2014 Conference on Test Security

Conference Program

September 30 – October 2, 2014
Iowa City, Iowa
ACT Main Campus, Ferguson Center

2014 Conference Hosts

ACT®

caveon™
Test Security

KU THE CENTER FOR EDUCATIONAL TESTING & EVALUATION
The University of Kansas

UNIVERSITY OF WISCONSIN SYSTEM
CENTER FOR PLACEMENT TESTING

Conference Sponsors

AICPA®

Alpine
Testing Solutions

ATP ATPSC
Security Committee

CollegeBoard

Dilworth Paxson LLP

ETS®
Listening. Learning. Leading.®

GMAC®
GRADUATE MANAGEMENT ADMISSION COUNCIL

National College Testing Association
NCTA

2014 Conference on Test Security
September 30 – October 2, 2014
Iowa City, Iowa

Tuesday, September 30

5:00 – 6:00 p.m. *Check-In/Registration* *Sheraton Iowa City Hotel, Lower Level Lobby*

6:00 – 8:00 p.m. *Welcoming Reception* *Sheraton Iowa City Hotel, Dean AB*

Wednesday, October 1

8:00 – 8:30 a.m. *Check-In/Registration* *Ferguson Center Lower Level (LL) Lobby*

8:30 – 9:30 a.m. *Welcome and Keynote Address* *Full Ferguson LL Conf Room*

Keynote Address: Getting Along Without a Truth Machine
Max Shron, Data Strategist and Founder, Polynumeral

Statistical methods for ensuring test security have existed for decades, but new technologies and methodologies for testing have proliferated the amount of data available to work with. This trend, of increasing volumes and varieties of data, is occurring across every industry. How do we orient ourselves in this deluge, and more importantly, how do we make the right use of it? Max Shron, author of *Thinking with Data* and founder of Polynumeral, a premier data science consultancy, will discuss the ways that other disciplines have dealt with these changes and how embracing a nuanced view of statistics, automation, and data analysis can put each in their proper place.

9:30 – 10:30 a.m. *Session 1: Plenary Session* *Full Ferguson LL Conf Room*

How Has Our Approach to Test Security Evolved and Where Are We Headed?
John Fremer, Caveon Test Security
Wayne Camara, ACT
Ray Nicosia, Educational Testing Service
Phil Dickison, National Council of State Boards of Nursing

10:30 – 11:00 a.m. **BREAK**

11:00 a.m. – 12:30 p.m. **Session 2:** Tracked Sessions

Session 2A: *Peer Review Session — Response Analysis* *Ferguson LL Conf B*

Using Response Times to Detect Test Fraud
Wim J. van der Linden, CTB/McGraw-Hill Education

Analysis of Answer Changes via Kullback-Leibler Divergence
Dmitry I. Belov, Law School Admission Council

Empirical Cut-Offs Using Item-Response Probability in Detecting Aberrant Item Responses
Kyoungwon Lee Bishop, Pearson
Brett Chaney, ACT

Discussion of Session
Deborah J. Harris, ACT

Session 2B: ***Practitioner Session — Test Security Planning*** *Ferguson LL Conf A*

Protecting Our Tests: The Past, Present, and Future of Test Security
Steve Addicott, Caveon Test Security
Rachel R. Watkins Schoenig, ACT
Walt Drane, Mississippi Department of Education

Test Security: Building a Comprehensive Framework
Ray Yan, Financial Industry Regulatory Authority (FINRA)
Barbara Graham, Financial Industry Regulatory Authority (FINRA)

Session 2C: ***Practitioner Session*** *Ferguson LL Conf C*

Hercule Poirot, Sherlock Holmes, Nancy Drew: Multiple Detectives Investigate Invalid Scores
Kim Brunnert, Elsevier
Dennis Maynes, Caveon Test Security
William Skorupski, University of Kansas
Keke Lai, University of California, Merced
Sarah Thomas, University of Virginia
J. Patrick Meyer, University of Virginia

12:30 – 1:45 p.m. **LUNCH**

1:45 – 3:15 p.m. **Session 3:** Tracked Sessions

Session 3A: ***Peer Review Session — Mixed-Format Testing*** *Ferguson LL Conf B*

Examining Individual and Cluster Test Irregularities in Mixed-Format Testing
Xin Li, ACT
Chi-Yu Huang, ACT
Deborah J. Harris, ACT

Impact and Detection of Gaming in Constructed-Response Items
Vincent Kieftenbeld, CTB/McGraw-Hill Education

Discussion of Session
Dennis Maynes, Caveon Test Security

Session 3B: *Practitioner Session — Item Protection & Analysis* *Ferguson LL Conf A*

Protecting Item Content via the Discrete-Option Multiple-Choice Item Type
Gail C. Tiemann, University of Kansas
Harold Miller, Brigham Young University
Neal M. Kingston, University of Kansas
David Foster, Caveon Test Security

Improving Exam Security with Performance Items
Jill Burroughs, Alpine Testing Solutions
Russell Smith, Alpine Testing Solutions
Patrick Irwin, Cisco Systems

Session 3C: *Practitioner Session — Data Analytics & Web Monitoring* *Ferguson LL Conf C*

Data Analytics — Identifying Cheaters Before They Cheat
Patrick Watts, CFA Institute

Evaluating Return on Investment (ROI) For Your Online Test Security Procedures
Jamie Mulkey, Caveon Test Security
Christy Frederes, Ascend Learning

3:15 – 3:45 p.m. **BREAK**

3:45 – 5:15 p.m. **Session 4:** Tracked Sessions

Session 4A: *Practitioner Session — State Assessments* *Ferguson LL Conf B*

Data Forensics Methods for the New Generation of State Summative Assessments
Jessalyn Smith, CTB/McGraw-Hill Education
Litong Zhang, CTB/McGraw-Hill Education
Furong Gao, CTB/McGraw-Hill Education

The Use of Growth Norms in K–12 Data Forensics Studies
Xin Lucy Liu, Data Recognition Corporation
Mayuko Simon, Data Recognition Corporation
Marc Julian, Data Recognition Corporation

Session 4B: *Practitioner Session* *Ferguson LL Conf A*

Follow-Up Study of an Empirical Method for the Detection of Potential Test Fraud
John Weiner, PSI Services LLC
Amin Saiar, PSI Services LLC
Greg Hurtz, PSI Services LLC

Answer Changing Patterns in Computer-Based Tests
Ardeshir Geranpayeh, Cambridge English
Language Assessment

Session 4C: *Practitioner Session —*
Strategies for Planning and Audits *Ferguson LL Conf C*

Building a Test Center Audit Program
Patrick Watts, CFA Institute
Heather Mullen, CFA Institute

"If You Fail to Plan, You Are Planning to Fail!" — Strategies and Techniques to Develop Your Test Security Plan
Joe Brutsche, Pearson VUE
Aimée Hobby Rhodes, CFA Institute
Michael Clifton, ACT

Thursday, October 2

8:30 – 10:00 a.m. **Session 5:** Tracked Sessions

Session 5A: *Peer Review Session —*
Protection vs. Exposure of Items *Ferguson LL Conf B*

Employing Statistics and Publication Countermeasures To Inoculate Exams Against Braindump Users
Dennis Maynes, Caveon Test Security

Using Answer Key Imputation for Making Bayesian Inference about Cheating on Tests
Marcus Scott, Caveon Test Security

The Impact of Compromised Items in CAT on Score Estimates
Qing Yi, ACT
Meichu Fan, ACT

Discussion of Session
Neal M. Kingston, University of Kansas

Session 5B: *Practitioner Session —*
Response Analysis *Ferguson LL Conf A*

Autonomous Detection of Cheating in the Presence of Aberrant Responses in OMR Documents
Raghuveer Kanneganti, CTB/McGraw-Hill Education
Randy Fry, CTB/McGraw-Hill Education (Retd.)
Lalit Gupta, Southern Illinois University Carbondale
Wim J. van der Linden, CTB/McGraw-Hill Education

Is Response Time a Good Indicator of Aberrance
Xiao Luo, National Council of State Boards of Nursing
Doyoung Kim, National Council of State Boards of Nursing

Session 5C:

Practitioner Session —
Test Security Investigation Techniques

Ferguson LL Conf C

Working a Case: Best Practices in Conducting Exam Integrity Investigations
A. Benjamin Mannes, American Board of Internal Medicine
Rachel R. Watkins Schoenig, ACT
John Fremer, Caveon Test Security
Marc Weinstein, Dilworth Paxson LLP

10:00 – 10:15 a.m. **BREAK**

10:15 – 11:45 a.m. **Session 6:** Tracked Sessions

Session 6A:

Peer Review Session —
Copying

Ferguson LL Conf B

A Comparison of Similarity Indexes in Detecting Copying in Cases with Significant Observational Evidence on the Multistate Bar Examination
Mark A. Albanese, National Conference of Bar Examiners
Cory Tracy, University of Wisconsin–Madison

