exam security. This session will explore ways to identify and engage the different parts of your testing ecosystem. Join security experts as they share how to effectively engage your unique ecosystems – and avoid tripping over your own laces – on your path to exam security success!

V9	Practical Considerations to Improve Test Security
Time:	12:05 PM - 1:05 PM
Speakers:	Joe Betts, William Muntean, Murat Kasli, Shonai Someshwar, William Muntean, Luping Niu, Zhuoran Wang, & Onur Demirkaya

Using Regression & Markov Transition Matrices for Identifying Anomalous Responses

(Joe Betts, William Muntean & Murat Kasli)

When candidates fail their licensure/certification exams, it is important to monitor retest scores for anomalous score changes that could have resulted from extra-curricular methods such as obtaining outside help or harvested items to prepare. This presentation will provide information about automated indices that can be used to identify anomalous score changes. In addition, it will provide a methodology for investigating the probability of those score changes using a regression-based model, residual analysis, and a Markov Transition Matrix method using both change in scores and change in responding times. It will be shown that the combination of excessive score gains with significant decreases in response times are a useful indicator of anomalous behavior that should trigger more direct investigations.

Person-Fit Monitoring in Polytomous CAT: Comparing CUSUM and Change-Point Approaches

(Shonai Someshwar, William Muntean, Luping Niu & Zhuoran Wang)

Computerized Adaptive Tests (CATs) are widely used in licensure testing due to their efficiency and psychometric precision. However, person-fit detection methods are limited, particularly in their applicability to variable-length tests and polytomous scoring. This study evaluated two promising methods—cumulative sum (CUSUM) and change-point (CP) statistics—for detecting aberrant response behavior in mixed-format, variable-length CATs. A simulation framework was developed using an item pool of dichotomous and polytomous items, with parameters generated under the Rasch Partial Credit Model (RPCM). Honest examinees were simulated using model-consistent response behavior, while aberrant examinees followed a three-phase model simulating warm-up effects, item pre-knowledge, and fatigue. CUSUM was adapted to polytomous items by computing standardized residuals based on category-level expectations, while CP statistics (Wald and likelihood ratio) were extended using likelihood comparisons under RPCM. The study introduces a practical approach for setting CUSUM decision thresholds based on test length. It further evaluates the performance of both methods using fixed-threshold metrics (Type I error and power) and threshold-independent Receiver Operating

Characteristic (ROC) analysis. The results demonstrated the feasibility and flexibility of person-fit statistics for modern licensure CAT programs.

Detection of Item Preknowledge Using Proportion-Correct Scores and Structural Change Tests
(Onur Demirkaya)
This study explores the use of structural change analysis with score-based statistical tests to detect item preknowledge in dichotomously scored assessments. Item preknowledge, when examinees gain prior access to test content, poses a significant threat to test validity by inflating scores and undermining fairness. In the literature, proposed methods for detecting item preknowledge often require item response theory (IRT) modeling. In contrast, this study proposes a more accessible approach using percent correct (proportion correct) as the focal statistic. Simulations varying in test length, percentage of compromised items, and item exposure-preknowledge correlation were conducted to evaluate Type I error rates and detection power. Results indicate that the performance of a score-based test is promising, especially when there is partial information about which items may be compromised. The method was also applied to operational data from two linear forms of a computerized licensure exam, demonstrating practical utility in real-world settings. The approach is particularly advantageous for tests using classical test theory (CTT) or small populations, as it avoids re-estimating item or examinee parameters and is a promising method for supporting test security through post-administration monitoring.

V10	**Ongoing Challenges of Remote Test Administrations for State Assessments: Recent Changes Implemented in 2025 to Improve Test Security**
Time:	12:05 PM - 1:05 PM
Speakers:	John Olson, David Ragsdale, Shaun Bates, Timothy Butcher, & Walt Drane

In 2024-25, many additional states included a Remote Test Administration (RTA) option for their assessment systems. Recently, more states have moved to using RTAs—this method is now a regular part of most assessment programs. A fast-growing new need in some states is using RTA for virtual charter schools. Other states tried RTA, only to subsequently drop its use. These new approaches, and related challenges, have required many changes, e.g., adopting new state policies/procedures to allow assessments to be monitored and administered

remotely. Most importantly, while the benefits of RTA are generally appealing, it cannot be implemented without maintaining test score validity and trustworthiness. The move to RTA also encourages that technology be better leveraged by states in creative ways to help combat test fraud, with multiple methods combined to maximize protection. There are many lessons learned by current and former users. This session promises to inform participants of the challenges and crucial changes made by states that have securely administered RTAs in recent years. Three state assessment directors and two test national-level security experts will describe how states do RTAs, or decided to stop their use, and the procedures that were implemented to maintain security during test administrations.

V11	**Unfinished Business: Navigating the Frustration of Unresolved Threats**
Time:	2:05 PM - 3:05 PM
Speakers:	Megan Rees & Heidi Green

In investigations, success is often measured by swift detection and decisive action. But what happens when a case stalls- when threats are identified, but evidence is insufficient, jurisdictions are uncooperative, or leads can't be followed any further? This session explores the challenge of unresolved cases—where threats are known but action is limited.

