Conference On Test Security 2017

September 6 – 8, 2017

Madison, WI

Conference Welcome

It is my extreme pleasure to welcome you to the 2017 Conference on Test Security (COTS). It is hard to believe that COTS is now in its sixth year. Since its inception, COTS has provided a venue for 273 different presentations, having grown from 19 presentations in 2012 to 75 this year, and over 400 conference participants.

When the conference was initially conceptualized by Neal Kingston at the University of Kansas, it was branded as the Conference on the Statistical Detection of Potential Test Fraud, and the goal was to provide a forum for fostering research into developing and improving statistical tools to identify cheating on tests. Unquestionably, the original intention of this conference has been realized. The presentations from the first two conferences provided the foundation for two new edited volumes dedicated entirely to test security methodology—*Test Fraud: Statistical Detection and Methodology* (Kingston & Clark, 2014) and the *Handbook of Quantitative Methods for Detecting Cheating on Tests* (Cizek & Wollack, 2017). In addition, the last six years has seen a dramatic increase in the number of cheating detection methodology publications in peer reviewed journals.

In 2014, the scope of the conference broadened to focus on all test security capabilities and enhancements that protect the validity of test results and brand integrity, and to encourage and foster dialogue between the different sectors of the test security community. This broadening of focus has coincided with a greater recognition among the testing community in the importance of test security. The most recent edition of the joint *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) places a greater emphasis on the impact of cheating on tests than did previous editions, and makes clear that test developers and providers have a responsibility in the name of fairness to take measures to maintain the security of exams.

One thing that has not changed since this conference began is that cheating on tests continues to persist and remains a significant threat to the validity of test score interpretations. This year's conference provides a host of technological, methodological, and organizational ideas for improving our ability to prevent, deter, impede, detect, investigate, and respond to cheating. It is my sincere hope that the 2017 Conference on Test Security provides you with an opportunity to learn about the newest strategies in test security, to expand your network of test security professionals, and to identify several areas in which you might be able to modify your operational practices to improve the security of your assessments.

The University of Wisconsin-Madison was the host for the 2013 conference, and is honored and excited to host the conference yet again. UW-Madison has long been a leader in the assessment world, housing the Quantitative Methods graduate program within the #1 ranked Educational Psychology Department in the country. In addition, the UW Center for Placement Testing and the WIDA Consortium, both national leaders in the development and delivery of assessments for college-level course placement and K-12 English Language Learners, respectively, also reside on the UW campus.

James Wollack
Professor, Educational Psychology Department
Director, UW Center for Placement Testing
Director, Office of Testing & Evaluation Services

Schedule At-A-Glance

Wednesday, September 6	
11:00 – 5:00	Registration
12:00 – 1:30	Workshop 1
1:45 – 3:15	Workshop 2
3:30 – 5:00	Workshop 3
5:15 – 6:45	Executive Committee Meeting
7:30 – 10:00	*BAD GENIUS* free screening
Thursday, September 7	
7:00 – 7:45	Buffet Breakfast
7:30 – 12:00	Registration
8:00 – 9:30	Opening Keynote Speaker
9:45 – 10:45	Session 1
11:00 – 12:00	Session 2
12:00 – 1:00	Lunch
1:00 – 2:30	Session 3
2:30 – 3:00	Refreshment Break
3:00 – 4:00	Session 4
4:15 – 5:15	Session 5
5:15 – 5:30	Poster Set Up
5:30 – 7:30	Poster Presentations with Networking Reception
Friday, September 8	
7:00 – 7:45	Buffet Breakfast
8:00 – 9:30	Session 6
9:45 – 11:00	Closing Keynote Speaker
11:15 – 12:15	Session 7
12:15 – 1:15	Lunch
1:15 – 2:15	Session 8

General Conference Information

Wi-Fi

Wi-fi is available for conference attendees throughout the entire Pyle Center. Please select the UWNET network, click on "Guest Access," complete the required fields and click register. After confirming the information, you will be able to log in. You may register multiple devices.

Conference URL

https://cete.ku.edu/2017-conference-test-security

Conference App

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

Social Media

COTS is excited to use social media to help connect conference attendees both during the conference and throughout the year. We invite you to join and participate in the LinkedIn discussion group at https://www.linkedin.com/groups/13542712. Also, please follow our Twitter handle @COTS_2017 and post about your conference experience on Twitter using the hashtag #COTS2017.

Conference Layout and Meeting Space

All conference sessions will be held in the Pyle Center (see maps on pages 60-62). Breakout Sessions will all be in rooms on the 2nd and 3rd floors. The opening keynote, both lunches, and the Thursday evening poster reception will be held in the Alumni Lounge on the 1st floor. Breakfast will be held in the Lowell Center Dining Room. The screening of *BAD GENIUS* will occur in the Marquee Cinema at Union South.

Information for Speakers

All presentation rooms are equipped with a PC laptop computer, data projector and screen, audio hookups, microphone, and a podium. VGA cables are provided, in case you wish to present from your personal laptop computer. Adaptors are available through the conference IT staff, as necessary. A Speaker Lounge has been set up in room 317 so that you can make sure your presentation works with the equipment being used in the presentation rooms. If you experience any problems with the technology or room accommodations during the conference, each presentation room is equipped with a telephone that you can use to connect with IT Services or the Pyle Center front desk. Those participating in the poster session should arrive at the room between 5:15 – 5:30 to set up their posters.

Food Allergies and Dietary Restrictions

All food allergies and dietary restrictions identified during the registration process have been communicated to the catering staff. By offering buffets for breakfasts and lunches, we have been informed that many of these dietary allergies and restrictions will not present issues. Individuals for whom the standard option will not suffice have been identified by the catering staff and should have received a special card upon check-in indicating which meals will require special accommodation. In these cases, the individuals are asked to simply identify yourself to the catering staff. They are expecting you and are preparing separate meals. To facilitate attendees making food selections, all buffet items will include information on the ingredients used in preparation; however, if at any point you should have a question about food selection, please speak with a member of the catering staff.

BAD GENIUS

The screening of *BAD GENIUS* will begin at 7:30. The movie is 2 hours and 10 minutes long, so should end around 9:45. Guests of conference attendees are welcome to attend.

The Marquee Cinema is located on the second floor of Union South, which is three-quarters of a mile from the Pyle Center. Two different groups will travel together to the theater. A walking group will leave from the Langdon Street entrance to the Pyle Center at 6:50 to take the 15 minute walk through the heart of the UW campus. A bus will also be available for those wishing to ride to the theater. The bus will leave from Langdon Street (in front of the Pyle Center) at 7:00. The bus will also be available after the movie to return attendees to the Pyle Center. If you would like to walk yourself to the movie, please pick up a copy of the Walking Directions to Marquee Cinema from the Registration Table.

Popcorn and soda/water will be provided. For those interested in purchasing other food items or alcoholic beverages to bring into the cinema (yes, it’s Wisconsin, so beer and wine are okay in the cinema), there are several restaurants and a small market on the first floor of Union South.

Madison

Madison is best known for being the capital of Wisconsin and home of the University of Wisconsin, one of the most illustrious and exceptional public universities in the world. However, Madison is also one of America’s most livable cities, featuring numerous outdoor activities, tremendous culture, and, of course, the best weather in the country! Within one mile of the conference venue, you may head southwest for a stroll through the heart of the campus en route to Camp Randall Football stadium or due east down State Street, where you will find dozens of restaurants, shops, bars, coffee houses, and clubs, en route to the state capitol building. If you are staying in town through Saturday morning, Madison features the largest producer-only Farmer’s Market in the country. The Market is open on Saturdays from 6:15 a.m. to 1:45 p.m, and takes place on the four streets that encircle the state capitol. Just a few steps west of the Pyle Center, you will find the Wisconsin Memorial Union, home of the renowned Union Terrace, with its breathtaking views of Lake Mendota and celebrated Babcock ice cream. If you are looking for a longer walk/run, feel free to head two miles due west along the waterfront to Picnic Point for 360 degree views of the lake and downtown Madison.

Local Contacts

Whatever it is that suits your fancy, there’s a good chance that Madison has it. Because it is impossible to list every event, restaurant, and attraction, those interested in directions, food recommendations, or more information about local amenities are encouraged to seek out one of the individual attendees from the Madison area. To facilitate identifying those individuals with inside knowledge about the city, look for attendees wearing one of the yellow “Local Contact” ribbons on their name badge.

2017 Conference Sponsors

Co-Hosts

ACT®

Alpine
Testing Solutions

caveon™
Test Security

cisco™

Pearson
VUE

2017 Conference Sponsors

Friends

AICPA®

ascend LEARNING

Castle
Intelligent solutions for quality testing

CollegeBoard

ETS®

psi®

WESTERN GOVERNORS UNIVERSITY
WGU
wgu.edu®

Testing Organization Sponsor

National College Testing Association
NCTA

ATP
ATPSC
Security Committee

WEDNESDAY, SEPTEMBER 6

11:00 – 5:00 **Conference Registration** **Pyle 3rd Floor**

12:00 – 1:30 **Workshop 1** **Pyle 325-326**

Implementing DOMC: Fast and Easy

David Foster, Caveon Test Security

1:45 – 3:15 **Workshop 2** **Pyle 325-326**

Planning and Responding to Test Security Incidents

Rachel Schoenig, Cornerstone Strategies, LLC
Mike Clifton, ACT, Inc.
Nick Charge, Cambridge English Language Assessment
Ray Nicosia, ETS
Bryan Freiss, Pearson VUE

3:30 – 5:00 **Workshop 3** **Pyle 325-326**

Using the New Credentialing Security Framework

Rachel Schoenig, Cornerstone Strategies, LLC
Jennifer Geraets, ACT, Inc.
Jamie Mulkey, Caveon Test Security

7:30 – 10:00 ***BAD GENIUS* Free Film Screening** **Marquee Cinema, Union South**

BAD GENIUS is drawing rave reviews on the international film festival circuit, and recently opened the Asian Film Festival (AFF) in New York. Given the commonality between the movie's plot line and the professional interests of attendees, the film's international distributor, GDH 559 Co. LTD, has graciously agreed to allow the Conference on Test Security to host a private, free screening of this film.

> Lynn is a scholarship kid and as smart as they come, but that doesn't mean she's immune from making some dumb decisions. When her best friend, Grace, asks for some help on an exam the dishonest act of kindness catches the eye of other students in need. Soon Lynn is cheating for an increasing number of wealthy classmates — in exchange for some substantial cash payments — and she justifies it as a way of helping her working-class dad. Their schemes grow in complexity and culpability leading up to an exam with international ramifications.
>
> ***BAD GENIUS*** is a sharply directed and edited look at the high cost of cheating, and while its action is far removed from life and death situations the film delivers some tense and highly suspenseful sequences. More than that though, the film finds the heart in young Lynn's situation and moves deftly between the expected narrative beats to leave viewers truly caring about her fate. (New York Asian Film Festival)

Getting to the Marquee Cinema

If you wish to walk as a group to the Cinema, meet in front of Pyle Center (on Langdon Street) at 6:50. A shuttle will also leave from the front of Pyle Center at 7:00 to drive people to the Cinema. The shuttle will be available after the movie to return attendees to the Pyle Center.

Thursday, September 7

7:00 – 7:45 Buffet Breakfast — Lowell Dining Center

7:30 – 12:00 Conference Registration — Pyle 3rd Floor

8:00 – 9:30 Opening Keynote Speaker — Pyle Alumni Lounge

Who Do You Trust? The Epidemic of Synthetic Identities

Paul Bjerke, LexisNexis Risk Solutions

What can we do when the very materials we rely upon for identification are being undermined? Synthetic identities are just one new method being used to perpetrate fraud and undermine trust in identification capabilities we take for granted. Today, synthetic identities are being used by individuals and criminal gangs to defraud the financial industry. Join Paul Bjerke, Vice President of Fraud and Identity Strategy at LexisNexis, as he shares how synthetic identities are created, why it's important for test publishers to understand, and what we can do to mitigate the risk.

Paul Bjerke leads the Fraud and Identity Management strategy for LexisNexis Risk Solutions. He sets the longer-term fraud solution framework, leveraging the company's vast data repository, analytic linking capabilities and proprietary high-powered technology, HPCC Systems. These innovations help clients optimize critical new account origination and account management activities that appropriately balance their risk/reward equation and support financial inclusion and financial transparency.

Mr. Bjerke has a passion for fighting fraud and has over 20 years of experience in retail banking risk, analytics, operations, and product and payment management. Prior to joining LexisNexis Risk Solutions in 2016, he was a fraud and risk product leader at Deluxe Corporation; provided fraud prevention and AML consulting solutions at IBM; led the credit card, gift card, debit card, e-commerce fraud and AML compliance policy at Target Corp; and was Vice President of Product at FIS ChexSystems. He has also held numerous roles within payments risk and retail banking at U.S. Bank and Wells Fargo and has been a member of the International Association of Financial Crimes Investigators since 1995.

Mr. Bjerke graduated with an MBA, with a Finance Concentration, from the University of Minnesota, and has a B.S. in Business Administration from North Dakota State University. He also has a diploma in commercial lending and general banking from the American Institute of Banking.

9:45 – 10:45 Session 1

RS1: MULTIPLE MEASURES APPROACHES TO PREDICTING AND DETECTING CHEATING **Pyle 309**

Forensic Profiling of Test-Taker Response Patterns Associated with Distinct Cheating-Related Behaviors

Greg Hurtz, PSI Services LLC
John Weiner, PSI Services LLC

Predicting Cheating Before It Happens

Elizabeth Amador, Western Governors University

Test-Cheating Risk Prevention: Developing Predictive Model Based on the Result of Cheating Detecting System (CANCELED)

Xuan Zhou, Beijing Language and Culture University
Xiang Kong, Beijing Language and Culture University

DP1: *Identifying and Implementing Test Security Enhancements for Your Testing Program* **Pyle 209**

Jennifer Geraets, ACT, Inc.
Emily Scott, ACT, Inc.

PP1: *Improving and Streamlining Proctor Training Through National Proctor Certification: An Initiative of the National College Testing Association* **Pyle 226**

James Wollack, University of Wisconsin – Madison
Rachel Hample, Temple University
Jarret Dyer, College of DuPage

PP2: *This IS a Drill! Using Tabletop Exercises to Plan for Test Security Disasters* **Pyle 325-326**

Jamie Mulkey, Caveon Test Security
Tara Miller, Amazon AWS

Program Legend of Session Types

RS = Research Session
DP = Demonstration Presentation
PP = Panel Presentation
FR = Facilitated Roundtable

11:00- 12:00 Session 2

DP2: ***The New SETI Project: Tools to Help the Search for Examination Treachery and Iniquity*** **Pyle 209**

Jennifer Davis, National Association of Boards of Pharmacy
Nathan Thompson, Assessment Systems Corporation

PP3: ***We Can Do Better: Security of Test Exams and Identifying Events in Near Real-Time*** **Pyle 325-326**

Michelle Barrett, Pacific Metrics
Walt Drane, Mississippi Department of Education
Michael Clifton, ACT, Inc.
Wes LaMarche, IMS Calipers Analytic Standard Working Group

PP4: ***Designing, Developing, and Implementing Security Policies and Practices for Formative Assessment Products*** **Pyle 226**

Steve Ferrara, Measured Progress
Steve Addicott, Caveon Test Security

PP5: ***Cheaters Say the Darnedest Things!*** **Pyle 309**

John Fremer, Caveon Test Security
Jarret Dyer, College of DuPage
James Wollack, University of Wisconsin – Madison
Ben Fortney, University of Wisconsin – Madison

ACT–
Leading the way in research and test development
ACT
MS3356
©2017 by ACT, Inc. All rights reserved.

