THE 11TH ANNUAL

CONFERENCE ON TEST SECURITY

OCTOBER 26-28, 2022

PRINCETON, NJ

Dear Colleague,

It is my pleasure to welcome you to the 11th annual Conference on Test Security in beautiful Princeton, NJ, to once again discuss methods and best practices in and around test security. I think I speak for all of us in expressing my enthusiasm to finally be able to convene in-person after our plans to do so were derailed by the pandemic in both 2020 and 2021. I can safely say that this year's program is not only delayed by two years, but wildly different than it would have been absent the pandemic. As an industry, we are no longer flying by the seat of our pants responding to the challenges imposed, but we are approaching a new normal in which we have learned some hard-earned lessons and established new guidelines and best practices for keeping our tests secure.

This year's conference will begin with a welcome reception and movie showing on the evening of 10/26, followed by a full day of sessions on 10/27. The evening of 10/27 will also showcase our traditional "Cocktails and Conversations" poster event. The conference will end with a half day of sessions on 10/28, including the ever popular Keynote Debates.

Finally, a special thank you to our sponsors for making this year's conference possible. On behalf of the entire COTS Executive Committee, I thank you for your generosity and commitment to the field of test security.

I hope that you enjoy your time in Princeton at the 2022 Conference on Test Security.

Sincerely,
Carol Eckerly
2022 Conference Chair

A SPECIAL THANKS TO OUR HOST

A SPECIAL THANKS TO OUR CO-HOSTS

ascend LEARNING

caveon™ Test Security

examity® BETTER TEST INTEGRITY.

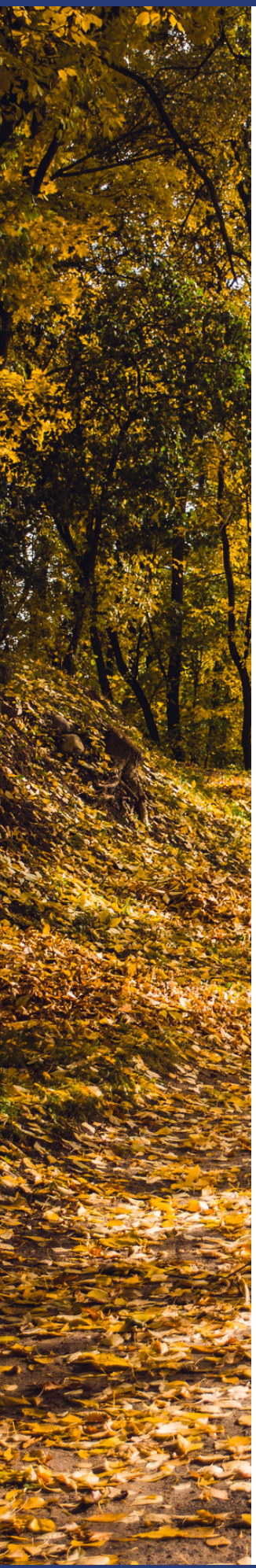
MEASURE LEARNING

proctorU yardstick

Pearson VUE

A SPECIAL THANKS TO OUR FRIENDS

duolingo
english test



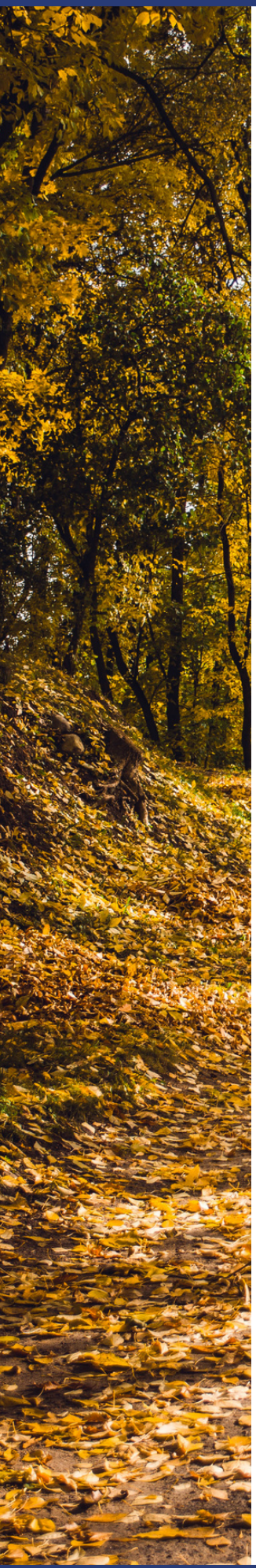
CONFERENCE AGENDA

WEDNESDAY, OCT 26

3:00 p.m. - 7:00 p.m.	Packet Pickup and Information *Assembly Area*
4:30 p.m. - 5:30 p.m.	Meet and Greet *Prince William Ballroom*
7:00 p.m. - 9:30 p.m.	Movie Showing *Prince William Ballroom*

THURSDAY, OCT 27

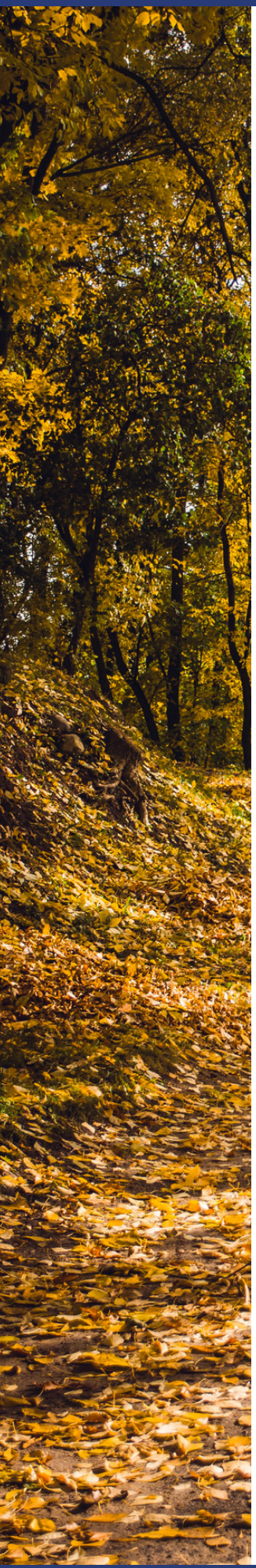
7:30 a.m. - 5:00 p.m.	Packet Pickup and Information *Assembly Area*
7:30 a.m. - 8:15 a.m.	Breakfast *Prince William Ballroom*
8:15 a.m. - 9:30 a.m.	Welcome and Opening Keynote *Prince William Ballroom*
9:30 a.m. - 9:45 a.m.	Break
9:45 a.m. - 10:45 a.m.	Sessions *Prince William Ballroom, Albert Einstein Room, Nassau Room, Palmer Room*
10:45 a.m. - 11:00 a.m.	Break



CONFERENCE AGENDA

(CONTINUED)

11:00 a.m. - 12:00 p.m.	Sessions *Prince William Ballroom, Albert Einstein Room, Nassau Room, Palmer Room*
12:00 p.m. - 1:00 p.m.	Lunch *Prince William Ballroom*
1:00 p.m. - 2:00 p.m.	Sessions *Prince William Ballroom, Albert Einstein Room, Nassau Room, Palmer Room*
2:00 p.m. - 2:15 p.m.	Break
2:15 p.m. - 3:15 p.m.	Sessions *Prince William Ballroom, Albert Einstein Room, Nassau Room, Palmer Room*
3:15 p.m. - 3:30 p.m.	Break
3:30 p.m. - 4:30 p.m.	Sessions *Senior Room, Albert Einstein Room, Nassau Room, Palmer Room*
4:30 p.m. - 5:00 p.m.	Extended Break
5:00 p.m. - 6:30 p.m.	Cocktails and Conversations Poster Presentations *Prince William Ballroom*