Detecting and Reporting Test Irregularities in Online Testing
E. Matthew Schulz, Pacific Metrics, Inc.
Xiujuan Yuan, Louisiana Department of Education

Discussion of Session
Andrew Mroch, ACT

Session 6B:

Practitioner Session —
Test Security Process & Framework

Ferguson LL Conf A

Measuring the Effectiveness of Your Test Security Program
Rachel R. Watkins Schoenig, ACT
Ray Nicosia, Educational Testing Service

Responding to Emerging Security Threats
Aimée Hobby Rhodes, CFA Institute
Camille Thompson, ACT
Joe Brutsche, Pearson VUE

Session 6C:

Practitioner Session —
Selecting Data Forensics

Ferguson LL Conf C

Selecting the Right Data Forensics Analyses for Your Program: A Review of the Literature
J. Carl Setzer, GED Testing Service
Kathleen A. Gialluca, Pearson VUE

Analysis of Online Answer Change Behavior with Survival Analysis Model to Detect Aberrant Behavior
Mayuko Simon, Data Recognition Corporation

11:45 a.m. – 1:00 p.m. **LUNCH**

1:00 – 2:30 p.m. **Session 7:** Tracked Sessions

Session 7A: ***Practitioner Session — Test Security Investigation Techniques*** *Ferguson LL Conf A/B*

So You Flagged a Cheater, Now What?
Aimée Hobby Rhodes, CFA Institute
Lorin Mueller, Federation of State Boards of Physical Therapy
Kellie Early, National Conference of Bar Examiners
A. Benjamin Mannes, American Board of Internal Medicine
Joy Matthews-Lopez, National Association of Boards of Pharmacy

Session 7B: ***Practitioner Session*** *Ferguson LL Conf C*

Things that Go Bump in the Night: What Should We Be Worried About?
David Foster, Caveon Test Security

A Framework for Policies and Practices to Improve Test Security Programs: PDIR
Steve Ferrara, Pearson

2:30 – 2:45 p.m. **BREAK**

2:45 – 3:45 p.m. ***Demo/Poster Session*** *Ferguson Main Dining Center*

Detecting Answer Copying in an Assessment When Multiple Forms Are Administered
Chi-Yu Huang, ACT
NooRee Huh, ACT
Hongling Wang, ACT
Qing Xie, University of Iowa

Discrete Option Multiple Choice: Preventing Cheating and Test Theft
David Foster, Caveon Test Security

Simple Exploration of Answer-Changing Behavior Relative to Item Response Time and Difficulty — TAD Categorization
Djibril Liassou, Data Recognition Corporation
Vincent Primoli, Data Recognition Corporation

CESP and me — Why Should I Get Certified as an Exam Security Professional?
Jamie Mulkey, Institute for Exam Security

A Practitioner's Approach to Improving Test Security
Jason Taylor, Project Lead the Way
Karoline Jarr, Project Lead the Way
Claudia Guerere, Project Lead the Way
Kristin Donlon, Project Lead the Way

Test Security: Not Just a "Bolt-On"
Jennifer Geraets, ACT

Graphical Representations of Test Fraud
Jennifer Lawlor, Law School Admission Council
Peter Pashley, Law School Admission Council

Testing the Efficacy of CleanSlate Cheating Prevention Paper
Max Brickman, CleanSlate
Hugh Watson, CleanSlate
Haley Gedek, CleanSlate

Legal Defense of Score Cancellation Decisions
Michael Clifton, ACT

Sticking to the (Investigation) Plan: Effectively Documenting and Monitoring your Test Security Investigations
Mikel Trevitt, ACT

Who's Cheating Who — A Look at Modern Cheating and Stealing Tactics
Christy Frederes, Ascend Learning
Tara Miller, Ascend Learning

Practical Ways to Enhance Test Security: Messaging to Candidates and Stakeholders
Chuck Friedman, Professional Examination Service
Rory McCorkle, International Credentialing Associates

The Development of an R Package dataForensic for Conducting of Test Security Analysis
Jiyoon Park, Federation of State Boards of Physical Therapy
Yu Zhang, Federation of State Boards of Physical Therapy

Disrupted Opportunity Analysis (DOA): A System for Detecting Unusual Similarity Between a Suspected Copier and a Source
Mark A. Albanese, National Conference of Bar Examiners
Cory Tracy, University of Wisconsin–Madison

4:00 – 5:00 p.m. *Keynote Address and Closing Remarks* *Full Ferguson LL Conf Room*

Keynote Address: Screen is NOT Paper. Story is NOT History. Visualization, Data, Analysis, and Other Hurdles
Nahum Gershon, Senior Principal Scientist,
The MITRE Corporation

Technology has enabled us to do a plethora of things we could not do before. We can represent data and information, for example, in ways unimagined before, not only in words or simple images, but also in automatically generated complex visual representations. This has enabled developers and other technically proficient people to generate by themselves visual and other representations of their data in a DIY fashion. However, making these representations effective frequently requires visual, design, representational, and analytical literacies. This poses a great challenge to many technology (and other) users. After all, the pencil inventor was not necessarily the best artist. This talk will focus on this challenge and chart some potential ways to overcome it.

Abstracts & Presentation Summaries

Opening Keynote Address

Getting Along Without a Truth Machine

Presenter: *Max Shron*

Statistical methods for ensuring test security have existed for decades, but new technologies and methodologies for testing have proliferated the amount of data available to work with. This trend, of increasing volumes and varieties of data, is occurring across every industry. How do we orient ourselves in this deluge, and more importantly, how do we make the right use of it? Max Shron, author of Thinking with Data and founder of Polynumeral, a premier data science consultancy, will discuss the ways that other disciplines have dealt with these changes and how embracing a nuanced view of statistics, automation, and data analysis can put each in their proper place.

Session 1: Plenary Session

How Has Our Approach to Test Security Evolved and Where Are We Headed?

Presenters: John Fremer, Wayne Camara, Ray Nicosia, Phil Dickison

In this session, the impact of four forces on test security practices will be reviewed.

- Standards—The influence of the AERA/APA/NCME and other formal standards. This presentation will be given by a measurement professional who has played a key role in developing and promoting testing standards in both a professional association and testing company settings.
- Move to CBT—The introduction of computer-based tests both in the United States and internationally. This presentation will be given by a measurement professional representing one of the first major testing programs to adopt computer-based testing as its basic method of delivery. The organization has been in the forefront of research and practice on maintaining test validity despite the challenges posed by testing misbehaviors.
- Development of Test Security Units within testing agencies—How have those roles developed and changed over time? This presentation will be given by a testing professional with long experience managing test security for a very large testing agency, but also one who is frequently called on as a spokesperson for test security issues with the media and government agencies.
- ormation of a Test Security Profession and Industry—How the concept of independent test security agencies has evolved and the roles they are playing. What agencies now exist and what do they do? What appears to be on the horizon?

In each instance, speakers will provide their view not just on how we got to our current status, but where we are heading and what the implications are for testing practitioners.

Session 2A: Peer Review Session — Response Analysis

Using Response Times to Detect Test Fraud

Author: Wim J. van der Linden

It is argued that response times are much more powerful to detect test fraud than responses on test items, provided the analysis is based on a model for the distribution of the response times of fixed examinees on fixed items. The claim will be illustrated for the case of the detection of collusion between examinees in group-based testing using a bivariate lognormal model for the response times. As the model has parameters for the time intensities of the items, it allows us to estimate collusion among examinees as the correlation between their response times adjusted for spuriousness due to differences in these time intensities. Without the adjustment, the procedure would have led to serious Type I errors. The effectiveness of the procedure is illustrated using empirical response-time data.

Analysis of Answer Changes via Kullback-Leibler Divergence

Author: Dmitry I. Belov

The statistical analysis of answer changes has uncovered multiple testing irregularities on large-scale assessments and is now routinely performed at testing organizations (Maynes, 2013). Recently, this analysis was deepened by intense research in different areas: modeling answer changes (Bishop, Liassou, Bulut, Seo, & Bishop, 2011; van der Linden & Jeon, 2012), fitting distributions of number of answer changes across individuals and groups (Wibowo, Sotaridona, & Hendrawan, 2013), addressing scanning errors (Wollack & Maynes, 2014), and others (Maynes, 2013).

The information about previous answers has uncertainty due to scanning error in paper-and-pencil testing (P&P) or, possibly, lengthy sequence of selected answer options before the final choice in computer-based testing (CBT), multiple-stage testing (MST), and computerized adaptive testing (CAT). Therefore, statistics used in practice (such as the number of WTR erasures) to analyze answer changes have errors and critical values from corresponding asymptotic distributions are difficult to interpret because they fit data with error; also, such methods do not model the answer-changing behavior of examinees.