12:00 – 1:00 **Lunch** **Pyle Alumni Lounge**

Presentation: An Update on Registering Secure Tests with the US Copyright Office

Jennifer Ancona Semko, Baker & McKenzie

1:00 – 2:30 **Session 3**

RS2: Research Session: Cheating in Practice **Pyle 226**

Online Proctoring – Best Practices

Adel Lelo, Western Governors University

Two Decades of Investigative Cheating Detection Research

Ardeshir Geranpayeh, Cambridge English Language Assessment

Investigating Multi-Year Cheating on State Assessments: A Case Study

David Ragsdale, Massachusetts Department of Elementary and Secondary Education

Cisco v. TestKing, Pass4sure & Test-Inside: Results and Underlying Strategies from a Recent Exam Piracy Case

Gerald Pia, Roche Pia LLC
Victoria Quinn-Stephens, Cisco Systems, Inc.

RS3: Response-Time Methods **Pyle 309**

Some Graphical Techniques for Presenting Aberrant Response and Timing Data Analysis Results

Richard Luecht, University of North Carolina at Greensboro
Terry Ackerman, ACT, Inc.

A Comparison of Pre-Knowledge Detection Procedures in CAT with Response Time Modeling

Jin Zhang, ACT, Inc.

A Mixture Model to Detect Item Preknowledge Using Item Responses and Response Times

Seo Young Lee, University of Wisconsin – Madison
James Wollack, University of Wisconsin – Madison

Is Time on Our Side? An Item-Level Latency Analysis to Detect Pre-Knowledge

Sarah Thomas, Caveon Test Security

PP6: ***Principles and Practices for Presenting Findings of Test Security Violations in Legal Settings*** **Pyle 325-326**

Dennis Maynes, Caveon Test Security
Rachel Schoenig, Cornerstone Strategies, LLC
Marc Weinstein, Caveon Test Security
William P. Skorupski, University of Kansas
Camille Thompson, ACT, Inc.

PP7: ***Peer Review Requirements for States on Test Security and Monitoring: What was Learned from the Reviews of Recent Submissions to USED*** **Pyle 209**

John Olson, Caveon Test Security
John Fremer, Caveon Test Security
Kathy Moore, Kentucky Department of Education
Peter Zutz, Nevada Department of Education
Timothy Butcher, West Virginia Department of Education

2:30 – 3:00 **Refreshment Break** **Pyle 3rd Floor**

3:00 – 4:00 **Session 4**

RS4: SECURE TEST DEVELOPMENT **Pyle 209**

Building a Secure Online Testing Platform in the Cloud
Jim Sherlock, Pearson

Making Sense of the Science of DOMC
David Foster, Caveon Test Security

Analysis of the Discrete Option Multiple Choice Item: Examples from IT Certification
Carol Eckerly, Alpine Testing Solutions
Russell Smith, Alpine Testing Solutions
John Sowles, Ericsson

DP3: *Copyright and Trade Secret Protection for Standardized Tests* **Pyle 309**
Emily Scott, ACT, Inc.
Tim Conlon, ACT, Inc.

PP8: *Testing in Hell: What to Do When Everyone Seems to be Cheating* **Pyle 325-326**
Steve Addicott, Caveon Test Security
Beverly Bone, IBM Certification Programs

Protecting Exams Across the Galaxy...
caveon™
Test Security
Consulting • Data Forensics • Investigative Services
Secure Exam Development • Technology • Web Patrol
caveon.com

4:15 – 5:15 Session 5

RS5: Collusion Detection **Pyle 209**

Ockham's Razor and the Selection of Collusion Indices: Variants of J2 Provide a Simple and Effective Diagnostic Tool
Greg Hurtz, PSI Services LLC
John Weiner, PSI Services LLC

Detection of Potential Test Collusion across Multiple Examinees: A Real-World Example
Mengyao Zhang, National Conference of Bar Examiners
Joanne Kane, National Conference of Bar Examiners

Effects of Local Testing Volume on Test-Center-Based Collusion Detection
Anne Thissen-Roe, PSI Services LLC

PP9: *Champagne Test Security on a Beer Budget--How Smaller Organizations Can Develop and Maintain a Holistic Test Security Program Despite Limited Staffing and Budgets* **Pyle 226**
Marc Weinstein, Caveon Test Security
Thomas Gera, The Enrollment Management Association

PP10: *Legal Tips for Responding to a Test Security Incident* **Pyle 325-326**
Rachel Schoenig, Cornerstone Strategies, LLC
Jennifer Ancona Semko, Baker & McKenzie
Camille Thompson, ACT, Inc.

PP11: *When Test Items Fight Back* **Pyle 309**
Susan Weaver, Caveon Test Security
John Sowles, Ericsson
Diane Long, OKTA, Inc.

5:15 – 5:30 Poster Setup Pyle Alumni Lounge

5:30 – 7:30 Poster Presentations and Networking Reception Pyle Alumni Lounge

Poster 1: ***It's a Bird; It's a Spider: No, It's the Owlbot!***
Carissa Pittsenberger, Western Governors University

Poster 2: ***Using the ω Statistic to Estimate an Unknown Flawed Answer Key***
Marcus Scott, Caveon Test Security

Poster 3: ***Detecting Compromised Items in CAT Using a Sequential Monitoring Procedure***
NooRee Huh, ACT, Inc.
Qing Xie, University of Iowa/ACT, Inc.
Chunyan Liu, ACT, Inc.
Chi-Yu Huang, ACT, Inc.

Poster 4: ***Test Security Ecosystem***
Michael Clifton, ACT, Inc.

Poster 5: ***A Study of Students' Item Review Behaviors in Computer-Based Testing***
Hongling Wang, ACT, Inc.
Chi-Yu Hunag, ACT, Inc.

Poster 6: ***A Hierarchical IRT Model for Identifying Group-Level Aberrant Growth***
Jennifer A. Brussow, University of Kansas
William P. Skorupski, University of Kansas
W. Jake Thompson, University of Kansas

Poster 7: Using Candidate Clusters to Identify Potentially Compromised Items
Yu Zhang, Federation of State Boards of Physical Therapy
Jiyoon Park, Federation of State Boards of Physical Therapy
Aijun Wang, Federation of State Boards of Physical Therapy
Lorin Mueller, Federation of State Boards of Physical Therapy

Poster 8: Enhanced Assessment Monitoring – Leveraging Technology to Streamline Monitoring Processes, Manage Data Flow, and Report Useful Results in Real Time
Marc Weinstein, Caveon Test Security
Benjamin Hunter, Caveon Test Security

Poster 9: Considerations for Detecting Test Misconduct in Real Time
Anna Topczewski, GED Testing Service

Poster 10: Graphical Imaging Methods for Detecting Potential Collusion for Test Centers with Unusual Score Gains
Mengyao Zhang, National Conference of Bar Examiners
Mark Albanese, National Conference of Bar Examiners

Poster 11: A Comprehensive Test Security Program for States
Walt Drane, Mississippi Department of Education
Sally Valenzuela, Caveon Test Security

Poster 12: Can You Outsmart a SmartItem™?

David Foster, Caveon Test Security
Jamie Mulkey, Caveon Test Security

Poster 13: Performance Comparison of GBT Based on Purified Real Data

Sakine Gocer Sahin, University of Wisconsin – Madison
James Wollack, University of Wisconsin – Madison
Selahattin Gelbal, Hacettepe University

Poster 14: An Item Selection Design that Optimizes Item Bank Usage and Estimation

Jing Yang, Northeast Normal University
Liwen Huang, University of Illinois at Urbana-Champaign
Leanne Zeng, University of Illinois at Urbana-Champaign
Hua-Hua Chang, University of Illinois at Urbana-Champaign

Poster 15: Empirical Study for Item Bank Replenishment of Computerized Adaptive Testing

Tong Wu, University of Illinois at Urbana-Champaign
Anqi Li, University of Illinois at Urbana-Champaign
Hua-Hua Chang, University of Illinois at Urbana-Champaign

FRIDAY, SEPTEMBER 8

7:00 – 8:00 **Buffet Breakfast** **Lowell Dining Center**

8:00 – 9:30 **Session 6**

RS6: ABERRANCE AND RESPONSE VALIDITY **Pyle 225**

Using Change Point Analysis to Detect Inattentiveness in Polytomous Survey Response Data
Xiaofeng Yu, University of Notre Dame
Ying (Alison) Cheng, University of Notre Dame

Robust Bayesian Estimation of Item Response Model Parameters Accounting for Aberrance
Kaiwen Man, University of Maryland College Park
Hong Jiao, University of Maryland College Park
Jeffery R. Harring, University of Maryland College Park

Data Quality in Assessment
Maxwell Hong, University of Notre Dame
Ying (Alison) Cheng, University of Notre Dame

Detecting Examinees with Aberrant Answer Changes in CBT Via Posterior Shift
Dmitry Belov, Law School Admission Council
Stephen Cubbellotti, American Board of Internal Medicine

RS7: DETECTION OF ANSWER COPYING/SIMILARITY **Pyle 209**

Visualizing Test Fraud Using Multiple Correspondence Analysis
Joe Grochowalski, The College Board

Detecting Answer Similarity Using Nonparametric Item Response Models
Xi Wang, Measured Progress
Wonsuk Kim, Measured Progress
Louis Roussos, Measured Progress

Detection of Answer Copying in China's College Entrance Examination via Kullback-Leibler Divergence and ω-Index
Yiqin Pan, Beijing Normal University / University of Wisconsin – Madison
Fang Luo, Beijing Normal University

Investigating the Performance of ω Index in Detecting Answer Copying
Önder Sünbül, Mersin University
Seha Yormaz, Mersin University

FR1: *Standards in Online Proctoring* **Pyle 226**

Adel Lelo, Western Governors University
Carissa Pittsenberger, Western Governors University

DP4: *Educator Coaching and Student Response Interference in K-12 Assessment Administrations--What It Looks Like and How to Detect and Stop It* **Pyle 309**

Marc Weinstein, Caveon Test Security
Walt Drane, Mississippi Department of Education

PP12: *State Strategies to Comprehensively Address Test Security Issues to Improve their Assessment Programs* **Pyle 325-326**

John Olson, Caveon Test Security
John Fremer, Caveon Test Security
David Ragsdale, Massachusetts Department of Elementary and Secondary Education
Elaine Themm, Michigan Department of Education

9:45 – 11:00 Closing Keynote Panel Pyle 325-326

DEBATING "APPROPRIATE" EXAM SECURITY

Michael Clifton, ACT
John Fremer, Caveon Test Security
Rory McCorkle, PSI Services LLC
Ray Nicosia, ETS
Jennifer Ancona Semko, Baker & McKenzie
William P. Skorupski, University of Kansas
Moderator: Rachel Schoenig, Cornerstone Strategies, LLC

As the assessment industry has matured, so has the debate concerning test security. Today, it is generally recognized that public trust in assessment results is dependent on more than psychometrically sound exams. Without appropriate exam security, the trust in exam results and reputation of our programs - and industry - is quickly eroded.

However, while the debate concerning the importance of exam security has been settled, the debate concerning what constitutes "appropriate" exam security remains on-going. Questions such as whether to cancel scores based solely on statistical improbability or whether good online proctoring is better than good in-person proctoring continue to frame the debate around what constitutes "appropriate" exam security.

In this fast-paced keynote, experienced professionals will present differing views on controversial exam security topics. Before and after each debate, audience members will be asked for their opinions on each topic. By the end of this keynote, you will have heard the positions of seasoned professionals and learned the collective wisdom of the crowd to help YOU ultimately decide what "appropriate" exam security means for your program.

11:15 – 12:15 Session 7

RS8: STUDENT DATA PRIVACY COMPLIANCE **Pyle 209**

US Student Data Privacy Compliance Landscape
William Wells, NCS Pearson

Integrating Compliance into Information Security Programs
William Wells, NCS Pearson

Measuring Information Security Risk in Quantitative Terms
William Wells, NCS Pearson

PP13: ***Communicating Through Conflict: How to Be the Calm in the Storm*** **Pyle 226**

Kim Brunnert, Elsevier
Richelle Gruber, Caveon Test Security
Marc Weinstein, Caveon Test Security
Walt Drane, Mississippi Department of Education

PP14: ***Cheater Cheater!*** **Pyle 309**

Rachel Schoenig, Cornerstone Strategies, LLC
Faisel Alam, Law School Admission Council
Ray Nicosia, ETS
John Fremer, Caveon Test Security
Ardeshir Geranpayeh, Cambridge English Language Assessment
Cody Shultz, Guidepost Solutions

PP15: ***The Importance of Having Proper Protocols in Place to Effectively Investigate an Exam Security Breach*** **Pyle 325-326**

Linda Johnson, National Association of Boards of Pharmacy
Jennifer Davis, National Association of Boards of Pharmacy

12:15 – 1:15 **Lunch** **Pyle Alumni Lounge**

1:15 – 2:15 Session 8

RS9: Detection of Preknowledge **Pyle 209**

Detection of Item Performance Changes Through Dynamic Programming
Dennis Maynes, Caveon Test Security

Modeling Item Sharing for High Stakes Tests - An Epidemiological Perspective
Alex Brodersen, University of Notre Dame

Strengthened Scale-Purified Deterministic Gated Item Response Theory Model
Xinhui Xiong, AICPA
Anqi Li, University of Illinois Urbana-Champaign

DP5: *Effective Interviewing and Interrogation Techniques* **Pyle 325-326**
Cody Shultz, Guidepost Solutions
Mikel Trevitt, ACT, Inc.

PP16: *Closing the Barn Door BEFORE the Horse Bolts – Performing an Internal Test Security Audit* **Pyle 309**
Cathy Koenig, American Board of Pediatrics
Katie Gottwaldt, National Board of Certification and Recertification for Nurse Anesthetists

FR2: *BAD GENIUS:* Reactions, Reflections, and Implications for our Industry **Pyle 226**

Presentation Abstracts

(alphabetical by last name)

Steve Addicott
Caveon Test Security

Beverly Bone
IBM Certification Programs

Panel Presentation 8: Thursday, 3:00 – 4:00

Testing in Hell: What to Do When Everyone Seems To Be Cheating

Many test programs in Certification/Licensure, Education, and Workforce Skills Credentialing find the rewards for expanding internationally to be compelling. Administering high-stakes exams in certain parts of globe, however, can present vexing challenges to the integrity of test results. Imagine the worst places to test, where all threats are real and programs operate at high risk.

Quite simply, administering tests in some countries is inherently problematic and risky. This can be attributed to many causes. While the source of problems varies in each country, some challenges are common across these geographies, and often involve the following dynamics: (1) cheating on exams is widely socially acceptable, (2) the use of sophisticated cheating devices and technology to circumvent test security protocols is rampant, (3) a strong tradition exists for using time-tested methods for stealing test items, (4) proxy test takers operate virtually unrestricted in many areas, and (5) in order to cancel a test score in certain cultures, undeniable proof of cheating is required.

The test security threats involving these overarching dynamics present daunting risks; risks that affect test score validity and create financial burdens. In order to build appropriate defenses, a deep understanding of threats and risks must be developed.

While test programs everywhere may contend with these same challenges to some degree or another, this session will explore WHY these issues are more pervasive and intense—and in turn more problematic—in parts of the globe. This session's presenters not only understand the impact of these threats and risks, but also their causes. Panelists will share these findings in an effort to educate other test program leaders who will or already are facing similar issues.

To overcome these hurdles, test programs are forced to innovate and experiment. Aggressive measures are required to administer tests securely in these countries. These measures must prevent, detect, and deter test fraud. Security should be built into testing processes, not bolted on.