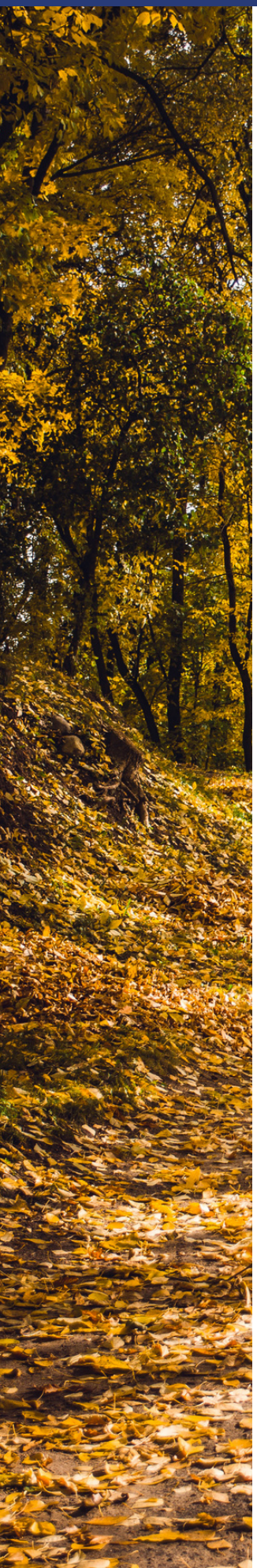


CONFERENCE AGENDA

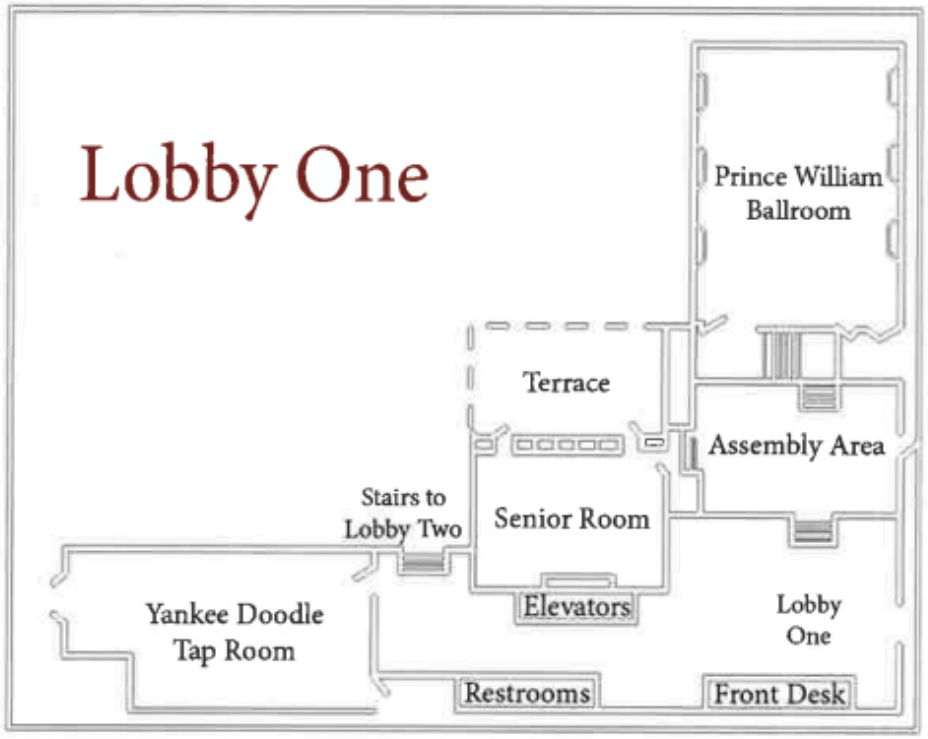
(CONTINUED)

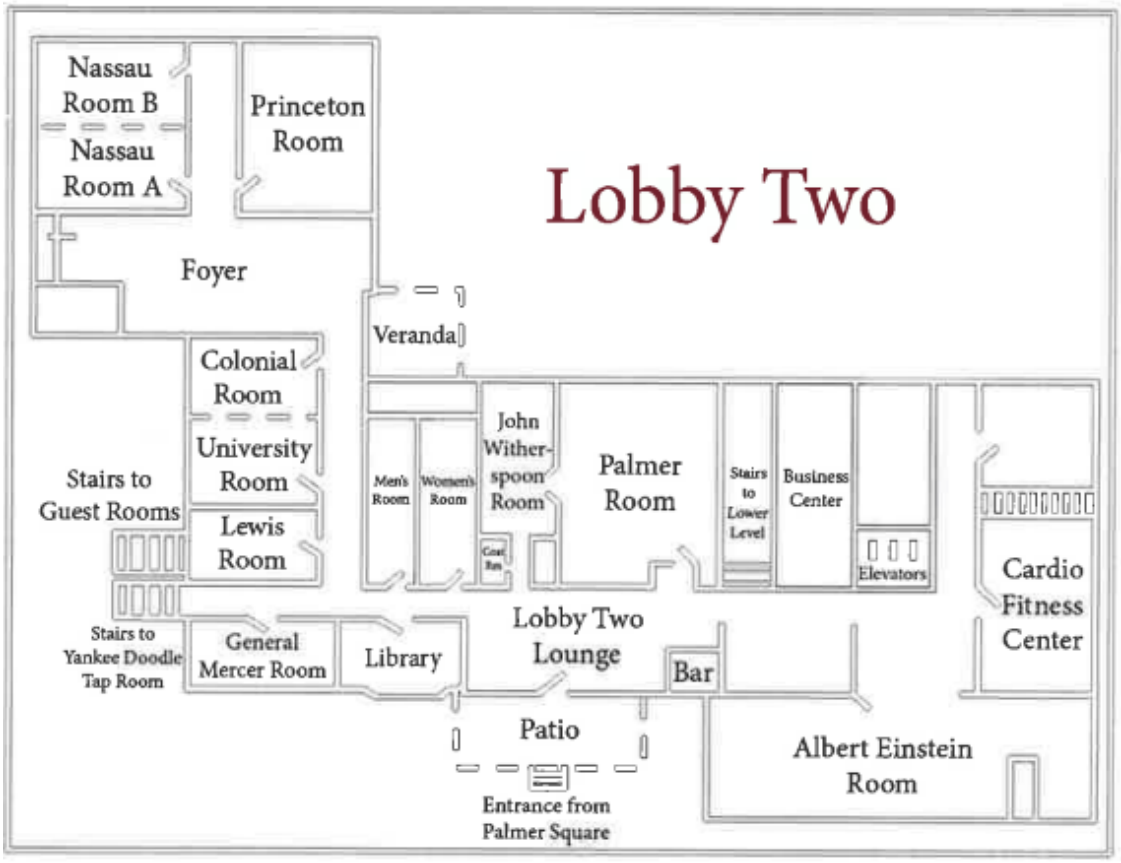
FRIDAY, OCT 28

7:30 a.m. - 8:15 a.m.	Breakfast *Prince William Ballroom*
8:15 a.m. - 9:15 a.m.	Sessions *Prince William Ballroom, Albert Einstein Room, Nassau Room, Palmer Room*
9:15 a.m. - 9:30 a.m.	Break
9:30 a.m. - 10:45 a.m.	Closing Keynote Debates *Prince William Ballroom*
10:45 a.m. - 11:00 a.m.	Break
11:00 a.m. - 12:00 p.m.	Sessions *Prince William Ballroom, Albert Einstein Room, Nassau Room, Palmer Room*
12:00 p.m. - 1:00 p.m.	Lunch *Prince William Ballroom*