This paper presents a conservative approach to analyzing answer changes on an individual level. The information about all previous answers is ignored, except for the partitioning of the response vector into two disjoint subsets: responses T_1 where an answer change did not occur, and responses T_2 where an answer change did occur. Following Belov, Pashley, Lewis, and Armstrong (2007), the idea is to compute the difference in performance between T_1 and T_2, where a size incompatibility between T_1 and T_2 is corrected. Then, for an aberrant examinee this difference should be unusually large. The new statistic is computed as follows (without loss of generality, let us assume that T_1 is larger than T_2):

1. Compute the posterior of ability $\mathbf{P}_2$ from T_2.
2. Estimate the ability $\hat{\theta}$ from T_1.
3. Sort responses $r_1, r_2,...$ from T_1 such that the corresponding items $i_1, i_2,...$have descending Fisher information at $\hat{\theta}$.
4. Add responses $r_1, r_2,...$ to a new subset T_1 until the corresponding posterior of ability $\mathbf{P}_3$ has a variance equal or smaller than the variance of $\mathbf{P}_2$.
5. Compute the difference between $\mathbf{P}_2$ and $\mathbf{P}_3$ using Kullback–Leibler divergence $\mathbf{D}(\mathbf{P}_2 \,||\, \mathbf{P}_3)$ (Kullback & Leibler, 1951).

Assuming normality of posteriors $\mathbf{P}_2$, $\mathbf{P}_3$ and no change in true ability from T_1 to T_2 (meaning no aberrancy), the distribution of $\mathbf{D}(\mathbf{P}_2 \,||\, \mathbf{P}_3)$ is chi-square with one degree of freedom, which holds for any population distribution of ability. Thus, one can perform a statistical test to detect aberrant answer changes that will be robust to multiple factors, such as sparse data, noisy data, or dependence of answer changes to score.

Answer-changing behavior was simulated using IRT, where realistic percentages of WTR, WTW, and RTW erasures were achieved and the area under the receiver operating characteristic curve was used as a measure of performance. Despite its conservatism, the presented method outperformed two popular statistics based on absolute and relative number of WTR erasures.

Belov, D. I., Pashley, P. J., Lewis, C., & Armstrong, R. D. (2007). Detecting aberrant responses with Kullback–Leibler distance. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 7–14). Tokyo: Universal Academy Press.

Bishop, N. S., Liassou, D., Bulut, O., Seo, D. G., Bishop, K. (April, 2011). *Modeling erasure behavior* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*, 79–86.

Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 173–199). New York: Routledge.

van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics*, 37, 180–199.

Wibowo, A., Sotaridona, L., & Hendrawan, I. (April, 2013). *Statistical models for flagging unusual number of wrong-to-right erasures.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wollack, J. A., & Maynes, D. (April 2014). Improving the robustness of erasure detection to scanner undercounting. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Empirical Cut-Offs Using Item-Response Probability in Detecting Aberrant Item Responses

Authors: Kyoungwon Lee Bishop, Brett Chaney

Classifying aberrant item response patterns in detecting cheating is an emerging method in test security research that needs further development. To this end, a method of using the probabilities for each possible item response vectors across the exhaustive combinatoric space of possible item responses is proposed. This algorithm finds reasonable cut-off points for unusual item response vectors, given the examinee's latent ability.

In test security studies, it was common that the decision of aberrant response item is performed based on a small set of items from an existing testing program or simulated data. However, it was uncertain in real settings how students respond in general to items if they are genuinely measuring examinees' ability. An empirical distribution for the likelihood of item responses was established for this purpose using a synthetic method: Tests were collected from various subject areas and grade levels in state testing programs. 9,500 items' responses from multiple state testing programs were collected and calibrated, concurrently. These items were composed of 190 tests with 50 items each. Each of the 190 tests generally had 1,000 examinees, with a few exceptions of no less than 700 examinees. Concurrent calibration allows all items in the same scale and this method can collect multiple item responses in each ability level. Each ability level has its own distribution of item responses, and the expected range of the reasonable item response patterns will be established in each ability level, such that this algorithm can be utilized in detecting unusual item response patterns.

Session 2B: Practitioner Session — Test Security Planning

Protecting Our Tests: The Past, Present, and Future of Test Security

Presenters: Steve Addicott, Rachel R. Watkins Schoenig, Walt Drane

In the past decade, we've witnessed incredible changes in test security—both good and bad. While technologies have created ever-greater risks to reliable test results, other innovations have empowered test program leaders with new tools to better protect them. Looking ahead, we know the challenges will not only be different, but even tougher. In order to keep up, we'll need creativity, technology, and of course, funding to invent new methods and tactics in the ongoing battle for test results that matter. This session will explore answers to the question: what will test security look like in the coming years?

Join a panel of test security veterans for a quick look back, and a glimpse into the "crystal ball" of the future of test security.

We'll explore interesting topics such as:

- How security threats have evolved in the past decade, and the most effective tactics to thwart those threats.
- Where has our industry made the biggest strides in protecting our exams? Where have we lagged?
- Looking ahead, how will security risks evolve? How can we protect our programs from these risks?
- Which new technologies will prove most damaging, and most helpful?
- Most importantly, what should our future test security plans include that they may not currently?

Test Security: Building a Comprehensive Framework

Presenters: Ray Yan, Barbara Graham

This presentation will examine FINRA's holistic approach to assessing, developing and improving test security in a complex, multi-test, client services environment. This presentation will discuss the technical and operational aspects of test security development, as well as the leadership and organizational development perspectives that foster a comprehensive test security strategy.

In October 2013, FINRA began evaluating its test security practices. To thoroughly review FINRA's test security, a practical framework of more than 20 functional areas was developed. These areas spanned the testing and continuing education lifecycle from authorship through delivery to retirement. This assessment provided critical insight into the major risks and opportunities facing the development of effective test security (e.g., test security strategy, analytics and forensics, internet content exposure, score reporting practices, vendor accountability programs, etc.).

Leadership and organizational development were also crucial to test security success. More than 75% of the identified control weaknesses were readily addressed or scheduled for mitigation during the assessment process using current resources and with little to no additional expense. The assessment provided a broad understanding of required managerial practices (e.g., project planning, roles and responsibilities, managing limited resources, teamwork and collaboration, strategic partnerships, holistic and systemic approaches, incident response, etc.). Many of these simple, yet effective strategies and practices will be shared.

Session 2C: Practitioner Session

Hercule Poirot, Sherlock Holmes, Nancy Drew: Multiple Detectives Investigate Invalid Scores

Presenters: Kim Brunnert, Dennis Maynes, William Skorupski, Keke Lai, Sarah Thomas, J. Patrick Meyer

Protecting the efficacy of a test is at the forefront of test security. When students, faculty, or proctors share information about the content of a test before or during a test the resulting score is invalid. The true score now includes an unknown but large amount of error. Identifying invalid scores is important for the utility and reputation of a test. Detecting invalid scores by using data alone (without anecdotal evidence from the testing session) is still developing. Some procedures have been showcased in the literature (e.g., Belov & Armstrong, 2011; Bolt, Wollack, & Suh, 2012; Rudner, 2009) and at previous conferences (e.g., Conference on the Statistical Detection of Potential Test Fraud, 2012; Association of Test Publishers, 2013; Conference on the Statistical Detection of Potential Test Fraud, 2013). However, the utility of these procedures can vary depending on the type of test, type of administration environment, and the type of data collected.

This session brings together types of analyses that represent directions that seem useful for analyzing one example of a high stakes, computer-based exam.

For this session invited participants will analyze the same data set that has a portion of the students possibly having pre-knowledge of the content. The goals of the analysis will be to identify any students who had previous knowledge of the content. The goal of this presentation is to assess the feasibility and strength of various methodologies for the detection of potentially fraudulent scores. There are 4 investigators (or teams of investigators) working on the data. They will come together to present their findings, the pros and cons of using their type of analysis, and recommendations or considerations for further investigations. The methodologies include Bayesian analysis, mixed modeling and/or structural equation modeling, mixture IRT, and a Classical Test Theory/IRT mixture.

Session 3A: Peer Review Session — Mixed-Format Testing

Examining Individual and Cluster Test Irregularities in Mixed-Format Testing

Authors: Xin Li, Chi-Yu Huang, Deborah J. Harris

It is an increasing common practice among testing programs to have a mixture of multiple-choice (MC) and constructed-response (CR) items to form so-called mixed-format tests for their assessments. One of the reasons is that MC items are easy, fast, and inexpensive in scoring, and can reliably measure a broad range of content, while CR items are capable for measuring complex skills such as forming a mathematical proof, writing a logically structured argument, etc. (Livingston, 2009), thus a mixed-format test could benefit from both. The purpose of this study is to examine the power of different statistical methods in detecting test irregularities when a mixed-format test is administered with examinees nested within test centers. Information obtained from MC items and compromised CR items, and as well as from CR items and compromised MC items, is used to detect individual and test center test irregularities.