The experiences of the panel include real-world examples of wins and losses in the international battle for trustworthy test results, and each presenter will share his or her top tips for dealing with geographies that are particularly rife with test fraud. At the conclusion of this session, high-stakes test program managers will be better informed and more powerfully equipped to administer their exams in "hell."

Elizabeth Amador
Western Governors University

Research Session 1: Thursday, 9:45 – 10:45

Predicting Cheating Before It Happens

What if we were able to predict cheating before it happened?

How can we verify that the security of online proctored assessments is reliable?

These questions sparked the genesis of the Security Index Project, brainchild of the assessment security team at Western Governors University. Currently, the index calculates risk based on a 25-point algorithm and determines a risk score for students' test sessions with many other points in development and testing. The security index focuses the attention of the assessment security team allowing for more efficient security reviews of online proctored (OLP) assessments, and higher quality feedback to our OLP partners.

Short-term plans include utilizing the security index to flag sessions within online proctored assessments. An in depth review and analysis of these videos allows institutions to verify protocol is followed for the security of their assessments. Data gathered by the index and the results of this further review will be recycled into the algorithm to continually keep on top of trends in test-taker behavior.

In the long-term, the project has potential to give institutions the advantage of knowing the statistical probability of non-approved behavior occurring before the exam takes place based on millions of exams taken.

Michelle Barrett
Pacific Metrics

Walt Drane
Mississippi Department of Education

Michael Clifton
ACT, Inc.

Wes LaMarche
IMS Calipers Analytic Standard Working Group

Panel Presentation 3: Thursday, 11:00 – 12:00

We Can Do Better: Security of Test Exams and Identifying Events in Near Real-Time

Wait times associated with recovering assessment data from vendors often prevent education administrators from achieving the level of insight desired; in fact, "perishable insights", those that occur at a moment's notice and must be acted on quickly within a narrow window of opportunity before they lose their value (Gualitieri & Curran, 2014), are seemingly absent in large-scale assessment. Yet this type of insight may be of great value to the assessment community. Change is on the horizon, however. In this workshop, we will first discuss a state perspective on a small pilot to explore perishable insights of importance to assessment stakeholders. From a test security perspective, we will address the use of this data in real time and what value it might add. Participants will be able to experience some real time analysis with mock scenarios. Finally, we will discuss advances with the IMS Global Caliper Analytics standard assessment profiles, which provide for near-real time emission of detailed assessment data from assessment platforms.

Dmitry Belov
Law School Admission Council

Stephen Cubbellotti
American Board of Internal Medicine

Research Session 6: Friday, 8:00 – 9:30

Detecting Examinees with Aberrant Answer Changes in CBT via Posterior Shift

The statistical analysis of answer changes (ACs) has proven to be helpful in identifying possible testing irregularities on large-scale assessments and is routinely performed at some testing organizations. The purpose of this study is to assess whether the addition of time spent changing answers improves the sensitivity and accuracy of identifying aberrant responses. In this study, time spent on changing answers is added into the analysis. In particular, for each examinee the response vector (including scored responses and response times) is partitioned into two disjoint sub-vectors: responses where answers were unchanged and responses where ACs occurred. The proposed statistic measures a difference in performance (in terms of score and speed) between these sub-vectors, where only final responses and final response times are used. For each examinee, the difference in performance is computed as a weighted sum of posterior shift between corresponding posteriors of ability and posterior shift between corresponding posteriors of speed. In other words, examinees that view some items multiple times, change their final responses with a short final response time, and gain scores on these items higher than on other items may be flagged by the new statistic. The performance of the new statistic on simulated and real responses to a high-stakes CBT will be compared with other popular statistics.

Alex Brodersen
University of Notre Dame

Research Session 9: Friday, 1:00 – 2:00

Modeling Item Sharing for High Stakes Tests - An Epidemiological Perspective

Compartmental models, such as the Susceptable-Infectious-Recovered (SIR) model (Kermack and McKendrick, 1927), have long been used in epidemiology as a method for modeling the rates of transmission and recovery of a disease in a population. More recently, these models have been applied in the context of modeling the spread of information, such as the spread of so-called viral videos (Cheng, Li, and Liu, 2013), or of ideas (Woo, Son, and Chen, 2015). A literature review suggests compartmental models have not been applied in the educational testing domain for modeling item sharing. There are several applicable analogies between traditional compartmental models and item sharing, such as defining potential test takers who have not yet been exposed to leaked items as 'susceptible individuals'. However, several modifications are required to formulate a reasonable model. For example, many compartmental models dictate a fixed population size whereas in most testing situations test takers enter and exit the population systematically. Also, the concept of 'recovered' individuals does not apply to testing as the knowledge of item content could realistically keep an individual in the 'infectious' stage indefinitely. The current study aims to accomplish three goals: 1.) propose modifications to existing compartmental models for their use in test security, 2.) formulate strategies for collecting relevant data and estimating model parameters, and 3.) suggest interventions based on model estimates.

Kim Brunnert
Elsevier

Richelle Gruber
Caveon Test Security

Marc Weinstein
Caveon Test Security

Walt Drane
Mississippi Department of Education
Panel Presentation 13: Friday, 11:00 – 12:00

Communicating through Conflict: How to be the Calm in the Storm

Test Security Professionals are often called upon to lead conversations about delicate topics (like aberrant scores, collusion, pirates, rogue sellers, security, or compromise) with internal and external customers who use accusatory and inflammatory words (like stealing or cheating) that often elicit strong emotions not only with the customer but also with the Test Security Professional. Communication plans and standard verbiage are critical but too often new incidents require more.

This panel session will be structured with examples and accompanying commentary. The panel experts will provide tips, considerations, and strategies for creating and fine-tuning communication plans and standard verbiage for everyday as well as for unique situations.

Topics could include: how to stay unemotional and thoughtful about your response even if you're more upset than the customer; how to find words to calm and reassure without making promises or saying too much; and how to be proactive (rather than reactive) in your response. The panel consists of three experts who are educated, trained, and experienced in crisis communication, speaking as if someone is recording your every word, and public speaking.

Jennifer A. Brussow
University of Kansas

William P. Skorupski
University of Kansas

W. Jake Thompson
University of Kansas

Poster Presentation 6: Thursday, 5:30 – 7:30

A Hierarchical IRT Model for Identifying Group-Level Aberrant Growth

As cheating on high-stakes tests continues to be an issue for standardized testing, approaches for detecting cheating proliferate. Approaches to cheating detection vary, with common strategies being detecting unusual levels of wrong-to-right erasures (e.g., Wollack, Cohen, & Eckerly, 2015), similarity of answer patterns (Karabatsos, 2003), and aberrant improvement over time (e.g., Bishop & Egan, 2016). However, the majority of research focuses on detecting cheating at the individual level. As recent events have shown (e.g., the Atlanta cheating scandal), cheating at the group level is also a threat to the validity of decision made from scores on high-stakes standardized tests.

The present study adapts the Bayesian Hierarchical Linear Model (BHLM) introduced in Skorupski & Egan (2013, 2014) and further developed in Skorupski, Fitzpatrick, and Egan (2016) to detect group-level aberrance within an IRT framework. Since many testing companies use a latent trait model to estimate examinee ability, this method may prove more compatible with operational testing programs' current approach to scaling.

For this study, data will be simulated to emulate two years of standardized test scores for students nested within classrooms. Examinees will be simulated within 300 total classrooms with group sizes ~U(5, 35) and with a mean increase in ability of 0.5 standard deviations. These conditions were chosen to mirror typical class sizes and growth rates observed in the American educational system and also to facilitate comparisons with the BHLM simulation in Skorupski, Fitzpatrick, and Egan (2016). Variables to be manipulated will include the size of the cheating effect (τ_g, either 0.5 or 1.0) and the percentage of groups simulated to be aberrant (1% or 5% of groups).

Michael Clifton
ACT, Inc.

Poster Presentation 4: Thursday, 5:30 – 7:30

Test Security Ecosystem

As part of the poster presentation format, I propose to display an interactive database of test security related information. The database benefits test security practitioners by connecting them with information that is sortable, timely, and targeted to their needs. Visitors will be encouraged to access the free database and contribute suggestions as to its evolution.

Jennifer Davis
National Association of Boards of Pharmacy

Nathan Thompson
Assessment Systems Corporation

Demonstration Presentation 2: Thursday, 11:00 – 12:00

The New SETI Project: Tools to Help the Search for Examination Treachery and Iniquity

When a test becomes compromised due to item pre-knowledge, proxy test-taking, item harvesting, or other treacherous and iniquitous behaviors, the validity of test-score inferences degrades. Cizek and Wollack's 2017 book Quantitative Methods for Detecting Cheating on Tests illustrates that many statistical methods are available to detect anomalous patterns of test results. However, due to the lack of easily-available software, how accessible are these methods to most operational testing programs?

The objective of this session is to discuss several tools for detecting test compromise that are relatively easy to use and do not require advanced knowledge of psychometrics or programming. We will cover ASC's SIFT and the R CopyDetect Package, both of which can be used to examine possible collusion among examinees. SIFT offers a range of intra-individual and group-level indices. The Outlier Detection Tool (ODT) developed by NABP will also be covered. The ODT is an Excel-based program designed to identify performance outliers and anomalous response patterns indicative of examinee treachery. The features of these tools will be discussed and compared, including some technical aspects of the statistical methods employed.

Simulated collusion and other cheating behaviors will be seeded into the results from a real examination, and each of the three tools will be used to analyze the resulting dataset. Results will be presented and compared across SIFT, CopyDetect, and ODT.

Attendees to this session will come away with an understanding of how they could utilize SIFT, CopyDetect, and/or the ODT framework in their testing program. Though there are many statistical methods to detect cheating, accessible software currently exists only for a handful of them. Another goal of the session is to promote sharing of tools among practitioners and provide impetus for the development of statistical screening tools that could be easily used by the broader test security community.

Walt Drane
Mississippi Department of Education

Sally Valenzuela
Caveon Test Security

Poster Presentation 11: Thursday, 5:30 – 7:30

A Comprehensive Test Security Program for States

Test security is an important component of ensuring the validity of assessment data used for accountability. As states are preparing new ESSA (Every Student Succeeds Act) plans, ensuring the integrity and fairness of test administrations is critical. Common threats to test security in the K-12 testing environment include unauthorized use of technology, inappropriate assistance to students during testing, and student pre-knowledge of test content from educator coaching. These and other threats carry the risk of creating spurious test data that can impact educational decision making for individual students, educators, and schools.

This poster will illustrate how one state has created a comprehensive test security program that addresses prevention of testing irregularities, deterrence of unwanted behavior, detection of potential breaches and anomalous test score data, test incident management and response, and overall test security policies and procedures evaluation.

Carol Eckerly
Alpine Testing Solutions

Russell Smith
Alpine Testing Solutions

John Sowles
Ericsson

Research Session 4: Thursday, 3:00 – 4:00

Analysis of the Discrete Option Multiple Choice Item: Examples from IT Certification

The Discrete Option Multiple Choice (DOMC) item format was developed by Foster and Miller (2009) as an alternative to the traditional Multiple Choice item format to limit examinees' exposure to complete item content. Rather than having access to the stem, key, and all distractors concurrently then choosing a response, examinees only gain access to response options one at a time as a series of dichotomous true/false responses which are randomly administered to each examinee. Options continue to be administered until an examinee either correctly identifies the key as correct or incorrectly identifies a distractor as correct. Limited research has been conducted to determine whether DOMC items are psychometrically comparable to traditional multiple choice items or whether response processes to DOMC items fit traditional measurement models (Kingston, Tiemann, Miller, & Foster, 2012; Foster and Miller 2009). We propose conducting analyses on real data from two separate IT certification programs to address these questions. In the first example, all items in an exam were initially administered as traditional multiple choice, then were converted to DOMC format in response to security threats. In the second example, a small number of unscored items on an exam were randomly assigned to examinees as either DOMC or traditional multiple choice. For each of these examples, we will compare the performance of the DOMC items to their traditional multiple choice counterparts as well as address measurement model fit to DOMC items. Both of these examples differ from previous research due to their high stakes nature. We also plan to include a simulation study informed by the real data analysis to address questions about fairness and decision consistency related to the DOMC item format.

Steve Ferrara
Measured Progress

Steve Addicott
Caveon Test Security

Panel Presentation 4: Thursday, 11:00 – 12:00

Designing, Developing, and Implementing Security Policies and Practices for Formative Assessment Products

Most peer published work on test security in educational testing focuses on detection of cheating on high stakes accountability tests. Less guidance is available on preventing, investigating, and resolving all types of test security violations for high stakes tests (Ferrara, 2017). Even less guidance is available for formative assessment products, even those with rigorous content design and psychometric characteristics. The reasons for protecting test security for high stakes tests—examinee data privacy, test data integrity, protection of copyrighted intellectual property—can be every bit as important for commercial formative assessment products used by school districts (e.g., ACT Aspire, eMPower, iReady, STAR). In this session, presenters from a formative assessment product and services provider and a test security services provider will describe (a) application of the framework for comprehensive test security systems for a new formative assessment product, (b) threats to security of formative assessments, and (c) potential protections and remedies. The session will address both paper-pencil and online test delivery. Some parts of the session will be interactive.

Presenter 1 will define and provide examples of formative assessment products; distinguish interim, benchmark, and classroom formative assessment products; describe a comprehensive framework for designing comprehensive test security systems and how the framework is being implemented for a new formative assessment product; and identify risks that are specific to the security of formative assessments. Presenter 2 will introduce a simple risk analysis methodology for prioritizing test security efforts and identifying threats and related risks specific to the security of formative assessment products. He will propose protections and remedies to help prevent, deter, detect, and respond to identified cheating and piracy threats and comment on the test security framework.

Session participants will be asked to identify other threats to security of formative assessments, propose protections and remedies, and comment throughout the session.

David Foster
Caveon Test Security

Workshop 1: Wednesday, 12:00 – 1:30

Implementing DOMC: Fast and Easy

Come to this DOMC workshop to learn how to use DOMC in your testing program for security and other reasons. DOMC is a 'hot' topic today as several programs have discovered these benefits and have begun sharing their experiences. The workshop will briefly cover what DOMC is, how it works, and what are its advantages. The workshop will then cover how to convert an existing multiple choice exam to the DOMC format, and how to create a DOMC-based test from scratch. A growing number of test development and test administration vendors are supporting the new item format, and details on those will be provided. Implementing DOMC is not without some challenges. While the challenges are relatively minor, direction on handling them well will be provided. Communication samples will be provided for introducing the DOMC concept to stakeholders and examinees, as will instruction in creating tutorials and DOMC practice tests. Those who are using DOMC today will be at the session to provide their experiences and take your questions. Whether you are planning to use DOMC now or sometime in the future, this engaging and fun workshop will provide all you need to know to begin using DOMC to make a quantum leap forward in the security of your tests.

David Foster
Caveon Test Security

Research Session 4: Thursday, 3:00 – 4:00

Making Sense of the Science of DOMC

This session will appeal to those who appreciate that testing innovations must have research and scientific support, and to those who believe that research can be explained in straightforward and useful ways. The research is all about the Discrete Option Multiple Choice (DOMC) which is a relatively new item format that can realistically replace traditional multiple choice questions. But what are the pros and cons of such a change? And how do we know that the pros are real and the cons handled well. Trained as a scientist, I have tremendous respect for research, even solid experimentation, to provide the most trustworthy direction on something new. This session presents summaries of the latest high-quality DOMC research studies, brings them all together, and sends the attendee on his or her way armed with useful information that is as close to truth on DOMC as science can provide.