NASSAU INN MAP





WEDNESDAY, OCT 26
4:30 P.M. – 9:30 P.M. EDT

4:30 - 5:30 *Prince William Ballroom*

Meet & Greet

Light refreshments and drinks will be served

7:00 - 9:30 *Prince William Ballroom*

MOVIE SHOWING: SETTERS

Join us to view the 2019 film **Setters,** loosely based on the true events of the Vyapam Scam. The Vyapam Scam was an entrance examination scandal that took place in the Indian states of Madhya Pradesh and Uttar Pradesh involving bribery, proxy test taking, item theft, and more. The film is shown in Hindi with English subtitles.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

Conference On Test Security 2022

7:30 - 8:15

Prince William Ballroom

Breakfast: A buffet is available in the Senior Room. Seating is in the Prince William Ballroom.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

8:15 - 9:30

Prince William Ballroom

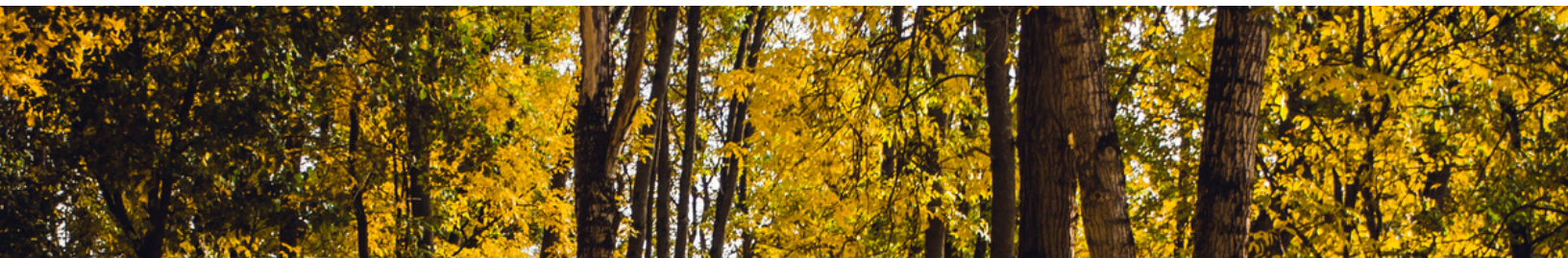
Welcome to COTS 2022

Jim Wollack, University of Wisconsin-Madison
Ida Lawrence, ETS

OPENING KEYNOTE

Camille Thompson, College Board | Marc Weinstein, Marc J Weinstein PLLC and Caveon Test Security | Rachel Schoenig, Cornerstone Strategies

A federal court in Ohio recently decided that a room scan conducted during a remote proctored online test at a state university amounted to an unreasonable government search in violation of the Fourth Amendment of the United States Constitution. The case has generated significant interest and commentary across the education and assessment space. Join experienced testing experts and attorneys as they discuss the case and key takeaways for testing programs.



THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

9:45 - 10:45

Palmer Room

Weathering a Test Security Storm

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

If your exam has value, it is not a question of IF you will experience an exam security incident but WHEN. How those incidents are handled can have a profound impact on your testing program. Respond well, and you can effectively protect score validity and your program's reputation; respond poorly, and you can find your program embroiled in litigation and going viral on social media. During this session, we will discuss the process for effective incident response planning, common exam security risks, and pitfalls to avoid. Join testing experts in an interactive and engaging workshop as they review exam security incidents and provide practical advice for test security incident response planning that can help your program weather a test security storm.

Prince William Ballroom

STANDARD PRESENTATIONS I

Visualizing the Prevalence of Collusion/Content Exposure in the Testing Population

Brett P. Foley Alpine Testing Solutions

The development of score similarity and answer similarity indices in recent years has added to our psychometric arsenal for detecting candidate collusion and content exposure. These indices have been shown to be effective and often are calculated by comparing the expected and observed number of matching item scores/responses for each pair of candidates. These results are useful for identifying pairs of candidates with questionable levels of agreement. However, as candidate volume grows, the number of candidate pairs grows quadratically, resulting in very large numbers of pairs for even modest sample sizes. Given the extreme number of possible comparisons, it can be difficult for non-technical audiences to understand the extent of an exposure problem. In other words, is a security issue negligible, isolated, or widespread?

In this session, the presenter will begin by briefly describing best practices for creating exam-security related visual displays. Next, we will present several examples of visual displays that use data from all candidate pairs to effectively communicate the prevalence and extent of collusion and content exposure issues in the testing population. The session is appropriate for both technical and non-technical audiences. Attendees will learn about important factors to consider when creating visual displays, be shown specific examples of effective tools for summarizing collusion/content exposure issues, and learn how scaling group frequencies can help highlight outliers or emphasize common patterns.

A Comparison of an Approximation Score Similarity Index to Other Similarity Indices

Amanda Wolkowitz, Alpine Testing Solutions | Russ Smith, Alpine Testing Solutions

This presentation will compare the results of multiple score similarity indices to an approximation score similarity index (aSSI) (Smith 2021, 2022). The goal of the study is to determine if aSSI produces similar results to other score similarity indices when applied to real data. Prior research has shown that aSSI does a reasonably good job of producing results similar to true score similar index (SSI) analyses (Smith 2021, 2022), the latter of which uses item response theory (IRT) and the generalized binomial model. However, research is lacking on the practical comparison of aSSI to other similarity indices on real data. The outcome of this study will be recommendations for practitioners on when to apply the aSSI method to real data and how it compares to other existing and more complicated methods.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

9:45 - 10:45 *CONTINUED...* *Prince William Ballroom*

STANDARD PRESENTATIONS I *CONTINUED...*

Digging Deeper into the Results of Similarity Analysis to Improve Test Security

Marcus Scott, ***Caveon***

Analyzing tests for similarity has proven to be a powerful technique for detecting preknowledge, collusion, proxy test takers, and other threats to test security. Because of its high power to detect security issues, similarity analysis should be an essential component of a program's data forensics analysis, provided the testing mode supports its use (e.g., similarity usually does not lend itself well to CAT because test instances typically have little item overlap). Although similarity analysis detects anomalous test instances, the findings from similarity analysis are not necessarily restricted to individual tests. This presentation will show how digging deeper into the results of similarity analysis can yield valuable insights into the security of a testing program and help identify security threats. Examples from real-life data sets will be presented, such as:

1. Using similarity analysis to identify examinees who took the exam with multiple registrations, allowing them to bypass the retake policy.
2. Using similarity analysis to infer that test content was disclosed and determine its quality.
3. Using similarity analysis to determine how long it takes for test content to be disclosed, which can help with republication strategies.
4. Using similarity analysis to track how quickly use of disclosed content spreads through the testing population (i.e., after content is disclosed, how prevalent is its use?)
5. Using similarity analysis to identify issues with exam structure, such as too much item overlap between forms.
6. Using similarity analysis to identify tests taken via proxy test takers.