Selected item parameters calibrated from NAEP 2000 mathematics are used. Data are generated using 3PLM for MC items and generalized partial credit model for CR items. To mimic the clustered structure that examinees are nested within test centers, the degree of cluster-related dependency (rICC: Koch, 1983) is considered in theta generation. Power of different statistical methods is assessed under variety of manipulated conditions, which include 2 levels of within-cluster sample size, 2 levels of cluster-related dependency, and five irregularity patterns. In addition to having examinees with test irregularities introduced, test centers are randomly selected to have test irregularities introduced, and thus different levels of test center irregularities are also considered. Fifty replicated datasets for all possible combinations of conditions are generated and analyzed.

Impact and Detection of Gaming in Constructed-Response Items

Author: Vincent Kieftenbeld

Automated scoring of constructed-response items is increasingly used in educational testing. For example, it is an integral part of the Common Core State Standards assessments developed by the two Race to the Top consortia (Smarter Balanced and PARCC). As large-scale, high-stakes assessments include more automated scoring, some examinees may attempt to take advantage of automated scoring by including construct-irrelevant material in their responses, aimed at increasing their score without being noticed ('gaming'). If undetected, gaming can pose a serious threat to score validity and raise public mistrust in automated scoring.

We will present an analysis of the potential impact of gaming in constructed-response items on scores assigned by different automated scoring systems and report on the efficacy of statistical outlier detection methods to identify gamed responses. Existing research on gaming has focused primarily on essays or writing samples. This presentation extends this line of research by also considering short-text constructed-response items in mathematics and English language arts. We investigated the susceptibility of automated scoring systems to gaming through the automated construction of gamed responses to items aligned with Common Core State Standards. The gamed responses were derived from existing student responses using combinations of four different gaming strategies: repeating a responses several times, adding paraphrases from item stimulus material, inserting academic words, and inserting content words. Gaming generally increased the score of low-quality responses between 0.25 and 0.5 points but lowered the score of high-quality responses. Outlier detection methods were developed in an attempt to identify gamed responses automatically based on atypical feature patterns (e.g., unusual length, repeated phrases or sentences, etc.). We will discuss implications for test security when automated scoring is used in large-scale assessments and suggest directions for future research and development.

Session 3B: Practitioner Session – Item Protection & Analysis

Protecting Item Content via the Discrete-Option Multiple-Choice Item Type

Presenters: Gail C. Tiemann, Harold Miller, Neal M. Kingston, David Foster

The computer-based Discrete-Option Multiple-Choice (DOMC) item type is a creative alternative to traditional selected-response multiple-choice (MC) testing. With DOMC, an item's stem is presented, and the answer choices are presented one at a time in a random sequence. As each choice is displayed, the examinee decides if the choice is correct. The item type has the potential to diminish the impact of coaching and reduce the feasibility of cheating based on the memorization and theft of items.

In this session, two complementary papers will report results of an empirical study of the DOMC item type's potential to thwart cheating in practical testing applications. Twenty examinees completed a 40-item high school math test consisting of items presented in both DOMC and MC formats. Ten examinees ("pirates") were asked to cheat prior to taking the test, and ten examinees ("passive cheaters") were asked to cheat after taking the test. Both groups of examinees recorded as many details as possible about the test content. Recollections were aggregated into two "cheat sheets." An additional group of 181 examinees ("cheaters") were each randomly assigned a cheat sheet and allowed to study for up to 30 minutes prior to taking the same test form that consisted of DOMC and MC items.

Paper one will present a comparative analysis of the two sets of cheat sheets utilizing an original taxonomy focusing on 1) whether a respondent's recollection was of the item, the answer, or both; 2) whether it was specific or nonspecific; and 3) whether it was correct or incorrect. The analysis will allow informal appraisal of the potential relative advantage to the test taker of access to cheat sheets.

Paper two will evaluate the usefulness of the cheat sheets in raising cheaters' test scores as well as their relative performance on DOMC and MC item types.

Improving Exam Security with Performance Items

Presenters: Jill Burroughs, Russell Smith, Patrick Irwin

Unintended advantage, regardless of source—exposure, prior knowledge, or cheating—presents persistent and pervasive threats to the validity of the interpretation and use of test scores resulting from certification exams. Research on various data forensic techniques has primarily focused on the susceptibility of select-type items to exposure, collusion, piracy, and other types of test fraud. However, the move to a computer-based delivery mode has led to the increased practice of developing and incorporating performance item types, such as simulations. The growing use of these performance type items could help to mitigate security concerns as the structure of these items demand that candidates "actually demonstrate their knowledge or skill" and allow for the incorporation of "a variety of other assessment approaches" on a singular exam (Cizek, 1999, p. 168).

There is currently a dearth of research focused specifically on the security of performance items. Some argue that these item types are memorable, while others contend that exposure has less impact as candidates still need to complete a task. This study will further existing research on data forensic techniques by investigating the use of commonly-used methodologies on performance type items as compared to selected-response item types. Moving averages, item parameter drift, and differential person functioning (DPF) will be used to investigate the relative security of performance type items versus selected-response items. This study will attempt to generalize these findings across multiple exams and programs to help practitioners determine next steps should their examinations display patterns of behaviors indicative of security issues based on item types. The presentation will also explore how to operationalize the information in this study and use it to build security into your development strategy.

Reference

Cizek, G. J. (1999). Cheating on tests: How to do it, detect it, and prevent it. Mahway, NJ: Lawrence Erlbaum

Session 3C: Practitioner Session — Data Analytics & Web Monitoring

Data Analytics — Identifying Cheaters Before They Cheat

Presenter: Patrick Watts

Cheating, in any form, drastically impacts both the reputation and efficacy of an exam. As such, it is necessary to use all of the tools at our disposal to accurately predict, monitor, and respond to problematic candidate behaviors. This session will address the role of data analytics as it applies to exam security. While analytics can not replace the role of first person observation by trained exam day staff, it can supplement and enhance the ability of a testing organization in its quest to curb cheating by 1) identifying and investigating suspicious candidates prior to the exam, 2) giving context to reports of unusual candidate behavior on exam day, 3) assisting in resource allocation based on threat levels, and 4) providing an objective threshold for determining the appropriate level of response to suspicious candidate behavior.

Evaluating Return on Investment (ROI) For Your Online Test Security Procedures

Presenters: Jamie Mulkey, Christy Frederes

What are the costs of exam exposure? Test security's goal is ensuring trustworthy test results, and protecting your investments in test development. How quickly will test development investments be eroded and how much of the test's lifetime revenue is lost if a test is exposed?

What are the risks of NOT monitoring the online landscape thoroughly? If there is online exam exposure how many candidates will gain unfair advantage? If just one exam is exposed it can be completely pirated and widely available on the web within weeks of publication without consistent web monitoring procedures and take-down efforts.

What are the costs of monitoring? Immediate and consistent actions are mandatory to prevent the spread of exposure. Can you afford to add this daily task to your plate? Or, is it more cost effective to outsource to a professional web monitoring company?

Final outcome Ultimately, after helping you answer these questions, we'll provide a system you can use in your organization to determine the ROI (Return on Investment) for the various Online Test Security options available to you.

Data Forensics Methods for the New Generation of State Summative Assessments

Presenters: Jessalyn Smith, Litong Zhang, Furong Gao

Breaches in test security are some of a countless number of threats to test validity. With more and more states implementing new assessments and adaptive administrations, it is important to examine current and new methods for detecting test security breaches at a test administration, school, or district level.

This simulation study evaluates the robustness and accuracy of several data forensics procedures that are commonly used in practice and are appropriate for analyzing the first year of a summative CAT Assessment (such as response time, response similarity analysis, etc.). Field test data from a K–12 testing program was used to determine the data structure and realistic values for all IRT item characteristics. Simulation study conditions that were varied include the number of common items between pairs of students and within a classroom, proportion of compromised items, the proportion of compromised examinees, and the test length. A state testing program was used as a model to determine a set of baseline characteristics for testing behavior.

For each case, different response time analyses and response similarity analyses were compared and evaluated to determine the most appropriate set of methods that could be used in practice. Results and recommendations will be presented to help identify any strengths and weakness of the data forensics methods, while focusing on the application of these methods in practice.

The Use of Growth Norms in K–12 Data Forensics Studies

Presenters: Xin Lucy Liu, Mayuko Simon, Marc Julian

Under the No Child Left Behind Act and the Race to the Top initiatives, teachers and principals are under intense pressure to show year-to-year academic growth within K–12 testing programs. Concurrently, incidents of test fraud have become more frequently observed at the teacher or the school level.

Within a population of students, we expect that there exists a normal growth trajectory that occurs most frequently. Students whose test scores are contaminated due to some test security violation may have a growth trajectory that differs from what is normally expected for that population across the state. In the study by Liu, Liu, and Simon (2014), the growth norm was estimated through a Bayesian polynomial model, and residuals of student observed year-to-year performance from the expected growth norm were computed, aggregated, and normalized for a school-level statistic for potential testing irregularities.