David Foster
Caveon Test Security

Jamie Mulkey
Caveon Test Security

Poster Presentation 12: Thursday, 5:30 – 7:30

Can You Outsmart a SmartItem™?
Generating enough items to limit item exposure and stop the use of test content pre-knowledge has always been a struggle for testing programs. Having the right resources to develop enough items in a short amount of time is a challenge if not impossible. Even if you could develop a lot of items, how do you know they would be valid measures of the knowledge or skill being tested? What if there were a way to generate an exponential number of construct-relevant items quickly with limited resources, and a means to measure item performance consistently? Come and see the new Caveon SmartItems.

Join us for the Caveon SmartItem Challenge! We will demonstrate how Caveon SmartItems™ are developed. Then we will provide a SmartItem experience, demonstrating how difficult it is to outsmart a SmartItem!

Research will be presented showing the effectiveness of SmartItems™ at protecting test content while at the same time contributing to psychometrically sound test scores.

John Fremer
Caveon Test Security

Jarret Dyer
College of DuPage

James Wollack
University of Wisconsin – Madison

Ben Fortney
University of Wisconsin – Madison

Panel Presentation 5: Thursday, 11:00 – 12:00

Cheaters Say the Darnedest Things!

What goes on in the minds of people who cheat on tests? Test center administrators were contacted and asked to provide some of their more memorable test cheater stories. The stories range from use of common “old school” cheating methods to sophisticated high technology tools, from test taker impersonation to throwing tantrums upon being found out. Understanding how these situations occurred can assist in developing policies and procedures to prevent such actions to undermine fairness and validity.

It is interesting to try to understand how cheating is rationalized; are individuals cheating because they need to ensure a successful outcome? Do they cheat because they can get away with it? Is it due to a perceived lack of value of the test itself? Do they miss nuances of what is considered cheating?

Additionally, there are many types of cheating that occur. Test administration modalities in both paper and computerized environments allow cheaters to get creative with smuggling content into the testing environment. Test administrators have seen handwritten notes on body parts, exchanges of information in restrooms, proxy test taking, as well as a prevalence of accessing information through electronic devices.

Our storytellers first look at some memorable proctor/test taker cheating encounters. We will use these stories to emphasize the importance of following the right test security processes and procedures. By making sure these fundamental test security elements are in place, there is a better chance of reducing inappropriate test taking behaviors.

Research on test cheating will then be discussed and a framework provided for helping set the right expectations and context for preventing inappropriate test taking behaviors.

Jennifer Geraets
ACT, Inc.

Emily Scott
ACT, Inc.

Demonstration Presentation 1: Thursday, 9:45 – 10:45

Identifying and Implementing Test Security Enhancements for Your Testing Program

Developing a test security plan can be a challenging process for any testing program. How do you determine what needs to be put in place to deter and detect misconduct? How do you decide what you should do if there is a test security incident? This session will introduce participants to a systematic process of analyzing their testing programs from test development through administration and the scoring and reporting process. Participants will learn how to identify risks and mitigation strategies throughout the entire lifecycle of a test, and will leave with suggestions for how to prioritize test security enhancements and work with others in their organizations to begin to integrate those enhancements into their testing program(s).

Ardeshir Geranpayeh
Cambridge English Language Assessment

Research Session 2: Thursday, 1:00 – 2:30

Two Decades of Investigative Cheating Detection Research

In this paper we review the security of test results over the last two decades. With increasing importance of the consequences of test performance for candidates in contexts such as immigration, access to higher education or job opportunities, the stakes associated with the use of test results have increased. We argue that as the stakes of test increase so does the level of cheating. We further argue that cheating is an inevitable consequence and a by-product of high stakes testing. Against this background, we look at how cheating practices in international examinations have been transformed and facilitated by technology in the last two decades. We share cheating detection methods to combat the increasingly new challenges for test security from an international perspective and propose comprehensive measures to address them.

Joe Grochowalski
The College Board

Research Session 7: Friday, 8:00 – 9:30

Visualizing Test Fraud using Multiple Correspondence Analysis

Clustering via Forced Classification (CFC) is a new method that enhances test fraud detection by visualizing test takers' response patterns. The purpose of this study is to introduce the CFC method and to illustrate its strengths and weaknesses using simulated data.

CFC uses Forced Classification, a method based on Multiple Correspondence Analysis, and plots each test taker as a single point on a low-dimensional map. Test takers with similar response patterns have close proximity on the map, making pairs or groups of test takers with unusually similar responses easy to detect. CFC emphasizes unusual response (i.e., aberrant) patterns, which are often important for detecting test fraud. CFC is similar to existing methods that compare aberrant and wrong answers, but it has a number of additional useful features: (1) CFC improves exploratory interpretation of suspected test fraud by visually mapping test takers' patterns of responses, rather than relying on statistics from pair-wise comparisons. (2) CFC identifies multiple dimensions for the similarity between incorrect responses, which can improve discrimination between suspected fraud and legitimate test-taking behavior. (3) Visual mapping helps to quickly identify groups of test takers by pairs, classes, schools, or test centers. (5) CFC results can be enhanced using control covariates, like test score history, to further improve the accuracy of fraud detection. CFC suffers from typical limitations, however, such as an inability to identify fraud when there are few incorrect answers, and reliance on ability estimates to detect patterns of cheating.

For this study, we simulated responses using the nominal response model, and added three contrived fraud scenarios: individual copying, collaboration among a small group of test takers at a specific location, and widespread fraud at a testing location. We compared the CFC visual detection results to similar fraud screening methods to illustrate its strengths and weaknesses.

Maxwell Hong
University of Notre Dame

Ying (Alison) Cheng
University of Notre Dame

Research Session 6: Friday, 8:00 – 9:30

Data Quality in Assessment

In the context of high-stakes tests, test takers who do not have enough time to complete a test would rush towards the end and may engage in random guessing behavior, when tests do not penalize guessing. Via mathematical derivations and simulations, Attali (2005) showed that such random guessing responses may lower reliability.

Meanwhile, some believe random guessing responses actually increase estimates of reliability. Supporting this belief, Wise & DeMars (2009) showed that random guessing does in fact inflate reliability estimates under certain conditions. Unmotivated participants who complete in low stake exams would omit or respond incorrectly to questions throughout the test. A greater proportion of these responses could occur and inflate reliability estimates. This issue is more common for low-stakes tests, such as psychological assessments or surveys.

Our research attempts to bridge the gap between these two positions. We will provide analytical and empirical evidence that random guessing responses will strictly attenuate estimates for: correlation amongst items and Cronbach's alpha, depending on the prevalence of such responses and how these responses are scored. Furthermore, we will extend previous research by reporting how such responses affect various forms of validity, test dimensionality, factor structure, item total correlations, and item rest correlations.

NooRee Huh
ACT, Inc.

Qing Xie
University of Iowa

Chunyan Liu
ACT, Inc.

Chi-Yu Huang
ACT, Inc.

Poster Presentation 3: Thursday, 5:30 – 7:30

Detecting Compromised Items in CAT Using a Sequential Monitoring Procedure

In computerized adaptive testing (CAT), items are selected from an item pool based on an examinee's ability estimate. Using the same item pool repeatedly may intensify item compromise because past examinees could share items with future examinees. When items are compromised, the items become easier for the cheaters, which will adversely alter the validity of test scores of the cheaters by overestimating their abilities. To protect the integrity of test scores, it is crucial to use effective statistical procedures to flag items as soon as they are compromised.

Zhang (2014) developed a classical test theory (CTT)-based and sequential procedure that targets monitoring items in real time. Zhang and Li (2016) developed an item response theory (IRT)-based sequential procedure and compared the performance of IRT-based procedure to the CTT-based procedure. They concluded that the IRT-based method has much lower Type I error rates and more power than the CTT-based method when the number of compromised items is small. The aforementioned sequential procedure research were based on the random-examinees assumption (Zhang, 2014) that items were administered to homogeneous ability groups throughout the administration. However, the ability levels of the groups may fluctuate across the administration in real testing situations; for example, a test may be administered at a school level, which could result in a school with lower (higher) ability examinees taking a test before a school with higher (lower) ability examinees, respectively.

This study will simulate a situation when the number of compromised items and cheaters increase gradually as more examinees take a test. The results of this study will provide helpful information on how the aforementioned sequential procedures detect compromised items effectively under different compromised-item conditions when an examinee group that takes a test first are somewhat different from a group that takes a test later in their ability levels.

Greg Hurtz
PSI Services LLC

John Weiner
PSI Services LLC

Research Session 1: Thursday, 9:45 – 10:45

Forensic Profiling of Test-Taker Response Patterns Associated with Distinct Cheating-Related Behaviors

"Traces" in forensic science are markers left by the occurrence of a (criminal) act. They are often framed as physical evidence left by the contact between objects, but by analogy when test-takers cheat in their interaction with test items they likewise leave trace evidence in their patterns of response data. Forensic data analysis can then be used to monitor and detect instances of this trace evidence. Specific types and strategies of cheating will leave different patterns of trace evidence (e.g., in correct answers, in errors, in response times, in interactions of these factors with item difficulties, etc.), so detection requires different indices that are sensitive to the different patterns. Our recent research has focused on methods for moving beyond the separate and independent use of multiple statistical indices to more of a forensic profiling approach based on distinctive patterns across multiple indices. This allows us to capitalize on relative strengths and weaknesses within groups of similar indices and account for different pattern sensitivities between different types of indices. Through this approach, model profiles of different cheating patterns across a suite of carefully selected indices are established, and each test-taker's similarity to each of several profile models is computed to determine which model (including a non-cheater profile) is the strongest fit for each test-taker. This approach has the advantage of allowing for theory development and model testing across virtually any collection of statistical indices for virtually any distinguishable pattern arising from a test-taker's response strategies and behavior. We will discuss our current selections for our suite of indices, our strategies for developing model profiles, computation of profile similarities, and classification accuracy in detecting test-takers whose actual responses are manipulated with different degrees of the focal trace patterns. We will also discuss future directions in expanding and potentially improving the index suite.

Greg Hurtz
PSI Services LLC

John Weiner
PSI Services LLC

Research Session 5: Thursday, 4:15 – 5:15

Ockham's Razor and the Selection of Collusion Indices: Variants of J2 Provide a Simple and Effective Diagnostic Tool

Ockham's razor presents the philosophy that among competing hypotheses or equally effective solutions to a problem, we should choose the simplest alternative that requires the fewest assumptions. With this in mind, we compared several alternative collusion indices for multiple-choice tests that are based at their core on the same underlying metric: The count of matching responses between pairs of test-takers. This study focused on a relatively simple regression model introduced as "J2" by Weiner et al. (2013) that operates at the level of match-counts and number-correct test-scores, comparing it and two new variants of it to two more widely-known but more complex indices by Frary et al. (1977; g2) and Wollack (1997; omega) that require probabilities for each person on each response option from an item response model. We also compared it to Sotaridona et al's (2006) conceptualization of Cohen's kappa as a test collusion index that incorporates item response models or empirically-estimated response probabilities in a simpler way, by recoding the response options into ordered categories. Because J2 is a regression model that operates entirely at the level of observed match-counts and number-correct scores, it makes no special item-level assumptions beyond those already made under classical test theory, and requires no item-level conditional probability computations. For our comparison we used N=1169 test-takers on a certification exam and repeatedly extracted random samples to manipulate subsets of test-taker responses, simulating patterns of (1) item preknowledge, (2) lower ability cheaters copying answers from higher ability sources, and (3) higher ability cheaters copying answers from lower ability sources. Results from ROC analyses indicated that a combination of two variants on the original J2 index provided strong detection of all three patterns while g2 and omega provided strong detection only for pattern 3. Results support the utility of variants of the simpler J2 model.

Linda Johnson
National Association of Boards of Pharmacy

Jennifer Davis
National Association of Boards of Pharmacy

Panel Presentation 15: Friday, 11:00 – 12:00

The Importance of Having Proper Protocols in Place to Effectively Investigate an Exam Security Breach

A security breach is anything that poses a threat to a testing organization’s greatest asset; its examination/assessment programs and intellectual property. Individuals involved with the theft of test questions use a variety of techniques to obtain content such as memorizing questions, collusion at testing centers, or capturing content with advanced electronic technology. There is no doubt participation in these wrongful activities is on the rise. The Internet and social media forums have become creative and lucrative platforms to disseminate item content. Accessibility to public sites that display exam content has made many examination and assessment programs more vulnerable to threats. Examination programs of all sizes, large or small, are not immune to harm that could result from security breaches. Organizations must continue to engage in strategic efforts to protect and prevent these threats from impacting their programs.

Organizations should have a test security plan or protocol in place identifying best practices for recognizing and responding to a breach. The plan should include steps relating to investigations of security incidents and the identification of key roles and responsibilities when it’s appropriate to launch an investigation. Pertinent questions should be asked when an alleged fraud occurs. Who was involved? When did it occur? What did they do? How did they do it? Will the results of the breach have a common outcome (i.e., score invalidations) or catastrophic outcome (i.e. loss of content, subsequently shutting down exam program)?

Participants in this session will receive guidance to develop a concise and detailed security plan and the approach used when investigating an examination security breach. We will introduce a case study and subsequent investigation that occurred at NABP. As the result of the breach, we will discuss what steps were executed appropriately, what we learned, and ways to improve the processes.

Cathy Koenig
American Board of Pediatrics

Katie Gottwaldt
National Board of Certification and Recertification for Nurse Anesthetists

Panel Presentation 16: Friday, 1:00 – 2:00

Closing the Barn Door BEFORE the Horse Bolts - Performing an Internal Test Security Audit

The time to assess the security of your organization’s examinations is not after a test security breach has occurred. Discovering and remediating security gaps is essential to test security. This session will show attendees how to find and use available resources to plan, conduct, and act upon the results of an internal test security audit.

Test sponsors will share their organizations’ experiences with planning, conducting, and acting on internal test security audits.

Seo Young Lee
University of Wisconsin – Madison

James Wollack
University of Wisconsin – Madison

Research Session 3: Thursday, 1:00 – 2:30

A Mixture Model to Detect Item Preknowledge Using Item Responses and Response Times

Although computer-based testing introduces some unique security vulnerabilities such as item preknowledge, it also offers unique detection opportunities through the collection of response times. While item response data provides information only about an examinee's problem-solving accuracy, response time data provides additional information about that examinee's problem-solving effort. Consequently, using both response time and item response data together may enhance the detection of examinees with item preknowledge.

Examinees with item preknowledge are different from honest examinees in that their answers to test questions are governed by some process in addition to the examinee's standing on the latent trait being measured by the test (?) that is likely to affect both response accuracy and response time. When a population is comprised of two distinct latent groups, it is often possible to distinguish them through the application of a mixture model.

In this study, we propose a mixture Rasch-LnRT model by extending van der Linden's (2007) model, which integrates an IRT model and a response time model within a hierarchical framework, to include a mixture component designed to detect examinees with item preknowledge. Simulation studies will be conducted to evaluate the performance of the mixture model to detect examinees with item preknowledge under various conditions. Preliminary results showed that the mixture model performed well to differentiate examinees with item preknowledge from honest examinees.

Adel Lelo
Western Governors University

Research Session 2: Thursday, 1:00 – 2:30

Online Proctoring - Best Practices

WGU has offered Online Proctoring since to its students since 2009. 8 years and more than 1 Million Online Proctored assessments later, we have learned lessons which we would now like to share with programs interested in entering into this space.

Attendees will learn about best practices in the following areas:

- Criteria to consider before picking an online proctoring service provider
- Technical integration
- How to verify security of assessment delivery
- Common cheating techniques used by test takers
- Next steps in online proctoring service delivery

We have made many mistakes over the years, mistakes we would avoid if we knew then what we know now. Armed with this information, programs who are considering utilizing online proctoring can avoid making the same mistakes.