Attendees will learn how to manage ongoing risk, communicate uncertainty, and navigate the emotional toll of unfinished investigations. Practical strategies and psychological tools will help professionals stay effective, ethical, and resilient in the face of lingering threats.

V12	**Identifying Security Vulnerabilities**
Time:	2:05 PM - 3:05 PM
Speakers:	Iain Holland, Yong-Siang Shih, Basim Baig, & Glenn Milewski

Red Teaming for Exam Security
(Iain Holland)
How do you critically examine your test security, identify unique and novel attack vectors, and convince people to enthusiastically participate in security training in just one 15-minute session? Just run a Red Teaming exercise.

Red Teaming is the process of taking on the role of a hostile actor and seeking to identify vulnerabilities in security systems from their perspective. Often used in police and military contexts, the concept can bring significant value to companies looking to improve their own security.

Iain Holland, the Assessment Security Lead at Oxford University Press will be running through how he delivered a Red Teaming exercise for the Assessment Team. He will look at the benefits of running these kinds of exercises as part of your own security program, what kind of outputs you can get, and some tips on running these yourself.

Using AI to Improve Secondary Camera Setup in Duolingo English Test
(Yong-Siang Shih & Basim Baig)
Setting up a secondary camera is a requirement for the Duolingo English Test, yet it sometimes leads to confusion and frustration for test takers. In this presentation, we share how the Duolingo English Test redesigned this process by introducing an AI-powered guided setup experience that supports test takers through real-time feedback. By dynamically detecting errors and guiding test takers toward successful camera placements, the system helps users complete the setup confidently and independently. We'll walk through the development and deployment of the model, and share insights into how it improves both usability and security. This case study highlights how investing in thoughtful, user-centered design can remove friction from high-stakes testing while strengthening the overall integrity of the process.

Caught on Camera: Lessons Learned from Real-World At-Home Security Incidents
(Glenn Milewski)
When the Independent School Entrance Exam (ISEE) transitioned to at-home testing during the COVID-19 pandemic, many predicted a temporary shift. Five years later, however, remote testing constitutes a third of ISEE tests administered per year. While this modality has significantly expanded market reach as well as customer expectations, it has also increased security challenges, prompting the addition of new solutions to mitigate threats. This session explores the ISEE Program's strategic decision to continue and enhance remote testing to fulfill mission objectives and maintain a competitive advantage. It highlights the security interventions that have been effective, such as dual-camera monitoring, personalized watermarks on question screens, a lower proctor-to-student ratio, and carefully controlled test environments. Through real-world examples of security breaches and the

countermeasures deployed, attendees will hear about humbling lessons learned and actionable ideas for strengthening their own programs.

Closing Keynote: COTS Debates	
Time:	3:10 PM - 4:10 PM
Debaters:	Jim Wollack, Claire McCauley, Isabelle Gonthier, & Steve Addicott
Moderator:	Rachel Schoenig
Join us for the closing COTS Debates! This always informative keynote session will feature industry luminaries as they debate topics of interest to testing professionals.	

SAVE THE DATE:

ADDITIONAL RESOURCES

Optional Group Dinners

UMass faculty and students have put together a list of recommended local restaurants, which you can find through this QR code. There are a few opportunities to join group dinners on both nights by filling your name in the spot provided. All guests are responsible for paying for their own meals.

Campus Walking Tours

Join UMass faculty and students for a short walking tour of the campus. Weather permitting, there will be a walking tour during the Tuesday lunch, and after the conference at 1:00 on Wednesday. Listen for announcements about where to meet for these tours.

Presenter Information

Presenters are responsible for managing their own time during the session. Please be mindful of the schedule and ensure your presentation, including any Q&A, stays within the allotted time.

If you encounter any technical issues during your presentation, please call (413) 577-8222.

Food Allergies & 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 questions, please speak with one of the hotel staff or visit the registration desk.

UMASS WI-FI INSTRUCTIONS

1. In the WiFi settings on your device, select the **UMASS network**.
2. Open up a web browser (IT suggests using Safari or Firefox).
3. Type in the address bar: http://login.wireless.umass.edu.
    a. Either a page telling you that this site is unsecure will present itself OR the UMass wireless login page will appear.
    b. If it is a page telling you the website is unsecure, click “More Details” or “Advanced” at the bottom and click on the link to visit the page anyway. You should then be on the UMass wireless network login page.
4. Enter the credentials below:
    - **Guest ID: 78474059**
    - **Password: 00443681**
5. Note: Guest accounts will not work on eduroam.

NOTES

NOTES

NOTES

THANK YOU TO OUR SPONSORS

CO-HOSTS:

Pearson VUE

Honorlock

caveon

Respondus

Integrity Advocate

duolingo english test

FRIENDS:

NCTA NATIONAL COLLEGE TESTING ASSOCIATION

ncme NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

TALVIEW

psi

PROMETRIC

WIDA UNIVERSITY OF WISCONSIN–MADISON

DEMONSTRATORS:

Surpass Powering Assessment

iTS