Further investigations are needed to examine the advantage or the accuracy of the growth norm approach that utilized student longitudinal data over the other approaches that utilized only two years of data. First, we propose to demonstrate this approach with real data that cover more subject areas and grade levels. Specifically, longitudinal data in reading as well as mathematics areas from both elementary and middle schools will be used. Secondly, the robustness of this approach given various degrees of cheating will be demonstrated in a simulated longitudinal data. Two factors were considered in the simulation study. The first factor is the proportion of cheating in a school (i.e., how widespread) has occurred. The second factor is the magnitude of suspicious growth (i.e., how deviant from the expected growth norm). The performance of the growth norm approach in these scenarios will be examined in comparison with the scale score change approach.

Session 4B: Practitioner Session

Follow-Up Study of an Empirical Method for the Detection of Potential Test Fraud

Presenter: Amin Saiar

This presentation examines a recently developed response similarity analysis (J2) that utilizes an empirical model to identify candidates with an abnormally high number of matching responses to other candidates (outliers), who may have formulated their responses by cheating. However, the identification of outliers using the empirical indices, such as J2, is limited by sample-dependence. Characteristics of the sample may lead to results that do not generalize. This follow-up study focuses on the comparison of response anomaly detection using both simulated (Monte Carlo) data and empirical data. The J2 method is compared to other methods of outlier detection. Implications for practice and further research are discussed.

Answer Changing Patterns in Computer-Based Tests

Presenter: Ardeshir Geranpayeh

There has been an increase in statistical cheating detection methods for answer changing in recent years. Most techniques address issues related to the paper-based answer changing using hand or optical scanner detection of changed item responses. There is little published research examining how answer changing can be detected in computer-based tests.

The purpose of this study was to explore the efficacy of combining response time with wrong-to-right and total answer-changing flagging rules. This study used a number of real data sets from a high-stakes English language proficiency test. Results show that about half of students changed one or more answers, an average of approximately 18 answer change per student. Seventy percent of students had at least five or more wrong-to-right changes. Right-to-wrong answer changes were slightly less frequent. 60% of students had at least six or more wrong-to-right answer changes. 95 percent of students had at least five or less wrong-to-wrong changes. Negative relationships were detected between total test score and total answer changes or between total test score and wrong-to-right changes. Score gain index was also computed for each examinee which shows that candidates gained fewer score as a result of answer changes. Two flagging rules (4 and 8) based on wrong-to-right answer counts were used to flag potential instances of cheating. The application of these cut points to the test under study yielded 64 flagged students while the same rules applied to the performance of the same students to a different test flagged 21 students. We will discuss how we arrived at setting up the cut scores using response time and present our recommendations for applying the same methodology to other similar computer-based tests.

Session 4C: Practitioner Session — Strategies for Planning and Audits

Building a Test Center Audit Program

Presenters: Patrick Watts, Heather Mullen

Regardless of whether your exam is paper-and-pencil or computer based, the administration can be one of the most vulnerable portions of the exam's lifecycle. The proper implementation of your company's policies and procedures is essential to maintaining your exam's integrity. As such, having a well-developed test center audit program is a crucial tool in any exam security practitioner's toolbox. This session will address how to build such a program from the ground up relying on the lessons learned by the CFA Institute in developing audit programs for both paper-and-pencil and CBT exams over the past four years. Topics to be discussed include best practices for audit programs, training of personnel, identification of test centers to audit, different types of audits, and what to do with the information after the audit is over.

"If You Fail to Plan, You Are Planning to Fail!" — Strategies and Techniques to Develop Your Test Security Plan

Presenters: Joe Brutsche, Aimée Hobby Rhodes, Michael Clifton

It is critical that testing programs put their resources where they will have the most impact. This presentation will introduce overarching strategies that should be considered when drafting your test security plan. With these principles in mind, the workshop will identify the key factors that go into developing a test security plan. Join experienced test security professionals and your colleagues for a hands-on opportunity to develop test security plan strategies, identify threats, and prioritize mitigation strategies.

Session 5A: Peer Review Session — Protection vs. Exposure of Items

Employing Statistics and Publication Countermeasures to Inoculate Exams Against Braindump Users

Author: Dennis Maynes

Large certification programs that administer tests globally on a continuous basis are susceptible to piracy and disclosure of their test items on braindump websites. Program requirements and resources often limit how often and how many items may be replaced. This research explores statistical methods and publication techniques, collectively referred to as "exam inoculation," for devaluing stolen content. In other words, the research in this paper tests the hypothesis that selective republication of test items guided by statistics will decrease the pass rates of braindump users and not impact the exam experience of honest test takers. Suggestions are provided for using statistics to determine how often and how many items should be republished. Detecting when the test should be republished is done using an "embedded verification test." (Maynes, D. & Burns, L., "To Catch a Cheat: Building Fraud Detection into Your Exams," presentation at the annual Association of Test Publishers (ATP) conference, Palm Springs, CA. 2012). Type I and Type II error rates of the embedded verification test will be studied by simulation. At the annual ATP conference in 2014, Maynes suggested using item clones for refreshing the exam items (van de Velde, B., Burns, L., & Maynes, D. "Exam Inoculation and Other Crazy Ideas to Stop Cheaters from Passing Exams," Scottsdale, AZ). This presentation will provide practical suggestions for developing clones in order to inoculate exams. Simulations will be used to evaluate the effectiveness of the proposed techniques when a known subset of the population has pre-knowledge of disclosed exam content. Finally, an item's increased service life through exam inoculation will be estimated which can be used to quantify the improvement to an organization's budget over current methods which require frequently republishing the entire exam.

Using Answer Key Imputation for Making Bayesian Inference about Cheating on Tests

Author: Marcus Scott

One method of cheating on tests is to steal the test questions in advance and determine an answer key. During the actual administration of the test, examinees who have had access to this stolen content respond to the items according to this predetermined answer key, which may differ from the real answer key. This paper presents a method for classifying the examinees into two groups: (1) those who most likely used the stolen content, and (2) those who did not. The method uses the EM algorithm (Dempster, Laird, and Rubin 1977) to iteratively determine to which group an examinee belongs. First, the answer key used by the examinees who are assumed to have accessed stolen content is imputed. If this imputed answer key differs from the true answer key, then each test is scored using both keys. Two-parameter item response functions (IRFs) are built using Measured Trait Item Response Theory. Each set of scores has its own set of IRFs. Probabilities of response patterns are calculated using these IRFs and are then used to make a Bayesian inference that an examinee belongs to a particular group. The updated groups are used to begin the next iteration. Data from a real-life security breach involving examinees who stole test questions beforehand and worked out a flawed answer key are used to demonstrate the method.

The Impact of Compromised Items in CAT on Score Estimates

Authors: Qing Yi, Meichu Fan

Computerized testing is gaining popularity in K–12 assessments. It provides several advantages over traditional paper-and-pencil testing. However, in many states there are limited numbers of computers or devices available in secure testing environments, which can extend the length of testing windows. This may increase the chance of items being compromised and can cause serious test security concerns. The goal of the study is to examine the impact of the locations of where the compromised items appear on score estimates in computerized adaptive testing (CAT). Five items are randomly selected as compromised items at four locations in the simulated CAT tests: Early (i.e., in the first 1/3 items), middle, late, and randomly across the test.

The item pool consists of 480 multiple-choice items. The three-parameter logistic item response theory model is assumed and the BILOG computer program is used to calibrate item parameters. Ten thousand theta values are generated. A fixed length CAT of 30 or 45 items is simulated. The maximum item information selection method with the Sympson-Hetter exposure control procedure incorporated is the CAT method in this study. Two different kinds of CAT simulations are conducted: in the absence and in the presence of the influence of compromised items. If a simulee encounters a compromised item, three different probabilities of getting this item correctly are simulated: 0.50, 0.75, and 1.00. Different percent of simulees are selected as those who have pre-knowledge of a test: 20, 40, and 50.

Several indices will be used to evaluate simulee ability estimation accuracy: (1) Pearson correlation of the maximum likelihood estimate of theta against true theta and (2) estimation errors: bias, root mean square error (RMSE), and absolute average difference (AAD) of the ability estimates.

Session 5B: Practitioner Session — Response Analysis

Autonomous Detection of Cheating in the Presence of Aberrant Responses in OMR Documents

Presenters: Raghuveer Kanneganti, Wim J. van der Linden

Optical Mark Recognition (OMR) is a well-known process of capturing human-marked data in documents. It is an extremely accurate and rapid form of data capture especially when each response can be entered as a single mark and for the same reason this process has been employed in various examinations/tests across the world. It is clearly understood that conducting these tests in an equitable manner is of the utmost importance. Sadly, in the past few years, there have been several cases in which teachers/administrators of elementary and high schools across the United States were identified for fraudulently correcting the answers written by their students in order to improve the success rate of their respective schools. In order to identify this format of cheating, a procedure was developed to autonomously determine if cheating has occurred by detecting the presence of aberrant responses in scanned OMR test books. The challenges introduced by the significant imbalance in the numbers of typical and aberrant bubbles were identified. The aberrant bubble detection problem was formulated as an outlier detection problem. A pool of features is initially selected by examining bubbles that are penciled by a large group of individuals and analyzing the differences between them. Several possible outlier detection methods were considered and a feature based procedure in conjunction with a one-class SVM classifier was developed. A multi-criteria rank-of-rank-sum technique was introduced to rank and select a subset of features from a pool of candidate features. Using the data set of 11 individuals, it was shown that a detection accuracy of more than 90% is possible. Experiments conducted on three real test books flagged for suspected cheating showed that the proposed strategy has the potential to be deployed in practice.