Adel Lelo
Western Governors University

Carissa Pittsenberger
Western Governors University

Facilitated Roundtable 1: Friday, 8:00 – 9:30

Standards in Online Proctoring

With the ever increasing number of online proctoring service providers and an even faster increase in online proctoring service consumers, it is time for the test security industry to start setting standards around the practice.

Having overseen the delivery of more than 1 Million online proctoring sessions over 8+ years, Western Governors University would like to share lessons we have learned so far and solicit others' thoughts on creating standards which will help protect the credibility and validity of assessments delivered via online proctoring.

Our goal is to start a discussion on this topic and work with interested parties in further defining these standards in months to come.

Richard Luecht
University of North Carolina at Greensboro

Terry Ackerman
ACT, Inc.

Research Session 3: Thursday, 1:00 – 2:30

Some Graphical Techniques for Presenting Aberrant Response and Timing Data Analysis Results

Aberrant response patterns and/or unexpected sequences of response times can provide compelling evidence of potential cheating and other test security problems. Examples include various person-fit statistics to evaluate the consistency of response patterns (e.g., Karabatsos, 2003; Meijer, 2002; Meijer & Sijtsma, 2001), likelihood-based statistics (Drasgow, Levine, & Williams, 1985; Levine & Drasgow, 1988), and various types of Bayesian aberrance detection methods (van der Linden & Lewis, 2015; Shu, Henson & Luecht, 2013). Response times aberrance detection methods have also been proposed (e.g., van der Linden, 2006; van der Linden & van Krimpen-Stoop, 2003; Meijer & Sotaridona, 2006; Marianti, Fox, Veldkamp & Tijmstra, 2014), including methods that simultaneously model patterns or item response along with Bayesian checks (van der Linden, 2007; van der Linden & Lewis, 2015). However, very little attention has been paid to how this “compelling evidence” is subsequently presented to non-technical audiences. This paper borrows for the graphical design literature to suggest an array of useful and informative visualization techniques for presenting comparative aberrance statistics (e.g., Meirelles, 2013; Börner & Polley, 2014; Tufte, 2001; Cleveland, 1994; Jacoby, 1998). Examples will be provided with data requirements and R code, where possible.

Kaiwen Man
University of Maryland College Park

Hong Jiao
University of Maryland College Park

Jeffery R. Harring
University of Maryland College Park

Research Session 6: Friday, 8:00 – 9:30

Robust Bayesian Estimation of Item Response Model Parameters Accounting for Aberrance

Many high-stakes decisions often rely on accurate and reliable scores from large-scale assessments. However, aberrant responding behaviors such as cheating, creative responding, may reduce measurement precision in estimates of latent parameters for item response models with maximum-likelihood estimation (MLE) method if the percentage of aberrances is large.

Mislevy and Bock (1982) has proposed an approach MLE framework to reduce the biases of ability estimates by weighting the contributions of the item responses with the Bisquare method to the log likelihood function. Later, Schuster and Yuan (2011) also proposed another method by replacing the weighting method with Huber function. However, both MLE based approaches are limited with the type of aberrance response pattern. Meanwhile, both methods had severe convergences issues.

In order to overcome the shortcomings of the MLE based robust correction method, the Bayesian based robust adjustment method is explored in this study. It is expected that this method will not suffer from the main problems with MLE based methods, like convergence issue.

Dennis Maynes
Caveon Test Security

Research Session 9: Friday, 1:00 – 2:00

Detection of Item Performance Changes Through Dynamic Programming

Item performance change is often hypothesized to occur through drift as exam content becomes more generally known. However, when items are disclosed, the performance changes are expected to occur rapidly. Hence, an accurate determination of performance change can provide critical information for learning more about item disclosure and compromise, such as when, where, and by whom the content was disclosed. This paper uses dynamic programming as a means to detect item performance changes. Dynamic programming has the ability to detect rapid change so that the effect and timing of disclosure events can be reliably estimated. Simulations are used to assess the size of detectable item performance changes and when those changes occurred.

Dennis Maynes
Caveon Test Security

Rachel Schoenig
Cornerstone Strategies, LLC

Marc Weinstein,
Caveon Test Security

William Skorupski
Universtiy of Kansas

Camille Thompson
ACT, Inc.

Panel Presentation 6: Thursday, 1:00 – 2:30

Principles and Practices for Presenting Findings of Test Security Violations in Legal Settings

Recently, there has been some debate concerning Bayesian versus frequentist frameworks of statistical inference for presenting test security findings before a fact finder (e.g., a judge). However, this debate has taken place without regard for rules of evidence that prevail in the judicial system. For example, any statistical findings presented in a legal setting must be reproducible, scientifically sound, and presented using accepted methodologies. While the body of case law documents a small number of statistical approaches which have been used, it does not cover many of the unique and case-specific situations practitioners encounter. Consequently, analysts often use prior approaches in order to make appropriate test security inferences. Frequentists rely upon traditional statistical techniques that have been adapted to identify anomalies and outliers. Bayesians have suggested that models which can compute the probability of cheating must be used. While both approaches have merit, the findings must be presented in such a way that the fact finder will find the expert's testimony to be objective, factual, scientifically sound, and based upon reproducible and accepted methodologies. The panelists will offer perspectives on these and other relevant questions. The session is intended to open a serious and well-reasoned dialog on this topic, which has often been argued from an absolutist position.

Jamie Mulkey
Caveon Test Security

Tara Miller
Amazon AWS

Panel Presentation 2: Thursday, 9:45 – 10:45

This IS a Drill! Using Tabletop Exercises to Plan for Test Security Disasters

When test security disasters strike, you want to make sure staff and personnel are in the right place, doing the right thing, at the right time. A well-trained test security team has the ability to react swiftly and with confidence, so that the right actions are taken and the situation is resolved.

This is where tabletop exercises come in.

We can take our queue from emergency response personnel who conduct drills to prepare teams in handling crises when they arise. As test security professionals, this may not be a physical drill, as much as it is one of mentally walking through the necessary steps to address a test security situation.

This session will teach participants the techniques of designing and implementing tabletop exercises for their organization. The importance of security breach preparedness will be discussed. A methodology for developing tabletop scenarios will be shared. Participants will then work in teams to identify test security incidents and develop a tactical plan for use in a tabletop exercise.

John Olson
Caveon Test Security

John Fremer
Caveon Test Security

Kathy Moore
Kentucky Department of Education

Peter Zutz
Nevada Department of Education

Timothy Butcher
West Virginia Department of Education

Panel Presentation 7: Thursday, 1:00 – 2:30

Peer Review Requirements for States on Test Security and Monitoring: What was Learned from the Reviews of Recent Submissions to USED

Last year the USED included new requirements on test security for ESSA-required peer reviews, with all states required to submit evidence that their assessment systems have integrity and are secure. Peer Review Critical Element 2.5 on Test Security asks states to prove they have "implemented and documented an appropriate set of policies and procedures to prevent test irregularities and ensure the integrity of test results." To fulfill this requirement, states had to document their policies and procedures in four categories of test security:

- Prevention
- Detection
- Remediation
- Investigation

In 2016, 38 states submitted documentation/examples of how they meet these requirements. Many received feedback from USED that identified areas where they were either lacking in evidence or in the procedures being used, and a few states got positive feedback from peer reviewers that their approaches were acceptable.

In this session, three states will describe the types of evidence they submitted for peer review that address test security requirements, the feedback they got from USED on their submissions, and their plans for 2017. An assessment expert and peer reviewer will discuss the different types of evidence found to be particularly supportive of valid, fair, and secure state assessment systems, and recommend various ways that states can provide better evidence and documentation of their programs.

In addition, information from an informal survey by Caveon of those states that had peer review issues on test security and/or monitoring will be shared. Time will be allotted for Q&A and group discussion on approaches that can be most helpful to states in passing peer review requirements. Note that the issues covered in this session are relevant for any and all evaluations of test security, not just ones carried out for Peer Review purposes.

John Olson
Caveon Test Security

John Fremer
Caveon Test Security

David Ragsdale
Massachusetts Department of Elementary and Secondary Education

Elaine Themm
Michigan Department of Education

Panel Presentation 12: Friday, 8:00 – 9:30

State Strategies to Comprehensively Address Test Security Issues to Improve their Assessment Programs

Increasing numbers of states are dealing with test security breaches and have found evidence of cheating by some teachers or school administrators.

Recently, many states have developed comprehensive strategies for improving testing integrity and security in their assessment programs. State strategies often include an approach that focuses on three key aspects of test security—prevention, detection, and follow-up investigations:

1. Prevention: Implementation of enhanced state policies/procedures for security, address issues specific to online assessments, develop targeted training materials, improve monitoring of test administrations, etc.

2. Detection: Successful implementation of data forensics, increased emphasis on affirming the validity of state assessment results for commonly used purposes, appropriate uses of results from forensics analyses, etc.

3. Follow-Up Investigations: Strategies for planning and conducting investigations, actions that need to be taken based on findings from investigations, procedures for conducting investigations in districts and/or schools, etc.

In this session, three states discuss the various steps they've taken to implement a comprehensive strategy for a strong test security system that includes many types of approaches, e.g., a strategic vision for testing integrity/security, comprehensive communications plan, standardized training design/plan for conducting investigations, secure computer-based testing system design, monitoring test administrations and the Internet, and regular use of data forensics. A well-known test security expert will discuss the various state approaches, provide his feedback, and recommend models of multifaceted security solutions for states. Attendees of this symposium will also receive a detailed outline for use in developing and implementing a comprehensive strategic vision and plan.

Yiqin Pan
Beijing Normal University / University of Wisconsin – Madison

Fang Luo
Beijing Normal University

Research Session 7: Friday, 8:00 – 9:30

Detection of Answer Copying in China's College Entrance Examination via Kullback-Leibler Divergence and ω-Index

The purpose of the paper is to detect answer copying in China’s College Entrance Examination, the most important test for all the high school students in China. The detection method builds on a two-stage method designed by Belov and Armstrong (2010). As the process suggests, the present research partitions the test into two subtests, objective items (multiple-choice items) and subjective items, considering the former as the performance after copying and the latter as the reflection of examinees' real ability, because there is lower difficulty for examinees to copy the responses in objective items than the ones in subjective items. The first stage uses Kullback-Leibler divergence to measure the difference between these subtests, sifting through examinees and retaining the individuals demonstrating inconsistent performance. For each examinee with aberrant behavior, the second stage applies ω-index to measure the agreement between the responses in multiple-choice items, detecting answer copying. The present research considers several conditions, including three copy percent level: 100%, 80% and 60%, three source ability level: exceeding 100% other examinees, exceeding 80% other examinees, exceeding 60% other examinees, 3×3=9 conditions in all. The detection result with simulated data shows that copy percent level has a significant influence on copying-detection rate, but source ability level not, type I and type II error rates are relatively low in all conditions. Therefore, the combination of Kullback-Leibler divergence and ω-index is effective in detecting copying in China’s College Entrance Examination.

Gerald Pia
Roche Pia LLC

Victoria Quinn-Stephens
Cisco Systems, Inc.

Research Session 2: Thursday, 1:00 – 2:30

Cisco v. TestKing, Pass4sure & Test-Inside: Results and Underlying Strategies from a Recent Exam Piracy Case

This session will examine the enforcement procedures utilized by Cisco to protect its Certification Exams in a recent litigation instituted against three major players in the internet test-preparation industry. Participants will gain an understanding of the options available to certification providers who want to reduce cheating by unscrupulous candidates, and who seek to protect and maintain the value of the certifications held by honest candidates. Courts have become more aware of the prevalence and impact of piracy and other misconduct transpiring in cyberspace, and they have demonstrated a willingness to issue modern forms of relief that impact those operating in anonymity on the internet, such as the freezing of financial accounts, and the impoundment of domain names and digital files. This session will focus on the relief available to certification exam providers and intellectual property owners combating online piracy in court, as well as effective and efficient anti-piracy (and anti-cheating) in-house programs. The presenters will discuss enforcement strategies available for all enforcement budgets, as well as options for providers and IP owners to consider in 2017 and beyond.

Carissa Pittsenberger
Western Governors University

Poster Presentation 1: Thursday, 5:30 – 7:30

It's a Bird; It's a Spider: No, It's the Owlbot!

Western Governors University (WGU) is a competency-based, student-focused, online, nonprofit university. The degrees awarded are based on a valid expression of competency determined by assessments. As such, a secure, dynamic, and reliable method to search for WGU high stakes assessment material on the internet was needed. WGU began searching for a tool that would provide this ability, but the team could not find anything workable; so, in true WGU fashion, it was developed: enter the Owlbot!

The idea for a specific WGU webcrawler was brought to a third party already working within the assessment delivery platform, Excelsoft. Excelsoft developed the webcrawler, known as Owlbot, within specific guidelines requested by WGU's Assessment Security and Academic Authenticity Team. The brand new technology was developed from scratch during the collaboration.

WGU uses the Owlbot to crawl and index specific websites. The application searches for assessment detail and shows the matches found. The information is searched within the application itself, ensuring that WGU assessment material is not exposed as part of the search process. The matches are then reviewed to determine if there is a clear concern regarding WGU copyrighted materials, if additional review of the material is needed, or if no concern is present. For those matches that do show issues involving WGU materials, a Digital Millennium Copyright Act takedown notification is processed, auto-populated within the application, and sent to the offending website. If there is no concern, the match is ignored and will not show up in future indexes. This ability to ignore irrelevant information will ultimately narrow the indexes and filter out the noise seen in manual internet searches.

Over 3,200 urls containing WGU information have been addressed using the super strength of the Owlbot!

David Ragsdale
Massachusetts Department of Elementary and Secondary Education

Research Session 2: Thursday, 1:00 – 2:30

Investigating Multi-year Cheating on State Assessments: A Case Study

This presentation will be a case study of an example of multi-year cheating on state assessments at an elementary school that the Massachusetts Department of Elementary and Secondary Education investigated. The following topics will be addressed.

- The data forensics that showed anomalous results at the school, leading to it being flagged as a data outlier
- The additional analysis and research done to develop the case, including cohort analysis, erasure analysis, rescoring constructed-response items, and tracking the performance of students who left the school.
- The information gained from interviews with teachers and administrators
- How the Department partnered with the school district to oversee the following spring's testing and ensure its integrity
- How the case was built to support invalidation of multiple years of test results, and licensure revocation for the educator involved
- Lessons learned from the investigation including what mistakes were made

This multi-faceted year-long investigation will be used to spark discussion of broader questions of identifying and investigating breaches of test security.

Sakine Gocer Sahin
University of Wisconsin – Madison

James Wollack
University of Wisconsin – Madison

Selahattin Gelbal
Hacettepe University

Poster Presentation 13: Thursday, 5:30 – 7:30

Performance Comparison of GBT Based on Purified Real Data

The overwhelming majority of research on cheating methods has involved simulated cheating, either within real-data contexts (e.g., Hanson, Harris, & Brennan, 1987; Bay, 1995; Wollack, 2003), or model-generated data contexts (e.g., Wollack, 1996; Wollack & Cohen, 1998; Sotaridona & Meijer, 2002; 2003). While these studies are useful for developing the methods, their generalizability to real world settings is limited because the assumptions made in simulating data/effects are often violated in practice to varying degrees. In addition, rarely has item preknowledge been studied within the context of exams for which all items have been compromised, and the existing methods are likely to struggle to detect well under these circumstances.