Nassau Room

State Panel Discussion on Maintaining and Improving Test Security for State Assessments during the Ongoing COVID-19 Pandemic

Walt Drane, Caveon Test Security | John Olson, Olson Educational Measurement and Assessment Services | Sandra Greene, Georgia Department of Education | David Ragsdale, MA Dept of Elementary and Secondary Education | Jessica Fenby, Michigan Department of Education

Panel Presentation

Over the past three years, states have learned many things on how to maintain and improve security for their state assessment programs. The COVID-19 pandemic significantly disrupted testing in 2020 and 2021, preventing normal standardized testing from taking place. Many states had to make numerous adjustments to plans for administering their assessments, and in the process became much more flexible in the security procedures implemented to minimize irregularities from occurring.

In this session, a panel of three states will discuss the various activities they implemented in the past three years to improve test security for their summative assessments. Details from a variety of state perspectives will be shared, with each state focusing on how they maintained the security of their assessments despite the ongoing challenges. The Panel, led by two nationally recognized experts on security for K-12 assessment programs, includes representatives from Georgia, Massachusetts, and Michigan, who will share information on their recent activities and describe effective strategies for securely assessing students in grades 3-12 in coming years. Panel discussions will focus on several important issues –

(a) Methods and approaches for maintaining and improving security in state assessments
(b) Various options for monitoring test administrations, both in the classroom and remotely
(c) Challenges in meeting USED requirements for Peer Review, especially for test security and monitoring of test administrations
(d) Important things learned when making changes to state assessment systems during the pandemic
(e) Best practices for minimizing test irregularities, preventing cheating, and maintaining the validity of state assessment results

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

9:45 - 10:45 *CONTINUED...* Albert Einstein Room

Raising your Exam Security

Cory Clark, Meazure Learning | Simon Welham, Mulesoft | Alison Kramer, Caveon Test Security
Panel Presentation

Cheating and content theft can seem like a never-ending cat-and-mouse game for global, high-stakes programs like MuleSoft. Online exam security is critical to prevent proxy testing, identify potential misconduct, and protect content from being compromised or shared, but security doesn't begin and end with the proctor. To ensure MuleSoft's exams are protected, consistent and innovative security processes must be used and evaluated for effectiveness. Collaborating with several responsive and forward-thinking exam security providers, MuleSoft has deployed a constant monitoring and enhanced security strategy that allows program leaders to not only catch "cheaters" but also identify integrity breaches and continually develop new tactics.

During this presentation, we'll review a case study in which Meazure Learning's enhanced security program detected data points identifying potential collusion, which led to a deeper analysis by Caveon's data forensics team. We'll state the analysis outcomes as well as MuleSoft's response and policy changes given the advancements in available detection tools. Hear firsthand how the collaborative approach between three industry leaders provides a balance of security and IP protection while also focusing on the overall candidate experience. These organizations work closely to identify security threats and create scalable solutions with highly defensible outcomes. Attendees will get a peek behind the enhanced security curtain and a broader understanding of the importance of partner collaboration when it comes to protecting global exam programs

11:00 - 12:00 Palmer Room

STANDARD PRESENTATIONS II

Time Keeps on Ticking: Advancements in the Automation of Data Forensics Allow Timely Action

Sarah Toton, Caveon Test Security | Carissa Redfield | PMI

This session presents a collaborative case study where a sophisticated and fully automated Near-Real-Time Data Forensics (NRTDF) system was developed to analyze PMI certification exam data on a daily basis. The NRTDF system automatically transfers and analyzes data, returning results well within 24 hours after administration, all without manual intervention. This dramatically reduces the delay between administration and data forensics analysis.

Timely data forensics analysis of test data is highly desirable, especially when candidate scores may be invalidated; however, strong actions require powerful statistics to produce reliable, contextualized, and defensible results. Significant test volumes are required to ensure statistical model stability, but high volumes are often not present in short timeframes. Moreover, streamlined and robust infrastructure must be in place to reliably perform computationally intensive analyses and produce outputs that can be manageably consumed by the client. Thus, historically, the frequency of statistical analysis of test data has been weighed heavily against the effort required to overcome these obstacles.

The NRTDF system represents a substantial step forward in overcoming these challenges by providing a streamlined process for quickly retrieving and analyzing data, combined with a creative approach to gaining the necessary statistical power to produce defensible results.

11:00 - 12:00 *CONTINUED...* *Palmer Room*

STANDARD PRESENTATIONS II *CONTINUED...*

Efficient Answer Similarity Analysis with Dichotomous Scores using Lookup Tables and Bitwise Operations

Carol Eckerly, ETS | Ben Babcock, Elsevier | Kylie Gorney, University of Wisconsin-Madison, Mridul Aanjaneya, Rutgers University

Similarity analysis is commonly conducted to detect security threats to exam programs. While similarity indices were designed to detect pairs of examinees that have unusually similar item responses, cluster analysis can be used to group examinees into clusters of three or more examinees using a similarity index to obtain measures of distance (Wollack and Maynes, 2017). However, 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 to enable clustering may be prohibitive for large testing programs. This presentation introduces a computational method that relies on the use of the Rasch IRT model in computing van der Linden and Sotaridona's (2006) GBT similarity statistic. In comparison to other IRT models, use of the Rasch model reduces the complexity of the problem by enabling the use of lookup tables to determine if the GBT is significant for each pair of examinees. Additionally, bitwise operations are used to efficiently count the number of matching responses between all pairs of examinees. Implementing both of these strategies results in a large decrease in computation time (e.g., for 1,100 examinees, computation time decreases from about 4 hours to 2.5 minutes). The method is illustrated using three real datasets with known collusion, and results are compared to those using competing IRT models in terms of cluster agreement and computation time.

Approximation Score Similarity Index (aSSI) Analysis: An Analysis of its Effectiveness Compared to True Score Similarity Index

Russell Smith, Alpine Testing Solutions

This presentation will describe continued research and development into a simple and effective approach to identifying pairs of test takers with an unusually high number of matching scores (Smith, 2021; 2022). True score similarity index (SSI) analyses leverage item response theory (IRT) and the generalized binomial model (van der Linden & Sotaridona, 2006; Zopluoglu, 2017) which require specialized software, are based on relatively strong assumptions, and, depending on the length of the exam and number of examinees, require substantial computational resources. The approximation score similarity index (aSSI) method is not intended to outperform true SSI. Rather, it provides reasonable estimates of pairwise probabilities without specialized software, without requiring the same computational resources, and (in many cases) without item-level data. The purpose of the presentation will be to describe the aSSI method, further explore the conditions in which it works reasonably sufficiently by comparing it to true SSI, and to make practical recommendations for its use.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

11:00 - 12:00 *CONTINUED...* *Prince William Ballroom*

What Newbies Need to Know about Test Security

Nicole Rhyne, Cornerstone Strategies | Jeff Marsh, Ascend Learning | Nicole Miller, NBME | Camille Thompson, The College Board

Panel Presentation

New-ish to testing or test security? Then this is the session for you! The reality is, test security doesn't start at administration of the exam, and the job is about way more than busting cheaters. But what's a new professional to do when trying to understand this unique aspect of testing? Attend this session, of course! Join test security professionals at different stages of their career as they discuss how to get up to speed on exam security and review some of the foundational principles. We'll share key information to know for new professionals, discuss helpful resources, provide insights around communicating with the rest of the organization, and answer questions you may have. Our goal: providing you a solid foundation for success in your role.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

11:00 - 12:00

CONTINUED...