Is Response Time a Good Indicator of Aberrance

Presenters: Xiao Luo, Doyoung Kim

Computerized-adaptive testing (CAT) allows practitioners to collect new types of response information from the test administration, such as response time (RT). Many post hoc statistical analytical procedures have been proposed to exploit RT for the aberrance detection purpose, intuitively assuming that either a too long or too short RT implies suspicious behaviors. However, it is unclear whether RT is a reliable indicator of some testing behaviors or merely a random variable capturing haphazard information unrelated to any key components of test-taking behaviors. Some other philosophical questions would arise too. For example, if a test is by design a power test where test takers are given sufficient time to exhibit their abilities, would the analysis of RT be relevant and useful? If some test-takers are naturally fast in reading or thinking than some other test-takers who are more cautious in thinking, would the utilization of RT introduce biases or unfairness?

To better understand essential properties of RT, the proposed study is intended to perform various statistical analyses on empirical RT data collected from a large-scale CAT examination in order to investigate whether RT is: (i) an individual-dependent variable and/or an item-dependent variable, (ii) related to other contextual variables, such as time of testing, testing centers, etc., (iii) consistent through a test-taking session, (iv) and related to the construct being measured. Additionally, we will fit the RT data with promising latent class models, such as van der Linden's (2009) hierarchical response and RT model, so as to investigate whether RT can be effectively modeled. In the end, we will advise how to better evaluate the RT for aberrance detection—namely, practitioners should be looking at the total value of RT, the overall pattern of RT, or the interaction between individuals and items.

Session 5C: Practitioner Session — Test Security Investigation Techniques

Working a Case: Best Practices in Conducting Exam Integrity Investigations

Presenters: A. Benjamin Mannes, Rachel R. Watkins Schoenig, John Fremer, Marc Weinstein

When confronted with evidence of irregular activities like brain dumping, proxy testing, or other forms of high-stakes cheating, test sponsors must investigate to determine the scope of the activities, the extent of the potential harm to the test sponsor and to identify the participants in the scheme. Now more than ever, test sponsors must be extremely vigilant in enforcing their security policies and take decisive action against those who compromise the integrity of their exams. Test sponsors can only achieve these goals by conducting thorough and effective exam integrity investigations that produce admissible and compelling evidence for later use in legal proceedings against cheaters and perpetrators of fraud. When a test sponsor determines that an investigation is warranted, how should it be planned and carried out?

Strategies for preparing for and conducting an exam integrity investigation have evolved over a number of years with the experiences of a number of different testing programs. This session will cover these strategies through a discussion of case studies and best practices from four noted experts in exam integrity investigations. These experts will walk attendees through planning the investigation and conducting incident triage, as well as discussing investigative methods, what evidence to obtain and how the results of an investigation may be used by the test sponsor.

Session 6A: Peer Review Session — Copying

A Comparison of Similarity Indexes in Detecting Copying in Cases with Significant Observational Evidence on the Multistate Bar Examination

Authors: Mark A. Albanese, Cory Tracy

The purpose of this study is to investigate methods for improving the detection of copying on the Multistate Bar Exam (MBE) in cases where significant observational evidence exists. Currently, based upon the work of Hansen et al,(1987), four indexes are employed as a screening tool to indicate if sufficient evidence of cheating exists to warrant further investigation: the number of identical responses (IR), the longest consecutive sequence of items with identical responses (IR-LCS), the number of identical responses on items with different keys (IR-DK), and the number of identical incorrect responses on items with the same key (IIR-SK). To compute the probability of an observed index value, benchmark data sets composed of random pairs of examinees tested in different locations are formed and the prevalence of the observed value in the benchmark dataset determined. The K–index (Holland, 1996), and Angoff's B index, and H index (Angoff, 1974) have the potential to be useful alternatives or additions.

The ω index (Wollack, 1997), which controls for the total performance of the suspected copier and source is more complex to compute, but generally has been found to be the most powerful index of those available (Sotaridona & Meijer, 2002; Wollack, 2003) and is generally computed when the screening analysis indicates a follow-up is warranted. As such, ω is considered the standard of comparison for the various indexes evaluated in this study. In this study, results for the MBE administered from February 2010 through February 2014 are used to compare the various indexes singly, and in combination, for examinees suspected of copying reported by proctors. ω is used as the standard for determining which index(es) function the best.

Detecting and Reporting Test Irregularities in Online Testing

Authors: E. Matthew Schulz, Xiujuan Yuan

Pacific Metrics is currently working with a state department of education to design reports on testing irregularities in a high stakes, online testing program. The reports will be used to establish policies that minimize irregularities and provide information back to the state and local administrators about cases of potential cheating or other threats to test validity. Students, teachers, schools, and districts will be flagged for unusual levels of one or more of the following types of irregular activity: 1) session reopens (whenever a student has to reopen a test session), 2) item previewing (whenever a student who reopens a test session answers items that they viewed when they were in the session initially), and 3) answer changing (whenever a student changes their answer to a test question). Details concerning item response times, session duration, and elapsed time between initial entry and session reentry will be made available to assist the state and/or staff at the school or district level in following up on, and providing explanations for, flagged instances of irregular activity. In the presentation at the conference on detection of test fraud, we will provide examples of the reports we designed and provide details for instances of flagged irregular activity from the spring 2014 administration. We will also discuss lessons learned in working with state and local administrators and suggest directions for further work in detecting and reporting test irregularities.

Session 6B: Practitioner Session — Test Security Process & Framework

Measuring the Effectiveness of Your Test Security Program

Presenters: Rachel R. Watkins Schoenig, Ray Nicosia

An effective test security program will help a test sponsor ensure score validity and protect test content by deterring and detecting threats to the program and the test sponsor's reputation. This presentation will discuss measurement techniques and metrics to collect related to the test security program, including metrics for your hotline or tip line, use of photos for identification and incident rates for surrogate or proxy test taking and the use of wands for test centers and the incident rates for electronic device use.

Responding to Emerging Security Threats

Presenters: Aimée Hobby Rhodes, Camille Thompson, Joe Brutsche

As technology continues to evolve, testing programs need a holistic approach to protect the integrity of their program. The presenters will highlight emerging technologies (now and future possibilities), their associated threats/risks, and various mitigation strategies to address their imposed risk. This interactive presentation will provide participants with an opportunity to share experiences, learn what others have encountered, and how to plan for these new threats. Join experienced test security professionals to identify future security threats that could impact your program.

Session 6C: Practitioner Session — Selecting Data Forensics

Selecting the Right Data Forensics Analyses for Your Program: A Review of the Literature

Presenters: J. Carl Setzer, Kathleen A. Gialluca

“How can data forensics help my program, and which analyses are right for me?” The answers to those questions, of course, depend on what specific security concerns you have, and where the biggest threats to your program occur.

The most common security-related concerns of testing programs include:

(1) Have my items been exposed, and are they now behaving differently?

(2) Which test-takers have benefited from advance knowledge of item content, so that their test scores no longer reflect their real abilities?

(3) Have certain test-takers copied from each other?

The presenters have conducted a literature review of the statistical data forensics techniques that have been developed to address those particular concerns, and they will present a summary of that review. The presenters will describe the main analytical techniques that have been reported in the professional literature, and will do so in a manner that practitioners can readily understand. They will also describe the scenarios in which these techniques are most appropriate, and will articulate the assumptions underlying the appropriate use of these techniques. The presenters will specify what conclusions are justified from the results of the data analyses and, more importantly, what conclusions are not.

It is not the intent of the presenters to offer a technically detailed training session on how to perform a variety of statistical analyses. Rather, the overriding goal of the session is to help the participants articulate their primary security concerns, to acquaint them with the variety of statistical techniques available to them, and to help make them better-informed consumers of data forensics analyses.

Analysis of Online Answer Change Behavior with Survival Analysis Model to Detect Aberrant Behavior

Presenter: Mayuko Simon

This study explores the application of survival modeling to analyze the answer change behavior in the context of online tests. The data for this study come from a computer-based assessment where all students had the same set of items. The data record events such as wrong-to-right change (WR), log-in, log-out, and pause, along with the precise timestamp of the event. Presumably, students follow similar answer-change behavior. The aim of this study is to detect a situation in which WR with many items happens later in the test, especially after pause/interruption.

Survival models apply to event data, wrong-to-right changes (WR) in our study. Cox model, our choice of survival modeling technique, assumes a time-dependent baseline rate of WR. This baseline rate is the instantaneous probability that a WR takes place exactly at time T into the test. Survival models can intrinsically handle different lengths of test-taking time.