In this study, we investigate cheating detection for a two-dimensional, 100-item exam that was administered to approximately 10,000 examinees. The exam is known to include test compromise, with 44 examinees having admitted to seeing a subset of questions prior to the exam. The 20 items measuring one construct are believed to be largely secure. All of the known cheaters have acknowledged having access to all 80 items in the other section. Additional examinees are believed to have had access to these same questions, enough to produce a negative correlation between scores on the two sets of items.

In this study, we examine the performance of the General Binomial Test (GBT; van der Linden & Sotaridona, 2006) and score differencing to detect item preknowledge under a variety of sample sizes and scale purification strategies. The performance of the different methods is evaluated based on the detection rates of the examinees with known preknowledge.

Rachel Schoenig
Cornerstone Strategies, LLC

Faisel Alam
Law School Admission Council

Ray Nicosia
ETS

John Fremer
Caveon Test Security

Ardeshir Geranpayeh
Cambridge English Language Assessment

Cody Shultz
Guidepost Solutions

Panel Presentation 14: Friday, 11:00 – 12:00

Cheater Cheater!

Now more than ever, cheaters have so many options at their disposal – high tech gadgets, low tech tools, internet access, collusion with fraud rings, convincing fake IDs. How successful are these tools? What can the industry better deter or detect cheating attempts? Join security practitioners for an engaging workshop showcasing new tools and methods and designed to raise awareness of cheating tools. Attendees will leave with the ability to better identify effective methods to deter or catch cheating attempts and options for how to respond when cheating is suspected.

Rachel Schoenig
Cornerstone Strategies, LLC

Mike Clifton
ACT, Inc.

Nick Charge
Cambridge English Language Assessment

Ray Nicosia
ETS

Bryan Freiss
Pearson VUE

Workshop 2: Wednesday, 1:45 – 3:15

Planning and Responding to Test Security Incidents

If your testing program has value, then it isn't a question of IF a security incident will arise but WHEN it will occur. Incident response planning is a critical part of ensuring your program is prepared to successfully manage incidents and maintain your program's reputation. This workshop will discuss key aspects of incident response plans and successful response processes. Attendees will then participate in tabletop exercises designed to put into practice learnings and highlight some of the critical moments of incident response. At the end of the workshop, participants will have learned some of the basic building blocks and procedures for incident response planning and have experienced the use of incident response plans in simulations that reflect real-life test security incidents.

Rachel Schoenig
Cornerstone Strategies, LLC

Jennifer Geraets
ACT, Inc.

Jamie Mulkey
Caveon Test Security

Workshop 3: Wednesday, 3:30 – 5:00

Using the New Credentialing Security Framework

The recently published Credentialing Security Framework is designed to help increase trust in workforce credentials. Recent survey results have provided insight into the concerns of credential earners and users. Because credentials have a great deal to offer individuals, employers, and our communities, several efforts have been undertaken in both the EU and US to help bring more transparency to assessment-based credentials. The Credentialing Security Framework is one such effort, intended to help provide greater information to all stakeholders in the credentialing space and help restore trust in workforce credentials. Join the presenters to discuss recent survey results and learn how the Framework can be used to increase awareness of test security within and across the testing ecosystem.

Rachel Schoenig
Cornerstone Strategies, LLC

Jennifer Ancona Semko
Baker & McKenzie

Camille Thompson
ACT, Inc.

Panel Presentation 10: Thursday, 4:15 – 5:15

Legal Tips for Responding to a Test Security Incident

Responding to test security incidents can be stressful and highly political. Managing stakeholders, gathering evidence, and communicating with the media during high profile incidents can present multiple legal landmines. How a testing program responds can mean the difference between a smooth resolution and positive public perception or a rocky resolution involving litigation and reputational damage. Presenters will provide practical do's and don'ts for responding to, investigating, and resolving test security incidents, including considerations involving contracting, candidate interactions and due process, interviews and evidence gathering, and media responses. Attendees will leave with materials and knowledge that better positions their programs for successful resolution of test security incidents.

Emily Scott
ACT, Inc.

Tim Conlon
ACT, Inc.

Demonstration Presentation 3: Thursday, 3:00 – 4:00

Copyright and Trade Secret Protection for Standardized Tests

Copyright and Trade Secret Protection for Standardized Test Content: This presentation will outline the steps test publishers must take under federal copyright law, state trade secret law, and industry standards to ensure legal protection of their copyrighted test content. The presentation will also discuss the very current topic of copyright protection for computer-based and technology-enhanced test content and will shed some light on the current political climate and the copyright office's position on protection of this content.

Marcus Scott
Caveon Test Security

Poster Presentation 2: Thursday, 5:30 – 7:30

Using the ω Statistic to Estimate an Unknown Flawed Answer Key

Test thieves who steal test items, but not the test's answer key, sometimes make errors when they attempt to impute their own, resulting in a "flawed answer key." Statistical methods exist for identifying examinees who used a flawed answer key to take the test. However, these methods only work when the response pattern in the flawed answer key is known. This research focuses on the problem of estimating the response pattern of an unknown flawed answer key. Because examinees who use a flawed answer key are essentially copying from a response pattern that is external to the testing session, it may be possible to use answer-copying statistics to identify the response pattern in the data that most closely resembles the unknown flawed answer key. Four methods of using the answer-copying statistic, ω (Wollack, 1997), to impute the response pattern of an unknown flawed answer key are presented and evaluated by applying them to both real-life and simulated test data. One method, denoted Common_Max, had near-perfect performance on the real-life data. Another method, denoted Most_Flagged, was more than 96% accurate at estimating the flawed answer keys in the simulations.

Jennifer Ancona Semko
Baker & McKenzie

Lunch Presentation: Thursday, 12:00 – 1:00

An Update on Registering Secure Tests with the US Copyright Office

Registering test content with the U.S. Copyright Office is an essential part of any test security program. For the last several decades, the Copyright Office has offered a "secure test" registration process, which allows the copyright holder to register tests confidentially, without the need to deposit a publicly available copy of the registered materials. This summer, the Copyright Office issued an interim rule making several significant (and some say troubling) modifications to the secure test rules and processes. Public comments to these changes, including many submitted by members of the testing industry, have expressed a number of concerns. Among other things, the new procedures call into question whether certain forms of technology-based tests can still be registered, whether the Copyright Office will continue to accept the registration of item banks, and whether certain test delivery methods (continuous testing, remote proctoring) jeopardize the ability to use the secure test registration process. During this session, attorney Jennifer Semko of Baker & McKenzie will provide an overview of the Copyright Office's interim rule, efforts by members of the testing industry to obtain changes to the new rules, and the status of communications with the Copyright Office about this issue.

Jim Sherlock
Pearson

Research Session 4: Thursday, 3:00 – 4:00

Building a Secure Online Testing Platform in the Cloud

Many papers and discussions have focused on online testing security as it pertains to the local testing environment or client software. This presentation will outline a framework for building a highly secure large-scale assessment platform in the cloud. Topics will include choosing the right cloud vendor, architectural considerations, building security into the Software Development Lifecycle (SDLC), and compliance and governance topics. This presentation is relevant for those building secure testing applications as well as those responsible for ensuring security of online testing platforms provided by external vendors.

Cody Shultz
Guidepost Solutions

Mikel Trevitt
ACT, Inc.

Demonstration Presentation 5: Friday, 1:00 – 2:00

Effective Interviewing and Interrogation Techniques

Taught by a former CIA counterintelligence agent and ACT's Senior Manager for Test Security, learn the differences between interviews and interrogations, effective techniques for conducting them, and hear of real world successes and failures. Attendees will also learn the basics of behavior analysis and elicitation, how to conduct interviews over the phone or through Skype. Learn answers to questions such as: Does furniture matter? Is there one way to tell for sure that someone is lying? What if the subject speaks a language other than English?

Önder Sünbül
Mersin University

Seha Yormaz
Mersin University

Research Session 7: Friday, 8:00 – 9:30

Investigating the Performance of ω Index in Detecting Answer Copying

Several studies can be found in the literature about investigating the performance of ω under various conditions. However no study for the effects of item difficulty, item discrimination and ability restrictions on the performance of ω could be reached. Current study aims to investigate the performance of ω for the conditions given below. For this purpose, b parameter range was restricted in two levels (−2.50 – 0.00, 0.01 – 2.50), a parameter range; in two levels (0.10 – 0.80 and 0.81 – 1.5). After crossing a and b parameter ranges, four different item parameter cells were obtained. 10000 examinee responses were generated for each item parameter cell for 20 items. After combining four data sets, an- 80- item-dataset was obtained. In order to obtain the effects of source's and copier's ability levels to the performance of ω, ability range was divided into four intervals (−3.00 – −1.50, −1.50 – 0.00, 0.00 – 1.50 and 1.5 – 3.00). By crossing the ability ranges of source and copier, sixteen different combinations were obtained. Each of sixteen ability pairs of source and copier cheating was investigated for item parameter crossing cells for power study of ω. For type I error study, no cheating data were investigated for the same conditions and levels. Type I error inflations were observed for the lower copier ability levels. The results of power study indicated that when high ability level copier copied answers of the low difficulty level and high discriminative items from high ability level source, power of ω got weakened.

Anne Thissen-Roe
PSI Services LLC

Research Session 5: Thursday, 4:15 – 5:15

Effects of Local Testing Volume on Test-Center-Based Collusion Detection

The detection element of an effective operational test security program should have two levels of focus: 1) detecting individual test-takers who cheat or compromise the test, and 2) monitoring for systematic threats and weaknesses, such as organized cheating, gaps in policy, or gaps in operational effectiveness of security measures. For example, in proctored testing, the proctor watches test-takers for suspicious behaviors (level 1), but the testing organization must also monitor the proctor (level 2). For example, a well-meaning but poorly-trained proctor can systematically fail to observe suspicious behaviors or objects, such as small cameras carried into the testing room; a less well-meaning proctor can assist in cheating or item harvesting.

Belov's (2013) divergence framework for collusion detection fits well within such a two-stage detection model. Under the framework, test center candidate populations, or other salient groupings of test-takers, are evaluated for group-level patterns of individual aberrant response patterns, using a test statistic derived from information theory; only within suspect sites are candidates individually evaluated for membership in the aberrant pattern. In addition to detecting large-scale answer-sharing, the method is useful for detecting other systematic candidate misbehavior, and some operational security gaps.

However, the method has primarily been developed under a paradigm of test administration in defined windows, with candidate populations evenly distributed across test centers. This presentation considers lessons learned from an operational test security program: a distributed CBT operation with continuous testing. Natural variation in test volumes across test centers causes differences in the false positive rate between larger and smaller centers. Complicating matters, continuous testing necessitates continuous or recurrent monitoring, and yet testing volume changes naturally over time, leading to instability in findings. Illustrative examples are presented, along with some functional solutions.

Sarah Thomas
Caveon Test Security

Research Session 3: Thursday, 1:00 – 2:30

Is Time on Our Side? An Item-Level Latency Analysis to Detect Pre-Knowledge

Computerized-adaptive testing (CAT) and linear-on-the-fly-testing (LOFT) are modern advances that are being adopted by testing programs around the globe. Test security measures for these test designs present special challenges as well as a particular opportunity for the field of data forensics to advance. Many current data forensics methods are not compatible with CAT or LOFT tests, despite increasing demand. Previous research on item response latency statistics indicates that tau, a correlation between expected and actual response times, was the best detector of examinees with pre-knowledge. However, the latency statistics in that, and other, research have mostly been calculated at the test-level. Item-level latency statistics may provide more granular information, which would in turn allow groups of items (i.e., compromised and uncompromised) and groups of examinees to be identified. An additional benefit of item-level statistics is that they would allow statistical indicators of unscrupulous behavior to be computed in real time. This research presents a specific methodology for analyzing item-level latency statistics to detect examinees with pre-knowledge, which could be applied to CAT or LOFT exams. The proposed method will be applied to several datasets where indicators of pre-knowledge were collected, ranging from a certification exam where Trojan Horse items were administered and scored to data from an experiment where some examinees were exposed to test content before taking the exam. The results will be discussed in terms of the overall performance of the method and the implications of the findings for data forensics investigations.

Anna Topczewski
GED Testing Service

Poster Presentation 9: Thursday, 5:30 – 7:30

Considerations for Detecting Test Misconduct in Real Time

Test misconduct puts at risk a test’s validity. Interpretations of test scores of those who do and do not engage in test misconduct are at risk, especially when the rank ordering of test scores is important. Therefore to support the validity of a test, test security should be an integral part of a testing company’s standard operations.

Different types of test misconduct will present themselves differently in psychometric data and as a result, a variety of test security analyses must be employed. Additionally, when a test is pre-equated and scores can be available in seconds, analyses must be developed before the test is administered and flagging criteria applied after test completion but before scores are released to flagged candidates.
This proposal provides guidelines for implementing misconduct flags for a continuous high-volume computer-based fixed-linear-form testing program that utilizes pre-equating and returns scores the same day of test completion. Although the context is specific, the general methods can be applied or adapted to almost all other testing contexts.

The proposed methodology to flag test misconduct focuses on identifying individuals and testing sites. For individuals, flagging criteria are determined based on past data. These criteria could include: unlikely score gains as determined by high percentile growth and low test time based on time percentile rank. For testing sites, the individuals’ flags are aggregated and the likelihood of the given number of individual flags for the site calculated. If the likelihood is unusually low, the number the testing site is flagged.

The flagging criteria can be implemented within a test scoring system and flagged tests can be held for further review. The hold allows the testing program time to further review the tester data and allows for any follow up analyses that are warranted, such as response similarity analyses.

Hongling Wang
ACT, Inc.

Chi-Yu Huang
ACT, Inc.

Poster Presentation 5: Thursday, 5:30 – 7:30

A Study of Students’ Item Review Behaviors in Computer-Based Testing

Item review behavior has been a focus in test security research for a long time. In paper-pencil testing (PPT), examinees are usually allowed to review and change their item responses. Many researchers have investigated the effects of item review and answer change on examinee performance in PPT. What we see from PPT may not provide a whole picture of item review behavior due to limited information recorded on answer sheets; on the other hand, computer-based testing (CBT) can record more information on item review behavior. Since most studies were for PPT and the item review behaviors under PPT and CBT may differ, more investigation of item review behavior in CBT is needed.

In this study, we will investigate examinees' review behaviors in CBT. Click data created from examinees' mouse-clicking in CBT can give a clear picture of item review behavior because the click data contains all the information of each item visit and answer changes. We will also explore how item review information may help us monitor for misconduct. For example, comparison of the item response timestamps of an examinee pair or comparison of their answer changes may give us new clues for cheating. The click data of four subjects from a CBT assessment is used in this study.

In addition, the results of this study have practical implications for Computerized Adaptive Testing (CAT). The change of students' test-taking experience with the transition from PPT to CAT could be bigger than that with the transition from CBT to CAT. Since item review could increase the possibility of item harvesting and cheating, the debate on whether to allow item review is still unconcluded. Results of this study will provide rich information for test publishers when they consider the option of item review for administering a CAT test.

Xi Wang
Measured Progress

Wonsuk Kim
Measured Progress

Louis Roussos
Measured Progress

Research Session 7: Friday, 8:00 – 9:30

Detecting Answer Similarity Using Nonparametric Item Response Models

Answer similarity analysis is widely used in operational testing programs to check the score integrity. Many analyses focus on the agreement between two examinees' response vectors after accounting for their ability levels. Unusually high response similarity suggests a violation of independent test-taking behavior, which could be caused by examinees copying from each other or test administer tampering.
Many answer similarity indices are based on a known response model, such as the ω index (Wollack, 1997), generalized binomial test (van der Linden & Sotaridona, 2006), and M4 statistic (Maynes, 2014). These model-based indices require estimating the probability of choosing each response option, which is typically estimated with a nominal response model (NRM). Based on our experiences, the NRM estimation is sometimes unstable: The estimation either does not reach a converged solution, or gives unreasonably large parameter estimates for low-discriminating items. Even if stable estimation is obtained, the model fit may sometimes be unsatisfactory. To overcome these problems, we propose to use nonparametric item response models to calculate the response probabilities. Nonparametric estimation is more flexible as it does not assume a functional form for the item characteristic curves (ICC). Douglas (1997) has proved that under mild assumptions, the smoothed ICC estimates could converge to their true values.