Nassau Room

How Investigations Have Changed in an ONLINE World

Brent Morris, Cisco | Bryan Friess, Pearson
Facilitated Roundtable

The adoption of online proctoring (OP) as a delivery methodology was already experiencing moderate growth, but since the global pandemic, that has accelerated significantly. As a result, the implications of moving exam volume to OP has become a hot topic across industry disciplines. Getting far less attention, however, are the differences found when investigating incidents of misconduct in online-proctored delivery versus traditional brick-and-mortar test center delivery. Unsurprisingly, the two can differ greatly. How are different IT sponsors handling these types of investigations? What are programs doing differently? From what we have seen, what should change and what do we expect in the future?

12:00 - 1:00

Prince William Ballroom

Lunch: A buffet is available in the Senior Room. Seating is in the Prince William Ballroom. Overflow seating is available in the University and Colonial Rooms.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

1:00 - 2:00

Palmer Room

Lessons Learned from COVID: Laying the Path for Future Test Security

Lestelle Schwab, Western Governors University | Carissa Pittsenberger, Western Governors University | Kurt Grabner, Examity
Facilitated Roundtable

The real life "COVID laboratory" gave us an unprecedented vehicle for developing and improving our test security programs. The constantly changing and evolving circumstances related to COVID pushed our testing methods and innovations to never-before seen levels. Having come through a crisis of such proportions, it is now time to look back and review our lessons learned: what worked, what did not, what did we plan for successfully, and what could we not have foreseen. There is power in collective knowledge and sharing this collective wisdom will help us better prepare for the future. We will take real time input during this discussion and aggregate what we have learned and share it with the group. Each participant will leave with a prioritized collection of solutions and innovations to take with them to help advance their own test security program.

Prince William Ballroom

STANDARD PRESENTATIONS III

Applications of Bayesian Decision Theory to Detect Test Fraud

Sandip Sinharay, ETS | Matthew Johnson, ETS

This paper suggests an approach based on Bayesian decision theory for the detection of test fraud. Guidance is provided on (a) the reformulation of the problem of detection of test fraud as a decision-theory problem, (b) the choice of appropriate loss functions and probability distributions, which are critical components in applications of statistical decision theory, (c) the computation of posterior expected loss, the measure that is minimized to determine whether an examinee should be flagged for possibly committed test fraud. The decision-theoretic approach will be applied to detect answer-copying and item preknowledge for two credentialing data sets described in Cizek and Wollack (2017, p. 14) and the results will be compared to those from the K-index (e.g., Lewis & Thayer, 1998) and a score-differencing approach (e.g., Finkelman et al., 2010) applied to the same data set.

Using Item Scores and Distractors to Detect Item Compromise and Preknowledge

Kylie Gorney, University of Wisconsin-Madison | James A Wollack, University of Wisconsin-Madison | Sandip Sinharay , ETS | Carol A. Eckerly, ETS

Recent years have seen an abundance of research focused on detecting examinees with preknowledge (EWPs) when the set of compromised items (CIs) is known (e.g., Shu et al., 2013; Sinharay, 2017). However, few studies have attempted to detect both the CIs and the EWPs when neither is known, despite the immediate threat that this type of situation poses to the security of a test. Moreover, those that do consider the simultaneous detection of both items and persons tend to overlook a critical piece of information that is available in all multiple-choice data: distractor selection. Such information may be useful in identifying preknowledge of an incorrect answer key (e.g., Scott, 2018). In this study, we address this gap in the literature by introducing a new procedure that uses both item scores and distractor selection to detect the CIs and EWPs when neither is known. We consider two types of preknowledge: one in which a correct answer key has been disclosed, and one in which an incorrect answer key has been disclosed. Results are evaluated using the Type I error rate and power with respect to both CIs and EWPs.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

1:00 - 2:00 *CONTINUED...* *Prince William Ballroom*

STANDARD PRESENTATIONS III *CONTINUED...*

A New Two-stage Regression Approach to Detecting Section Score Inconsistency

Yi-Hsuan Lee, ETS | Charles Lewis, ETS

For an assessment with multiple sections measuring related constructs, test takers with higher scores on one section are expected to perform better on the related sections. When some sections are exposed while the others are not, test takers with preknowledge of the exposed sections may score unusually high on those sections compared to their performance on the unexposed sections. We propose a two-stage regression approach to detecting score inconsistency among different sections of a test. It assumes that the assessment has historical data that are representative of the test-taker population with relatively little evidence of cheating. In Stage 1, an outlier analysis is conducted that leverages information from the historical data to identify potential outliers in a new administration based on a specified criterion. In Stage 2, these outliers are temporarily removed from the new administration, and linear regression models are built with the laundered data to predict scores on the exposed sections from scores on the unexposed sections. The resulting linear regression models are applied to the scores of all test takers in the new administration. Externally studentized residuals and their p-values are calculated to flag individuals with inconsistent section scores based on a specified criterion. The criteria used in Stages 1 and 2 are derived analytically and empirically to produce a predetermined overall Type I error rate and to optimize detection power for the assessment under study. Borrowing information from historical data for the initial outlier analysis is found to be crucial when test security is a concern.

Nassau Room

Developing an Effective Playbook for Responding to Brain Dumps; Organizational, Psychometric, and Legal Perspectives

Rebecca Moden, CLARB | Adrienne W Cadle, Credentialing Professional Testing, Inc. | Marc Weinstein, Marc J Weinstein PLLC and Caveon Test Security

Panel Presentation

Candidate brain dumping of confidential test content is one of the most ubiquitous and insidious test security problems that every program must be prepared to address. It is not a matter of if it will occur; it is only a matter of when. What can you do when you learn that your program has been the victim of a brain dump? Does your program have a brain dump playbook?

In this session, you will learn how to develop a playbook to effectively respond to brain dumps. The Chief Operations Officer, psychometrician and attorney for a national licensure testing program will share the lessons of their experiences responding to brain dumps. Examples of these lessons include: (1) the importance of having a plan to respond to brain dumps; (2) understanding and mitigating the scope of the breach in terms of exposed test content and potential impacts on the validity of past and future test scores; (3) understanding the range of legally supportable actions your program can take in relation to candidates based on their role in the incident; (4) identifying the information and evidence required to inform program decision making; and (5) determining how your program will communicate with candidates, score users and possibly the public in relation to the incident. By the end of the session, attendees will be better prepared to respond promptly and confidently when they discover a brain dump.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

Conference On Test Security 2022

1:00 - 2:00 *CONTINUED...* *Albert Einstein Room*

This session has been canceled.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

2:15 - 3:15

Palmer Room

Lessons Learned from Federal Litigation of Cheating Involving a Test Preparation Company: Security is a Pre-requisite for Validity

Jon S. Twing, Pearson Assessments | Justin M. Keen, Assistant US Attorney | Phil Canto, Florida Department of Education | Bryan Friess, Pearson VUE
Coordinated Symposium

Owners of a test preparation company were sentenced to federal prison following a criminal prosecution stemming from theft of proprietary content from the State of Florida's Teacher Certification Exam program. U.S. Attorney Jason Coody stated that "The defendants' profiteering scheme is an insult to the dedicated public-school teachers and administrators of Florida, who studied and worked hard to become certified in their professions, Floridians expect and deserve to know that the public schools to which they entrust their children to learn are being led by teachers and administrators who properly earned their certifications. Today's sentence reiterates a valuable, but basic lesson. Notably, that hard work and diligence are rewarded, but acts of theft and dishonesty, as demonstrated by these defendants, are to be punished." This series of presentations will explore the facts and circumstances of the investigation and litigation, including due diligence and vigilance maintained prior to the theft of content, forensic and other investigations used to detect and document the theft of content, and measurement considerations regarding the impact associated with the exposure of content, and the session will conclude with summary of legal actions and the disposition of this successful prosecution. The audience will learn the complex and interdependent nature of security investigations, breech protocols and actions, the severity of the validity threat such actions have on assessment results, and the due diligence required to formally prosecute.