In this study, survival models will be used to predict the expected number of WR for each student during the course of the test. The difference between the expected and the observed number of WR, termed the martingale residual, represents the excess number of WR that cannot be explained by the length of the time the student spent taking the test nor by other predictors of interest. These predictors include the student's ability, the probability that the student would have answered this item correctly anyway given his ability, or that this WR take place after the first, second, or third pause/interruption.

We will pay attention to the coefficients of the model and the residuals. The coefficients signify the importance of predictors, which show you what; and excessively large (positive or negative) residuals are suggestive of aberrant behavior.

Session 7A: Practitioner Session — Test Security Investigation Techniques

So You Flagged a Cheater, Now What?

Presenters: Aimée Hobby Rhodes, Lorin Mueller, Kellie Early, A. Benjamin Mannes, Joy Matthews-Lopez

As psychometricians continue to develop and refine methods for flagging unusual candidate behavior, practitioners must determine whether and how to respond to these flags. This is not an easy task. In fact, asking non-statisticians to understand and appreciate the reasons for a flag can be a real challenge.

In this session, a panel of test security professionals will review the issues involved when presenting statistical analyses to various stakeholders, including senior management, external stakeholders, and candidates and their representatives.

The panel will begin by discussing senior management. Panel members will share their experience, both good and bad, explaining similarity analyses and statistical anomalies to senior leaders whose approval is required before action can be taken. The panel will review effective presentations of the data, and approaches to consider when addressing concerns of senior leaders.

Next, the panel will review ways to present this information to external stakeholders. Score recipients—whether a jurisdiction looking to award a license or school seeking to make an admissions decision—will have a number of questions about the review process and the integrity of the score. These questions can be difficult to answer, especially when a candidate has not granted the test sponsor permission to discuss the situation, but the score recipient needs to understand whether the score is valid and transparency about review process is essential.

Finally, the panel will discuss how best to explain the data to the flagged candidate and how that explanation might change when representatives of the candidate get involved. For high-stakes tests, invalidating a score can have life-changing consequences for the candidate. The wrong presentation or explanation can incite the candidate or his representative and may lead to legal action. The panel will discuss how to present the evidence to the candidate's attorney or in response to legal action.

Session 7B: Practitioner Session

Things that Go Bump in the Night: What Should We Be Worried About?

Presenter: David Foster

In the world of test security we usually react, or more accurately, overreact. When we hear of a breach or even a possible breach we jump immediately into the fray, adrenalin coursing through our veins and worry on our face. We involve others, create turmoil, interrupt normal activities, make quick decisions, and spare no expense to right the ship. Then we deal with the damage for months, perhaps years. This all-too-typical scenario, of course, is not the ideal way to manage security for a testing program.

It is common also to decide on the security we need—proctoring, legal agreements, background checks, web monitoring, data analysis, tip lines, and many, many more—before we even know we need it. We make these decisions based on tradition, advice of others, knee-jerk response to a breach, what others are doing, and many, many more. Most of the time these security methods are not needed and are not appropriate.

Security should start at the beginning. It should flow from the actual demons that can hurt us. This session will present and discuss the major categories of security threats to a testing program, provide examples of each, present how to evaluate and put each into perspective with an analysis of risk, and how to set up an effective defense against the more dangerous.

A Framework for Policies and Practices to Improve Test Security Programs: PDIR

Presenter: Steve Ferrara

The goal in maintaining test security is to ensure that the test data we generate for evaluating student achievement, teacher evaluation, and school performance is trustworthy and that our investments in secure test material are protected. Generally speaking, (a) most testing programs appear to be doing marginally well in protecting test security—shocking stories about lax protection of secure materials, loosely managed test administrations, and appalling cheating incidents represent the behavior of fewer than 1% of all educators, according to one estimate; (b) research on statistical methods to detect cheating is producing excellent enhancements—though those methods currently detect only a narrow range of cheating behaviors; and (c) our policies and practices in investigating potential security violations and resolving those cases are quite weak. In this presentation I will offer a framework for building and enhancing test security policies and practices that more rigorously address prevention, detection, investigation, and resolution, or PDIR.

The presentation will provide logical argument and some evidence to support the assertions in (a)-(c) above; compare and contrast the knowledge and skill requirements of professional investigators (e.g., in law enforcement and journalism) and educators to support a proposal to develop a cadre of investigators that specialize in educational testing; and argue that entities who, by their role in educational testing, should take lead responsibility for establishing and enforcing test security policies and practices. Those entities include state legislatures, boards of education, and departments of education and the US Department of Education; and supporting entities such as the Association of Test Publishers, Council of Chief State School Officers, National Association of Test Directors, National Council on Measurement in Education, and NRC's Board on Testing and Assessment.

Demo/Poster Session

Detecting Answer Copying in an Assessment When Multiple Forms Are Administered

Authors: Chi-Yu Huang, NooRee Huh, Hongling Wang, Qing Xie

Many indices for detecting answer copying have been developed and applied in situations where a single form was administered to all examinees. However, to prevent examinees copying answers from examinees sitting next to them, many testing programs administer multiple forms in one administration. This paper will explore the possibility of applying some well-established answer-copying detection indices in situations where multiple forms are administered simultaneously.

Three indices, w index (Wollack, 1997) and the Pair 1 and Pair 2 indexes (Hanson, Harris, & Brennan, 1987), will be used with a necessary modification in this study.

In this study, the definition of w index will remain the same regardless the suspected copier and source take the same or different forms. One statistic in Pair 1 will be extended to two indices: the total number for which the suspected pair and source picked the same incorrect response on the items with the same keys (JI1I2_SK) and on the items with different keys (JI1I2_DK). The definition of another statistics in Pair 1, which is the number of the longest sequence of consecutive identical responses (STRINGL) will remain the same. One statistic in Pair 2 will be extended to another two indices: the total number for which the pair picked the identical response on the items with the same keys (TJOINT_SK) and on the items with different keys (TJOINT_DK).

This study will use simulation data to examine the detection power and errors for these indices. Different factors will be investigated: form similarity, number of different forms used, percentage of items that the suspected copier copied, copying patterns, and the ability difference between the suspected copier and source.

The results from this study will provide helpful guidance for detecting response similarity, not only between examinees who take different paper-and-pencil forms in a test center, but may extend to those who take computer-adaptive tests, where each examinee could receive different items in a test.

Discrete Option Multiple Choice: Preventing Cheating and Test Theft

Author: David Foster

It's not often in the testing industry that a small change in design can have large security benefits. Converting traditional multiple-choice tests to the Discrete Option Multiple Choice, or DOMC, provides an opportunity to make such a change. The DOMC format presents the options of a typical multiple-choice question in a new way, one at a time, randomly. Test takers respond by clicking on YES or NO buttons depending on whether they believe the option is the correct one or not. Options stop being presented when the question is answered correctly or incorrectly, resulting in fewer options being exposed unnecessarily.

This session will demonstrate DOMC, illustrate how it solves major security problems, summarize the research on it to date, and answer your questions.

Simple Exploration of Answer-Changing Behavior Relative to Item Response Time and Difficulty — TAD Categorization

Authors: Djibril Liassou, Vincent Primoli

This paper uses data from multiple state testing programs and years to provide empirical results and comparisons related to the type and frequency of answer-changing (AC) behavior by item response time and item difficulty in the computer-based testing domain. So, simple descriptive statistics and correlations will be provided for the three variables of interest.

Also, TAD (item Time, Answer-changing, item Difficulty) categorizations will be introduced. Using confidence intervals, the item data will be split into three categorizations (low, medium, high) per variable at both the state and school level to see what trends emerge. For example, an item with a TAD flag of "LHL" means on average there is a low (L) time spent on the item, a high (H) answer-change rate, and the item has a low (L) difficulty. These categorizations could potentially help assess test score validity at the content, state, district, and school levels.

It is hoped that this paper furthers the field's understanding of AC behavior as an assessment construct and the factors associated with it. In addition to providing base-rate information regarding the prevalence of AC and its covariates, it is possible such information and analysis might help improve related data forensics procedures.

CESP and Me - Why Should I Get Certified as an Exam Security Professional?

Author: Jamie Mulkey

Test security is now a true profession. Individuals need to be recognized and receive a credential for their test security expertise. The Certified Exam Security Professional (CESP) Certification program will do just that.

Come learn about the new CESP program and how to prepare for the CESP-Generalist exam. This poster session will discuss:

- The CESP program structure
- How the CESP-Generalist fits into the overall CESP credential
- The competencies required for the CESP-Generalist
- The use of DOMC (Discrete Option Multiple Choice) items in the exam
- How to navigate the training and reference materials to best prepare for the exam

This poster session is a must for those seeking the CESP credential.