A series of simulation studies are conducted to evaluate the type-I error and power of two well-known indices, the ω index and the generalized binomial test, when nonparametric estimation is used. The detection is evaluated both at the individual pair level and at the group level. For pair-level evaluation, a source-copier pair is generated by considering different ability levels of the source and copier. For group-level evaluation, groups with different ability distributions and sample sizes are generated. The study results will have implications on the feasibility of using nonparametric estimation in answer similarity detection.

Susan Weaver
Caveon Test Security

John Sowles
Ericsson

Diane Long
OKTA, Inc.

Panel Presentation 11: Thursday, 4:15 – 5:15

When Test Items Fight Back

Let's face it, test items have been bullied for years. Test takers guess their correct responses. Test items are stolen, snap-chatted, and shared with others. Far worse, many test items are sold on the black market to make a profit on their good works. Simply, the test item has been taken advantage of for far too long.

IT Certification programs are no strangers to having their test items bullied. They have experienced the heartbreak of proffered item banks and the hard work of having to rebuild and replace a certification test. But no more. It's time to fight back. IT certification programs are now implementing the DOMC (Discrete Option Multiple Choice) item type. This item format is a rising star to restore the good name of test items. Its powerful format protects it from exposure and more accurately measures an individual's competence, not just their ability to memorize item content from a cheat sheet.

Join three IT certification testing programs as they discuss their implementation and use of DOMC items within their programs. They will discuss why they are choosing to implement DOMC in their certification programs, the benefits they hope to gain, and how they believe DOMC will protect their exam content from senseless item bullying and the bullies who try to steal it.

Marc Weinstein
Caveon Test Security

Walt Drane
Mississippi Department of Education

Demonstration Presentation 4: Friday, 8:00 – 9:30

Educator Coaching and Student Response Interference in K-12 Assessment Administrations--What It Looks Like and How to Detect and Stop It

Educator coaching and interference in student responses during assessment administrations continue to undermine the validity of test results used by states and districts to measure student achievement and school performance. Despite the transition of many assessment administration platforms from primarily paper and pencil tests to online computer-based tests, some educators continue to administer and proctor assessments in ways that influence student responses. Although educator conduct that influences student responses is sometimes intentional, as in coaching, it is sometimes the result of unintentional administration and proctoring practices. Investigations of testing irregularities have revealed that educators sometimes engage in conduct during assessment administrations that causes students to think or suspect that the educator is sending 'signals' that the student has the wrong answer or should otherwise modify a response to a particular item. This phenomenon has been demonstrated by statistical evidence that is explained by student statements during interviews about what happened during the assessment administration that could have caused statistical flags. This workshop will draw on these data and experiences to demonstrate what coaching and response interference looks like during actual assessment administrations, how it can be detected by monitors or auditors during assessment administrations and statistical analysis of assessment response data following testing. Finally, the presenters will suggest that assessment sponsors and stakeholders rethink policies and guidance for test administration practices to mitigate the threat posed by coaching and response interference. The presentation will include live demonstrations of actual coaching and response interference based on examples of documented incidents, as well as administration practices that would better minimize these threats.

Marc Weinstein
Caveon Test Security

Benjamin Hunter
Caveon Test Security

Poster Presentation 8: Thursday, 5:30 – 7:30

Enhanced Assessment Monitoring – Leveraging Technology to Streamline Monitoring Processes, Manage Data Flow, and Report Useful Results in Real Time.

Whether through formal monitoring, site observations, assessment audits, or another mechanism, one of the foundations of a valid assessment and the delivery of useful test results is a robust assessment monitoring program.

Although “monitoring” can take many forms, physical on-site monitoring provides multiple angles from which to attack problems, whether it be stopping a bad actor before it becomes a widespread breach, preventing malfeasance simply by being present, or providing useful feedback to stop a future issue from ever evolving.

With limited resources, organizations are increasingly relying on technology to do the heavy lifting in storing, filtering, segmenting, retrieving, and analyzing data. This session will provide attendees with recommendations for systems requirements for their monitoring tools, suggest potential configuration options, and will make the presenters available to discuss their experience in assessment monitoring and technology configuration for monitoring purposes.

Marc Weinstein
Caveon Test Security

Thomas Gera
The Enrollment Management Association

Panel Presentation 9: Thursday, 4:15 – 5:15

Champagne Test Security on a Beer Budget--How Smaller Organizations Can Develop and Maintain a Holistic Test Security Program Despite Limited Staffing and Budgets

Smaller organizations that sponsor high-stakes tests face many unique challenges in protecting the integrity of their programs, not the least of which are limited staffing and budgets. Learn how one such organization developed and implemented a holistic test security program over a three-year period that has completely transformed its security policies and practices. The organization's evolution will be presented as a case study from which participants can glean important lessons to consider applying to their own programs. The organization that will be the focus of the case study has a staff of fewer than fifty people, yet administers more than 80,000 paper-based tests each year, at more than 700 test centers located in more than 75 countries throughout the world. Participants will learn how the organization assessed the greatest vulnerabilities and threats to its program and then identified and prioritized corresponding security solutions to develop a holistic, comprehensive program to deter, detect and respond to incidents that could threaten the validity of test results and the important decisions made based upon those results. The solutions adopted by the organization included changes to policies, legal agreements, communications to stakeholders, revised training for staff and partners, along with statistical analysis of test response data, Internet searches, announced and unannounced monitoring of test administrations, and many other practices designed to increase security and reduce risk all while not breaking the bank. The presentation will be followed by a question and answer session to allow for discussion of the issues with participants.

William Wells
NCS Pearson

Research Session 8: Friday, 11:00 – 12:00

Integrating Compliance into Information Security Programs

Traditional information security programs focus on confidentiality, integrity, and availability data. And while there is usually a tacit nod toward the need for privacy and compliance, the programs themselves often leave privacy and compliance to the legal and internal audit teams. In so doing, they put their companies at greater risk of non-compliance by not building these skillsets in-house. This presentation focuses on the types of skills and areas of knowledge that need to be integrated into information security programs to provide a more fulsome and holistic set of services to their organizations.

William Wells
NCS Pearson

Research Session 8: Friday, 11:00 – 12:00

Measuring Information Security Risk in Quantitative Terms

Information security used to be a discipline that focused almost exclusively on technical controls designed to protect data from being accessed without authorization. Today, however, information security has become a discipline that is increasingly more focused on measuring the risks posed to the security of information. Measuring those risks, particularly now that the control environment typically includes administrative and physical controls, can be particularly challenging for information security organizations whose focus has been primarily on whether or not (yes or no) technical controls were in place. This presentation describes a methodology for describing information security risks in quantitative terms, which can then be used to present those risks to management stakeholders in clear and meaningful ways.

William Wells
NCS Pearson

Research Session 8: Friday, 11:00 – 12:00

US Student Data Privacy Compliance Landscape

Increasingly, States' Departments of Education are upping the ante for student data privacy. States like Colorado, California, and Arizona are paving the way in laying down more stringent requirements for the security of student data. Companies servicing these jurisdictions need to be aware of this changing landscape in order to both prepare for and adhere to these emerging requirements. The presentation provides an overview of the student data privacy landscape and opines on the future direction of this intensifying compliance landscape.

James Wollack
University of Wisconsin – Madison

Rachel Hample
Temple University

Jarret Dyer
College of DuPage

Panel Presentation 1: Thursday, 9:45 – 10:45

Improving and Streamlining Proctor Training Through National Proctor Certification: An Initiative of the National College Testing Association

Thoroughly trained proctors are critical to maintaining test standardization and protecting exam security. Currently, every testing program develops its own training and proctors are required to separately become certified for each unique test they administer. While it is necessary for proctors to gain familiarity with certain aspects of each program prior to administering the test, many of the core elements and underlying principles are identical across programs. However, because programs currently have no way of ensuring that proctors possess the requisite knowledge, they have no choice but to develop their own extensive training materials, and proctors have no choice but to undergo lengthy certification and re-certification processes for each program.

The National College Testing Association (NCTA) has recently begun working on a proctor certification process that aims to ensure that certified individuals are well versed in industry best practices for test administration, test security principles and strategies, and approaches to best manage the nuances of interpersonal dynamics in highly stressful testing environments. Come learn about the value of certification (to proctors, test centers, and testing programs) and NCTA's vision for how test certification might help shape the testing landscape. We will discuss our work to date on this project, including what a certified proctor might be expected to know, and discuss next steps. We will conclude with an open forum for participants to ask questions and share their thoughts about the concept of proctor certification and how to best build support within the testing industry.

Tong Wu
University of Illinois at Urbana-Champaign

Anqi Li
University of Illinois at Urbana-Champaign

Hua-Hua Chang
University of Illinois at Urbana-Champaign

Poster Presentation 15: Thursday, 5:30 – 7:30

Empirical Study for Item Bank Replenishment of Computerized Adaptive Testing

Computerized Adaptive Testing (CAT) is an important mode of testing in educational assessment. With the development of technology, test security of CAT has been greatly threatened. Therefore, it has received attention from different perspectives including item selection, cheating detection and online calibration (Chang & Ying, 1999; Chen, Wang, Xin, & Chang, 2017; Guo, 2016; Wang, Xu, & Shang, 2016; Zhang & Li, 2016; Zheng, 2014). As the item pool has been repeatedly used in CAT, pool replenishment becomes a necessary process to maintain an item pool because of item parameter drift and item overexposure (Ban, Hanson, Wang, Yi, & Harris, 2001). Most previous studies on item pool focus on fully usage of item pool and pretest items while very few of them relate to item pool replenishment (Hau & Chang, 2001). In this study, assuming the pretest items have been calibrated, we will focus on item bank replenishment method of CAT.

In this study, we apply the item bank replenishing method in a real Chinese language test, which is a 45-item test including reading and listening sections. The real item pool with 1445 items is applied. 1000 examinees following standard normal distribution are simulated for each test administration. The item pool is replenished every time after a test is administered. The item replenishment rule is to replace items with exposure rates greater than 0.21 and to maintain the item structure of parameter *a* in a certain range and improve the difficult level of parameter *b*. The simulation study shows that the highest item exposure rate decreases and there is around 50% decrease in number of items with exposure rates larger than 0.3. Therefore, this item bank replenishment method in CAT could decrease the item exposure rates of high quality items and improve test security of the item pool.

Xinhui Xiong
AICPA

Anqi Li
University of Illinois Urbana-Champaign

Research Session 9: Friday, 1:00 – 2:00

Strengthened Scale-Purified Deterministic Gated Item Response Theory Model

Test security is one of major concerns in the testing field, especially for high-stake tests. “Threats to test security are omnipresent” (G. Cizek and J. Wollack, 2017). Though it is crucial to prevent items from being compromised after tests are designed and constructed, identifying compromised items accurately in time is equally important to ensure tests validity. Generally, detection methods fall in four categories (G. Cizek and J. Wollack, 2017): (1) detect examinees with pre-knowledge at individual level; (2) detect compromised items; (3) detect both examinees with pre-knowledge at individual level and compromised items; (4) detect examinees with pre-knowledge at group level. This study will propose a new method, strengthened scale-purified deterministic gated Item response theory model (DGM), to detect examinees preknowlege at individual level and compromised items for a licensure test. Shu et al.’s (2013) simulation study showed the sensitivity and specificity with DGM method. Eckerly et al. (2015) proposed a scale-purified DGM to minimize the scale drift issue and showed improved results compared with results using DGM method. The proposed new method, based on the scale-purified DGM method, is to further minimize the scale drift issue in order to obtain a more accurate detection results. Real item and examinee data from a licensure test will be used.

Jing Yang
Northeast Normal University

Liwen Huang
University of Illinois at Urbana-Champaign

Leanne Zeng
University of Illinois at Urbana-Champaign

Hua-Hua Chang
University of Illinois at Urbana-Champaign

Poster Presentation 14: Thursday, 5:30 – 7:30

An Item Selection Design that Optimizes Item Bank Usage and Estimation

Item selection, as the most important component of computerized adaptive testing (Cheng & Chang, 2009), aims at maximizing testing efficiency while maintaining test security (Cheng & Chang, 2009; Davey & Parshall, 1995). Nonetheless, striking a balance between measurement efficiency and test security is challenging. Cheng and Chang (2009) proposed Maximized Priority Index (MPI), an item selection method that excels at exposure control, content balancing and estimation accuracy, but underuses the item pool. In order to overcome this limitation, they further proposed *b*-matching as an adaptation of MPI (Cheng Y., Chang, Douglas, & Guo, 2009). However, this method was yet to be verified by simulation studies. In this operational project, we implemented *b*-matching with an ascending *a*-stratification on an empirical item bank in a cloud-based system for one of the world's largest language proficiency tests, the Chinese Proficiency test (HSK).Through simulation studies, we compared *b*-matching's performance with that of MPI, evaluating them in terms of indices of constraint management, exposure rate and measurement accuracy. Results show that *b*-matching optimizes item bank usage while retaining good estimate accuracy and exposure control, compare to MPI.

Xiaofeng Yu
University of Notre Dame

Ying (Alison) Cheng
University of Notre Dame

Research Session 6: Friday, 8:00 – 9:30

Using Change Point Analysis to Detect Inattentiveness in Polytomous Survey Response Data

Carelessness or inattentiveness is one of the most encountered aberrant behavior in survey responses which can distort the test score and lead to erroneous conclusions. Detection of carelessness is therefore very important and crucial. Change point analysis (CPA) is a statistical process control method which can not only be applied to detect unusual response pattern, but also to identify the position from which a participant starts to respond in an inattentive fashion. This paper evaluates the performance of CPA procedures to detect inattentiveness in survey data using simulation studies. Results showed that when item parameters were known, CPA methods resulted in high power while keep the type-I error low. When item parameters were unknown, the partial removal method (i.e., removing only the responses flagged as inattentive) and the complete removal method (i.e., listwise deletion if some responses from a participant are flagged as inattentive) were compared. Their implication on power, latent trait estimation, item parameter estimation, reliability and other psychometric properties were investigated.

Jin Zhang
ACT, Inc.

Research Session 3: Thursday, 1:00 – 2:30

A Comparison of Pre-knowledge Detection Procedures in CAT with Response Time Modeling

Two Bayesian procedures based on a lognormal response time model are compared in detecting response time patterns indicating pre-knowledge. A simulation study is conducted to investigate the effectiveness of the methods in conditions where proportions of items and persons affected by pre-knowledge varied.

Mengyao Zhang
National Conference of Bar Examiners

Joanne Kane
National Conference of Bar Examiners

Poster Presentation 10: Thursday, 5:30 – 7:30

Graphical Imaging Methods for Detecting Potential Collusion for Test Centers with Unusual Score Gains

Cheating on standardized tests erodes the validity and fairness of scores and compromises decisions based on those scores. Cheating can take many forms, including individuals acting alone or colluding with others. In this study, we explore different graphical imaging approaches to detecting potential collusion among examinees sitting for a national licensing examination at different test centers that were triggered because the mean scores at the test centers showed unusual score gains (+ 1 SD). We use plots and other imaging approaches for several indices to depict unusual pairwise response similarity, such as GBT (van der Linden & Sotaridona, 2006) and M4 (Maynes, 2014) indices based on item response theory (IRT), and IR (i.e., number of identical responses) and IR-LCS (i.e., longest consecutive sequence of items with identical responses) indices relying on empirical null distributions. Because center-level collusion often involves multiple examinees, the cluster analysis approach suggested by Wollack and Maynes (2017) combining different similarity indices is employed, in order to detect potential collusion within and across centers. We also use heatmaps and dendrograms as graphical representations that could be useful for conveying test collusion results to a broader group of audiences who may or may not fully understand the underlying mathematics.