Prince William Ballroom

Security Considerations in an Online Testing World: Leveraging Opportunities while Maintaining the Integrity of your Exam Program

Isabelle Gonthier, PSI Services | Rachel Schoenig, Cornerstone Strategies
Facilitated Roundtable

Credentialing and testing organizations have faced a lot of changes within the last two years, with unprecedent pressure to reconsider how examinations are developed and administered in a world that had to transition to online meetings and online proctored exams. With that rapid movement to online development and delivery, many questions related to the security of the exam content and process have been raised, with a heavy focus on security considerations on online proctoring. Although these questions regarding online proctoring are critical, we need to address security across the entire assessment life cycle and implement adequate measures every step of the way, from working with subject matter experts, item development and item bank management to exam assembly, rendering and delivery through different modalities.

In this roundtable discussion, we will discuss opportunities that arose through this rapid movement to create more sustainable online exam development and delivery, including broader access to groups of subject-matter experts and increased accessibility. We will also discuss key considerations in leveraging technology to maintain the integrity of exams in this ever evolving digital world.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

2:15 - 3:15 *CONTINUED...* *Nassau Room*

Standard Presentations IV

Comparison of Forensic and Psychometric Statistics for Remote vs. In-Person Proctoring for Several Testing Programs

George W. Waugh, HumRRO | Philip Hinson, HumRRO | Amy McKee, HumRRO

Remote online testing has dramatically increased during the last couple of years—particularly for high-stakes testing for credentialing, selection, and education. In theory, it is easier to cheat and steal test items for remotely-proctored vs. on-site tests. This study compared remote vs. on-site proctoring with respect to psychometric equivalence and forensic statistics. Response and latency data were analyzed for testing programs of six of HumRRO's clients: one selection test, and five credentialing tests. Each testing program uses both remote and on-site testing for the same exam. The analyses use classical statistics (e.g., mean, standard deviation, score gain, factor analysis), Rasch (IRT) statistics (e.g., item misfit, person misfit, DPF, DIF), and collusion statistics. The results, which will be available this summer, will show the degree of measurement equivalence and degree of test fraud for on-site vs. remote proctoring.

Practical Applications and Operational Considerations for the Detection of Item Preknowledge and Compromised Content With Real Data

Linette P. Ross, NBME

Item preknowledge leads to compromised content and threatens the integrity of testing programs and the validity of the scores reported. Monitoring methods for detecting item preknowledge are needed to detect unexpected group behavior and identify potentially compromised content. This study examined the effectiveness of using data forensics and sequential monitoring of item performance with real data to detect item preknowledge and known compromised content. The study also examined the effect of compromised content on item performance and examinee ability estimates. Operational considerations included subgroup populations, exam administration type, and group-based threats associated with prior security breaches, multiple administrations, and testing centers. Recommendations for the use and setup of sequential monitoring of item difficulty for operational testing programs are provided.

Follow-up Analysis of Score Anomalies

Timothy Sares, ANCC | Lidia Dobria, Wright College

The Scored-Diff Statistic compares performance on scored items versus non-scored items
to detect examinees who potentially had pre-knowledge of the scored items. In this presentation, we will investigate score difference anomalies that appear below a typical chance threshold but that warrant further investigation.

We observed in a couple of reporting instances where a small number of examinees were routinely being flagged as having better performance on the scored items than the non-scored items. Although the number of examinees flagged did not exceed the 1 in 1,000 chance threshold, the score differences of at least 3.49 standard deviations were unusual nevertheless and warrant follow-up investigation into the why.

We explore through alternative quantitative DIF analysis some possible explanations for why these anomalies are repeatedly being flagged in the analyses. We hope to engage in a discussion on some interesting aspects for interpreting unexpected score variations.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

Conference On Test Security 2022

3:30 - 4:30

Albert Einstein Room

Solutions Both Old and New: Once-and Done Items and Test Forms to Prevent Cheating

David Foster, ***Caveon Test Security***

Demonstration

What if we were able to give completely unique test forms to every student, college applicant, or job candidate using the simple process of randomly sampling the items from a much larger pool of them? It would be exactly like shuffling the cards of a deck just before dealing them out to the players. No matter what the game is, everyone gets a unique "hand", and each player's skill is eventually measured by how well their hands are played. Switching from games to tests, the random selection of items would help protect the usefulness, security and fairness of test scores.

This idea of only using randomization to produce unlimited numbers of unique test forms is not new, first proposed by Frederic Lord of ETS in 1955, as he also noted its test security value. Lord also provided the theory and statistical tools to eventually make such testing a reality.

With computers today, unique test forms can be built and used. Also, enough items can be created so that each one may only be used a single time. These are once-and-done items, used on a single person's once-and-done test form, and never again. Test security is virtually assured, but are the test forms psychometrically solid?

This session will present Lord's ideas in more detail, show how today's technology is being used to create unique forms without duplicating items. The session will provide evidence of test quality in the form of case studies and simulations, among other benefits.

Standard Presentations V

Detecting a Change in Response Patterns in a Continuous Assessment Program

Aaron Myers, American Board of Internal Medicine | Derek Sauder, American Board of Internal Medicine

The medical certification industry has recently experienced a shift from point-in-time maintenance of certification assessments to a type of longitudinal assessment. These longitudinal assessments allow examinees to complete relatively few items at relatively frequent, small intervals (e.g., month) at a time of their convenience. Given the vast majority of previous test security research has focused on traditional point-in-time assessments, little research exists that focuses on the unique challenges associated with these longitudinal assessments. In contrast to point-in-time assessment, practitioners of longitudinal assessments may expect changes in estimated ability across items when items are administered essentially continuously (i.e., frequent, small intervals). Additionally, whereas point-in-time assessments generally consist of a relatively large number of items that can be evaluated for anomalous response patterns at once, it may not be practicable to aggregate items administered continuously because aberrancies may be confounded by changes in examinee ability. In this study, we propose using a change point detection model as a method of detecting anomalous behaviors in the context of continuous assessments. The change point model is able to identify whether an examinee is exhibiting anomalous response behavior and, if so, when an examinee begins exhibiting such behavior. We evaluate the performance of the change point model in a simulation study across different conditions that may impact its performance (e.g., test length, proportion of anomalous responses, location of change point). The change point model is then fit to data with known anomalous behavior. Implications and limitations of the change point model will be addressed.

Identifying Statistical Trends for Potentially Exposed Items

Richard Feinberg, NBME | Merve Sarac, University of Wisconsin-Madison | Chunyan Liu, NBME | Linette Ross, NBME | Justin Wolczynski, NBME

Testing organizations often investigate if secure test material has been exposed and consequently invalid for scoring and inclusion on future assessments. In the current study, we present an approach for longitudinally modeling both response accuracy and time-intensity to compare against items previously flagged as exposed from subject matter expert review. Preliminary results highlighted normatively extreme items that have substantially drifted from their initial deployment. Further, there did not appear to be a strong correspondence between the statistically unusual items and those flagged for exposure. Thus, decisions on the extent to which items are compromised may benefit from a more comprehensive approach utilizing both qualitative and quantitative methods to ensure that concerning items are discovered.

Test Security Challenges and Potential Solutions for Continuous Assessments

Derek Sauder, American Board of Internal Medicine | Aaron Myers, American Board of Internal Medicine

Many medical certification boards are beginning to offer some sort of continuous/longitudinal assessment to fulfill physicians' maintenance of certification (MOC) requirements. Continuous assessment programs benefit physicians by being more flexible and partially formative in nature, as compared to traditional, point-in-time MOC assessments. Most continuous assessment programs are able to be completed at home, away from a structured testing environment (e.g., testing center) and with no proctor. Additionally, many programs feature some sort of immediate feedback to inform physicians of whether they answered the item correctly or not. Given the end goal is to make a summative, pass/fail decision, such flexibility in the assessment program, albeit beneficial to learning, poses many additional test security challenges not faced by more traditional MOC exams. We will discuss some of the unique test security challenges associated with continuous assessment programs, identify guidance on how to approach test security in these programs, and provide an example from a medical certification board of some of the techniques used to help ensure a valid and reliable pass/fail decision for physicians.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

Conference On Test Security 2022

3:30 - 4:30 *CONTINUED...* Senior Room

When Star Trek meets Testing: How Technology is Changing Test Security

Craig Mills, Mills Consulting | Nikki Eatchel, Mursion | Kimberly Swygert, NBME | Ray Nicosia, ETS | Rachel Schoenig, Cornerstone Strategies

Panel Presentation

Technology continues to provide new opportunities for how we interact with one another, collect data, and make decisions. We're not testing with holograms (yet), but there are important technology advances on the horizon that hold significant promise for assessment and, in turn, can have a significant impact on exam security. From virtual reality and longitudinal assessment to machine learning, we will explore opportunities for testing programs to boldly go where no testing programs have gone before. Join a panel of experts as we discuss some of the newest areas being explored in testing and what that can mean for your testing program. We won't be able to "beam you up" so you'll need to join us the old-fashioned way for an engaging and thought-provoking look at our what our future may hold.

Improving Test Security for State Assessments in 2022: A New and Updated Guide for States

John Olson, Olson Educational Measurement and Assessment Services | Walt Drane, Caveon Test Security| Timothy Butcher, WV DOE | Lynn Schemel, IN DOE | Sarah Toton, Caveon

Coordinated Symposium

In recent years, many new things have been learned by states for improving test security, including the implementation of online assessments/CATs, monitoring test administrations more effectively, and attempting various methods for conducting secure testing during a pandemic. Among the latest challenges for states are the increasing use of nonpermitted technology and devices by students during testing, disclosure of test content on social media, and doing remote testing securely. In 2022, a team of state assessment representatives, independent test security experts, and Caveon staff collaborated to develop a new report, titled "Test Security Guide for State Assessments: Updated 2022" that addresses many of the latest issues and challenges. This document is a useful resource for improving security practices, preventing test irregularities, detecting test piracy, and investigating suspected instances of improper testing behavior by students or educators. In this session, information from the new report will be presented along with effective state strategies and current procedures for improving security. Two nationally-recognized experts on test security for state assessments and two state assessment directors will provide many examples of effective state practices, e.g., implementing stronger policies, procedures for conducting remote testing, better monitoring of the Internet and social media, expanded use of data forensics analyses, and using innovative test designs to increase security. Copies of the new Caveon report will be shared with all attendees.

THURSDAY, OCT 27
7:30 A.M. – 6:30 P.M. EDT

5:00 - 6:30

Prince William Ballroom

COCKTAILS & CONVERSATION

Light hors d'oeuvres will be served

Poster Presentations

Item Preknowledge Detection in CAT with the Impact of Preknowledge Varying across Persons and/or Items
Jianshen Chen, College Board | Kylie Gorney, University of Wisconsin-Madison | Luz Bay, College Board

Use of Statistical Evidence to Flag Potential Compromised Items
Cecelia Alves, Maxim Morin, Qi Guo, Medical Council of Canada

Improved Detection of Abnormal Testing Behavior through More Complete Modeling of Response Time Profiles
Regi Mucino, PSI Services LCC | Gregory M. Hurtz, California State University, Sacramento and PSI Services LLC

Examining Preknowledge Detection Statistics based on Likelihood Ratio Test for Use in Real-Time
Merve Sarac & James Wollack, University of Wisconsin-Madison

Exploring Item Response Trajectory Similarity
Hongling Wang & Chi-Yu Huang, ACT Inc.

Creating an Automated Proctoring Calibration Tool to Enable Decision Agreement Transparency & Ensure Quality Control
Rose Hastings Keith & Will Belzak, Duolingo

Effects of Multidimensionality and Difficulty on the Modified Caution and U3 Person-Fit Indices: New Insights and Modified Indices
Steven Svoboda, PSI Services LCC | Gregory M. Hurtz, California State University, Sacramento and PSI Services LLC

Detecting Item Pre-knowledge on Re-used Forms
Aijun Wang & Yu Zhang, The Federation of State Boards of Physical Therapy

Improving Proctoring through National Certification: An Initiative of the National College Testing Association
Rachel Hample, National College Testing Association

FRIDAY, OCTOBER 28
7:30 A.M. – 1:00 P.M. EDT

7:30 - 8:15

Prince William Ballroom

Breakfast: A buffet is available in the Senior Room. Seating is in the Prince William Ballroom.

8:15 - 9:15

Palmer Room

When Insiders Go Rogue

Jennifer Semko, Baker McKenzie | Ray Nicosia, ETS | Rachel Schoenig, Cornerstone Strategies
Panel Presentation

Attempts by third parties to circumvent testing rules, steal content, or provide proxy test takers are well-known. As testing professionals, we spend significant resources attempting to deter and detect these types of activities to protect our testing programs. But what happens when the threat isn't coming from outside of our organization but is instead a trusted insider? The reality is, insider threats are a significant risk to testing organizations, and when insiders go rogue, the impact can be devastating. Join test security experts to take a look at what can happen when trusted insiders engage in negligent or intentional behaviors that put your testing program at risk. We will share case studies, discuss what you can to do protect your organization, and how you can respond if an incident arises.

Prince William Ballroom

Balancing Test Security and Accessibility - Ensuring the Validity of State Assessment Results for Students with Disabilities

John Olson, OEMAS | Sheryl Lazarus, National Center on Educational Outcomes | Mari Quanbeck, National Center on Educational Outcomes | Jessica Fenby, Michigan DOE | Walt Drane, Caveon Test Security
Panel Presentation

There is a need to balance test security with the accessibility of assessments for students with disabilities. The policies and procedures that are in place by states to minimize test security risks may at times affect accessibility. As such, there is a need to consider how different test security measures may affect or interfere with the accessibility of state assessments for students with disabilities. This panel session will provide an overview of the issues as well as share the findings of a recent analysis of states' test security policies and practices that examined how accommodated tests, alternate assessments, and related issues for students with disabilities were addressed in the policies. A state test security director will share state perspectives and describe how that state is including students with disabilities in their test security policies. The session will include suggestions and considerations for how to maintain test security while ensuring the accessibility of assessments. We will share what we know, what we think, and what we still just cannot figure out. Please join us in a lively discussion as we explore how test security and accessibility can be better balanced!