Mengyao Zhang
National Conference of Bar Examiners

Joanne Kane
National Conference of Bar Examiners

Research Session 5: Thursday, 4:15 – 5:15

Detection of Potential Test Collusion across Multiple Examinees: A Real-World Example

This project explores methods of detecting potential collusion in an operational, real-world context. Whereas many previous studies focused on identifying collusion for exploratory or screening purposes, we sought to apply a battery of methods in a confirmatory context. Further, whereas many currently available methods of cheating detection do not include information about the likely interrelationships among suspects, in our real-world context we had such information and sought to capitalize upon it both within the similarity analyses and by supplementing those analyses with additional investigation.

Several similarity indices are currently available for identifying collusion between two examinees. Fewer approaches to detecting collusion across multiple examinees are available, though a new approach based on cluster analysis (Wollack & Maynes, 2017) seems to be a promising tool for flagging multiple examinees with similar responses. These statistics can differ in terms of the theoretical framework, the reference distribution, and both the power and control of the Type I error rate. We took a multi-pronged approach, including both these pair-level and group-level approaches.

In our investigation, one examinee (primary suspect) was suspected of colluding with other examinees (secondary suspects) at his test site. We first compared responses between the primary and secondary suspects using several similarity statistics at the pair level, e.g., IR and IR-LCS suggested by Hanson et al. (1987), and M4 (Maynes, 2014). Next, we employed the cluster analysis approach to identifying groups of secondary suspects with similar responses.

Our study represents a practical real-world application of a portfolio of similarity measures which, taken together, could provide compelling and actionable evidence of collusion.

Yu Zhang
Federation of State Boards of Physical Therapy

Jiyoon Park
Federation of State Boards of Physical Therapy

Aijun Wang
Federation of State Boards of Physical Therapy

Lorin Mueller
Federation of State Boards of Physical Therapy

Poster Presentation 7: Thursday, 5:30 – 7:30

Using Candidate Clusters to Identify Potentially Compromised Items

The approach described in this poster was developed to identify items that were potentially shared by candidates. This approach starts with identifying clusters of candidates who show an unusually high degree of similarity in responses to test items through applying similarity index with additional information. The following procedures split items to re-estimate ability and redefine clusters based on the stability of ability estimates. Finally, items that the clusters have advantage in answering them correctly are considered as potentially compromised.

2017 Conference Volunteers

COTS Executive Committee

Mark Albanese	National Conference of Bar Examiners
Kim Brunnert	Elsevier
Dave Foster	Caveon Test Security
Richelle Gruber	Caveon Test Security
Rachel Schoenig	Cornerstone Strategies, LLC
Sonya Sedivy	University of Wisconsin
Billy Skorupski	University of Kansas
Jim Wollack	University of Wisconsin

Program Committee

Mark Albanese	National Conference of Bar Examiners
Carol Eckerly	Alpine Testing Solutions
Seo Young Lee	University of Wisconsin
Andy Mroch	National Conference of Bar Examiners
Sonya Sedivy	University of Wisconsin
Jim Wollack	University of Wisconsin

Administration

Mark Schroeder	University of Wisconsin

Registration Table

Mandy Fortney	University of Wisconsin

Webmaster

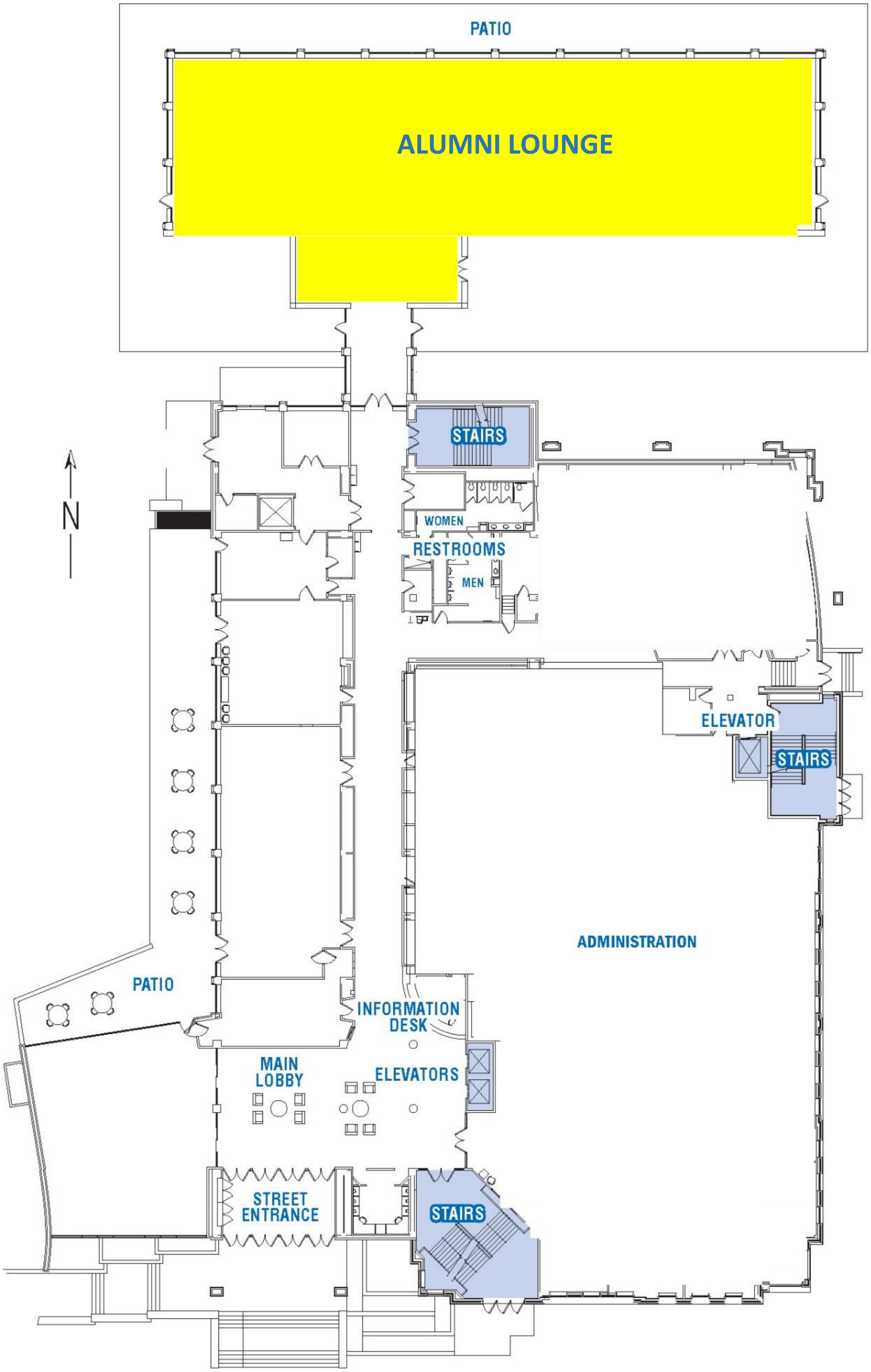
Clay Larson	University of Kansas

Special Thanks to…

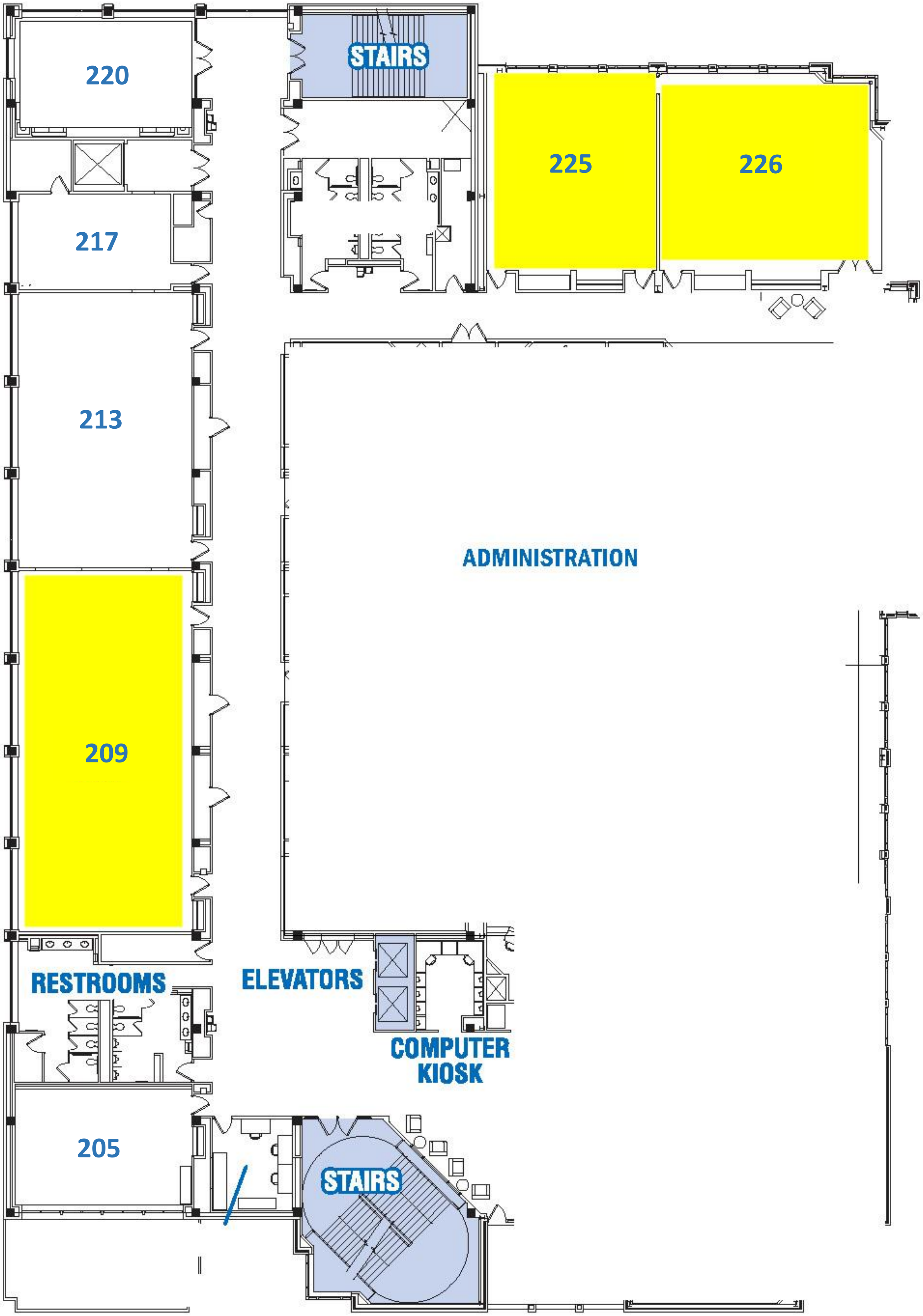
Pearson Education for the use of their abstract collection system, conference registration system, and conference app, and to Deandrea White, Mary Beth Hayes, and their team at Pearson for many hours spent updating and customizing these various systems for our use.

THE PYLE CENTER FLOOR MAPS

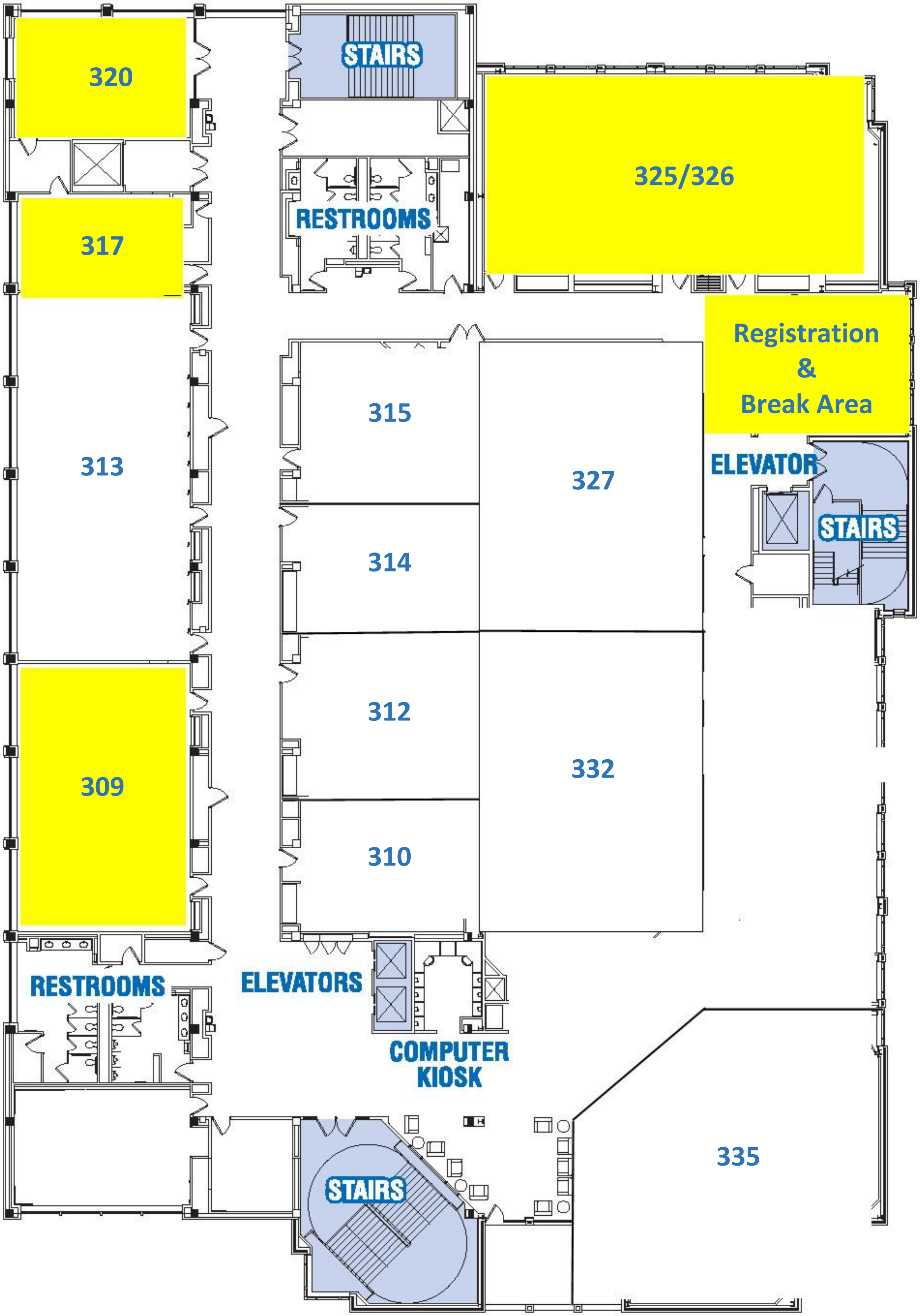
Pyle Center-Floor 1



Pyle Center-Floor 2



Pyle Center-Floor 3



NOTES

Save the Date!

The Conference on Test Security

October 10-12, 2018

Grand Summit Hotel - Park City, Utah

Hosted by Caveon Test Security

caveon™
Test Security