FRIDAY, OCTOBER 28
7:30 A.M. – 1:00 P.M. EDT

8:15 - 9:15 *CONTINUED...* *Nassau Room*

Standard Presentations VI

Secret Shopping: Exposing and Addressing the Inner Workings of Cheating Sites

Carissa Pittsenberger, WGU | Lestelle Schwab, WGU

The availability of websites offering to complete individual tests and even full courses, or degrees is on the rise. As most current processes include mainly reactive approaches to addressing the bad actors, we decided to explore a proactive solution.
Academic integrity is critical to relevant and valid assessment design and the subsequent delivery. As part of Western Governors University's (WGU) efforts to improve assessment delivery security, a secret shopping approach has been taken to seek out and identify the methods for proxy test-taking sites in order to stop the potential ability of these predatory sites to subvert the academic process. In addition, a focus on maintaining our integrity is as important as exposing the lack of integrity seen in these cases. The process used will be outlined and the results for a specific site will be discussed. Come find out what we have learned!

FRIDAY, OCTOBER 28
7:30 A.M. – 1:00 P.M. EDT

8:15 - 9:15 *CONTINUED...* *Albert Einstein Room*

Standard Presentations VII

Hybrid Threshold-based Sequential Procedures for Detecting Compromised Items in a CAT Licensure Exam

Chansoon Lee, ABIM | Hong Qian, NCSBN

Using classical test theory and item response theory, this study applied sequential procedures to a real operational item pool in a variable-length CAT to detect items whose security may be compromised. Moreover, this study proposed a hybrid threshold approach to improve the detection power of the sequential procedure while controlling the Type I error rate. The hybrid threshold approach uses a local threshold for each item in an early stage of the CAT administration, and then it uses the global threshold in the decision-making stage. Applying various simulation factors, a series of simulation studies examined which factors contribute significantly to the power rate and lag time of the procedure. In addition to the simulation study, a case study investigated whether the procedures are applicable to the real item pool administered in CAT and can identify potentially compromised items in the pool. This research found that the increment of probability of a correct answer (p-increment) was the simulation factor most important to the sequential procedures' ability to detect compromised items. This study also found that the local threshold approach improved power rates and shortened lag times when the p-increment was small. The findings of this study could help practitioners implement the sequential procedures using the hybrid threshold approach in real-time CAT administration.

Experimental Comparison of the Security Value of Randomly Parallel Tests and Strictly Parallel Tests

David Foster, Caveon Test Security

Creating a unique test form for each test taker is an important security measure for today's high-stakes exams. If every test is unique, then sharing the content of any one or all will not give any other test takers an advantage. Theft of test content will be a useless exercise, and most forms of cheating, will go away. A unique but equivalent test form for every test taker was Frederic Lord's vision. He called these tests Randomly Parallel Tests (RPTs). Because they were tests that were difficult to use in either research or practice at the time, very little empirical data exists on the design. Certainly, there are no comparative studies of RPTs versus "strictly parallel" or "rationally parallel" tests, the test design in predominant use then as it is today.

This session will present the results of empirical research directly comparing RPTs and test forms created more traditional ways. The test forms, whether RPTs or those created traditionally, will be provided to samples of test takers, balanced by ability in the domain. Performance on the test, that is, test scores, will be compared, along with other psychometric and security measures.

FRIDAY, OCTOBER 28
7:30 A.M. – 1:00 P.M. EDT

Conference On Test Security 2022

9:30 - 10:45

Prince William Ballroom

CLOSING KEYNOTE DEBATES

It's gloves off once again for the COTS Keynote Debates! Join industry experts as they debate the impact of rapid item development on item theft and whether high stakes summative testing (and its attendant exam security needs) will be irrelevant by 2032. Bring your cell phones, as you will be called on to provide additional context during these fast paced and engaging debates. Learn from the experts, shape and inform your own opinion, and gain from the wisdom of the crowd as we explore these important topics in assessment. It's a throw-down you won't want to miss!

Debaters: Russell Smith, James Wollack, David Foster, and Carol Eckerly

FRIDAY, OCTOBER 28
7:30 A.M. – 1:00 P.M. EDT

11:00 - 12:00

Palmer Room

Detection of Cheating in Writing in the Age of Advanced Technology

Duanli Yan, ETS | Jakub Novak, ETS | Ikkyu Choi, ETS | Jiangang Hao, ETS | Chen Li, ETS | Michael Fauss, ETS | Wenju Cui, ETS | Mo Zhang, ETS | Paul Deane, ETS

Coordinated Symposium

With the advances in technology, the environments for learning and assessment have improved tremendously both in classroom and at large-scale assessments. These enabled learners and test takers to improve their knowledge and skills in their subjects, but also provided opportunities for gaming the assessments. In this coordinated session, we gathered three presentations to show several approaches on how to detect potential cheating in writing in the current age of advanced technology. The three presentations include automated essay detection solution, detection of AI-generated essays, and keystrokes in detection. The goal is for researchers to think and develop methodologies to address these issues to ensure the quality of writing assessments.

Nassau Room

The Evolving Legal Landscape of Test Security

Jennifer Semko, Baker McKenzie | Donna McPartland, Han Santos | Camille Thompson, The College Board | Rachel Schoenig, Cornerstone Strategies

Panel Presentation

Legal guardrails underscore our relationships with test takers, vendors and third parties, and impact how we protect our content, implement security tools, and respond to security incidents. As laws around data privacy, artificial intelligence, biometrics, accessibility, and copyright protection continue to evolve, so too must our test security framework. How will these changes impact testing programs and the tools used to secure tests? What can testing programs do to ensure they are prepared to address and respond to legal and regulatory changes? Join experienced legal professionals as they provide insight into the evolving legal landscape around testing and practical tips on preparing your testing program to evolve as well.

FRIDAY, OCTOBER 28
7:30 A.M. – 1:00 P.M. EDT

11:00 - 12:00 *CONTINUED...* *Albert Einstein Room*

Standard Presentations VIII

Exam Performance Audits: Identifying and Resolving Issues With Your Exams Before They Become Emergencies

Stephanie Cobty, Caveon Test Security | Susan Weaver, Caveon Test Security

Exam Performance Audits are critical for low and high-stake certification programs to collect and analyze actionable data about the exam's performance. The test development process is rigorous and costly for the organization. Actions should be taken on active and future examinations to protect that investment. Exam Performance Audits present an opportunity to identify potential issues using classical statistics (including point biserial and p-value) and other data. The audit then digs deeper to find and propose actionable solutions to those concerns. Many issues can be identified with as few as 40 exam takes, though the more data available the more thorough the results. This presentation will address the issues detected in an audit, the benefits to identifying issues early on, and provide general guidelines for analyzing this data.

The Formation of ANCC's Test Security Task Force and Its Role in High-Stakes Test Security

Tracy Getselman, American Nurses Association | Shawn Amaker, American Nurses Association | Caitlin Gdowski , American Nurses Association

Within the American Nurses Credentialing Center (ANCC), the Measurement Services department develops and validates 17 certification exams. ANCC's Test Security Task Force (TSTF) draws membership from each of Measurement Services four workgroups to form a multidisciplinary team that addresses potential threats to program and test development security. The TSTF has developed foundation documents that guide our work as we respond to incidents, such as potentially compromised items and falsely copyrighted materials. Our presentation will describe the formation of the task force, the roles and responsibilities of task force members, and an overview of our guiding documents and procedures. In addition, we will provide insights into our approach to resolve hypothetical threats to our program's security.

12:00 - 1:00 *Prince William Ballroom*

Lunch: Boxed lunches are available in the Senior Room. Seating is in the Prince William Ballroom.

SAVE THE DATE

THE CONFERENCE ON TEST SECURITY

FALL 2023

TEMPE, AZ

HOSTED BY
ARIZONA STATE UNIVERSITY