A Practitioner's Approach to Improving Test Security

Authors: Jason Taylor, Karoline Jarr, Claudia Guerere, Kristin Donlon

The poster will outline the steps Project Lead The Way (PLTW) has undertaken to create a more secure testing program. PLTW is a non-profit organization that provides STEM curriculum, assessment, and professional development to more than 5,000 schools in the United States. As part of PLTW's balanced assessment program, End of Course (EoC) assessments are administered to high school students in 11 engineering, biomedical, and computer science courses.

Over the last four years, PLTW has made a concerted effort to improve the security of the EoC assessments, utilizing various techniques to identify and reduce security threats that impact score validity. This poster presents a practitioner's perspective and lessons learned as the testing program changed its policies and practices to improve test security as recommended in current literature.

A comprehensive approach to test security that focuses on prevention and deterrence is being implemented by the PLTW Assessment team. Improvements to item-writing, test development, test administration practices, policy, operations, and communication plans will be described. Methods include the transition from a paper-and-pencil to online test administration model, the development of a paperless item writing and review process, option randomization, computer-timed sessions, item expiration dates, and inclusion of "Forward-Only" test sections.

The PLTW test security approach also focuses on communication and policy development. This includes a series of checks and balances in the student rostering process, a standardized script for all test proctors to use, as well as a testing guidelines policy that all teachers must electronically agree to prior to test assignment.

Test Security: Not Just a "Bolt-On"

Author: Jennifer Geraets

Sometimes test security is thought of as something that is only included at the end of the testing process. The testing program is designed, distributed, delivered, and then "test security" happens when the test data are analyzed to determine if there is any question regarding the validity of test results. This presentation will explain why test security needs to be part of a testing program's lifecycle, and how ACT is using the program analysis process to ensure that test security is built into each program from the beginning, and is not an afterthought that is "bolted-on" at the end.

Graphical Representations of Test Fraud

Authors: Jennifer Lawlor, Peter Pashley

The simplest approach can sometimes yield the greatest results. The use of graphical displays of data can be a useful tool for identifying statistically significant anomalies that may indicate cheating, test collusion, or item pre-knowledge. More generally, graphical displays of data can supplement and inform all types of statistical analyses. For example, the influence of outliers or nonlinearities on correlation coefficients can be easily exposed by way of simple scatter plots. Similarly, graphics can help detect and verify cases of test fraud. They can also provide visual evidence of test fraud that can be very convincing to nontechnical audiences. This presentation will discuss some new and innovative test-fraud graphics and the usefulness of these graphics in test-fraud detection. Data from a high-stakes, large-scale assessment will be used. Specifically, sections that are susceptible to copying will be compared graphically to those that are not. The methods used to create and analyze the resulting graphs will be presented in detail. This presentation should suggest the potential for graphical representations in a variety of other assessment settings as well.

Testing the Efficacy of CleanSlate Cheating Prevention Paper

Authors: Max Brickman, Hugh Watson, Haley Gedek

Cheating is a major problem in the academic setting, and each year more and more students admit to looking off of other students' tests. As the prevalence of this issue continues to increase, the accuracy of our testing results is decreasing, creating the need for a more cheating-resistant testing process.

The current study aims to compare the efficacy of CleanSlate testing sheets to traditional multiple-choice sheets in terms of the capability to cheat. CleanSlate is a new testing form designed to eliminate cheating by disabling the ability for students to see their classmates' answers. When viewed at an angle, CleanSlate exams appear very darkly tinted, rendering the testing sheet indecipherable. However, to the student taking the test, it is translucent, allowing her or him to take the test as normal without the fear of wandering eyes. Preliminary data from focus groups and smaller research trials have led us to believe that CleanSlate testing sheets will greatly reduce students' ability to cheat on multiple-choice exams.

To test this hypothesis, participants will be assigned to one of two groups; one given a standard multiple-choice test and the other given CleanSlate testing sheets. Participants will then be asked to take a mock test and cheat off of nearby peers. We believe those given CleanSlate exams will perform significantly worse that those given traditional multiple choice, thus supporting the belief that CleanSlate testing sheets can help eliminate cheating in academic settings. Additional data regarding ease of use and time taken to complete exams will be collected, and we will discuss the possibility of implementing similar products into the classroom as a means of securing the integrity of our testing methods and the accuracy student data.

Legal Defense of Score Cancellation Decisions

Author: Michael Clifton

The poster will highlight issues that are litigated/arbitrated in situations involving the cancellation of test scores, such as: the importance of the examinee agreement and its establishment of the standard required for score cancellation; case law and statutes applicable to score cancellations; the importance of sound statistical evidence in the absence of eyewitness testimony; due process; contracting with minors; compelling arbitration; and appropriate messaging.

The poster will invite conversation related to the aftermath of a score cancellation decision, and how such aftermath informs decisions made during the investigation.

Sticking to the (Investigation) Plan: Effectively Documenting and Monitoring your Test Security Investigations

Author: Mikel Trevitt

“Everyone has a plan ‘til they get punched in the face” - (Mike Tyson)

While Mr. Tyson’s not-so-subtle quote may seem harsh, it does speak to a common challenge facing investigators and managers when planning, monitoring, and keeping multiple investigations moving forward. That is, how do you keep investigations on task, organized and driven toward conclusion once you’ve launched them and the “newness” wears off? Even the best intended plans can get derailed or “lost in the shuffle” of other daily work, priority shifts, staffing changes, new discovery, or other unforeseen roadblocks once the actual investigation process is put in motion. Having a standard, effective Investigation plan designed around specific goals, clear objectives, checkpoints, evidence collection, and action items is imperative to the life of the investigation. This presentation will provide an opportunity to discuss best practices for creating your investigation plan and the subsequent documenting and managing of the investigative process within your plan document. Taking time to carefully document, and stick to your plan allows better case organization, visibility, and clarity for anyone involved with the investigation process and creates a single “source of truth” repository for the evidence, findings, theories, and outcomes associated with the investigation.

Who’s Cheating Who — A Look at Modern Cheating and Stealing Tactics

Authors: Christy Frederes, Tara Miller

This display is for attendees to understand the current trends in test misconduct and how to place best practices within the process to diminish misconduct before, during, and after test administration.

Demonstration of current technology used to ”cheat” and “steal” content, display of gadgetry, video of remote proctoring stunts by cheaters and stealers. Handout of proctoring best practices and response program outline.

During Display will discuss and share information on the following topics:

Test Security Program- our processes and tools in place
Our Mission: The prevention, detection, correction and investigation control processes to ensure the validity of results for all exams and assessments administered.

- Test Security — Combating “cheaters” and “stealers”
- Current Trends for Test Misconduct
- Fighting Technology with Technology
- Proctoring Best Practices

Practical Ways to Enhance Test Security: Messaging to Candidates and Stakeholders

Authors: Chuck Friedman, Rory McCorkle

Test security guidelines and practices cannot be viewed in isolation, and any group faced with security issues must look at different ways to deter test cheating threats and test theft threats. Test sponsors must be sensitive to security risks before, during, and after the testing process to best protect their intellectual property.

This poster session will draw upon the publication of the Association of Test Publishers Security Committee and updated information to present practical ways to message program test security to multiple audiences.

Messaging to candidates, test users, stakeholders, educators, and the public cannot be minimized as a way to deter cheating and test theft. In the social media age and plethora of breaches in the public media, it is incumbent upon test sponsors to be proactive in publicizing the importance of maintaining test security, protecting intellectual property, and warning candidates repeatedly of examples of inappropriate behavior and its potential consequences.

The poster show will show participants examples of legal agreements, tips to be included in candidate handbooks, the use of video in the registration process, information to be presented on the day of testing, on the score report, and in other public venues.

The underlying theme is test sponsors must be vigilant in spreading the work to candidates and stakeholders. While a candidate may sign a non-disclosure agreement, it does not mean that he/she has really read the full statement nor understands the implications of inappropriate behavior.

The Development of an R Package dataForensic for Conducting of Test Security Analysis

Authors: Jiyoon Park, Yu Zhang

Statistical analysis of test results is the most widely used approach employed by test sponsors to capture the signs of security breaches and to evaluate the validity of test scores. The proposed R package (dataForensic) is a statistical tool that provides systematic and comprehensive analyses in test security. The dataForensic package provides applications of a variety of existing methods for analyzing test results and detecting cheating behaviors. The package can be used for: (1) screening spuriously low scores, (2) identifying potentially compromised items and identifying examinees who had an advantage in testing from pre-knowledge of those items, (3) identifying aberrant responses through use of several person-fit indices, and (4) identifying pairs or clusters of examinees who present similar response patterns through use of several IRT and non-IRT similarity indices.

The person-fit indices provided in dataForensic have been shown to be effective and reliable based on the simulation studies in the literature. Both parametric (e.g., l0, lz, CUSUM, and ECI) and nonparametric person-fit indices (e.g., rpbis, C, U, NCI, and HT) are included.

The dataForensic package also includes similarity indices that have shown high performance in detecting spuriously similar responses in the literature, including IRT-based indices (e.g., w [omega], GBT index) and non-IRT-based indices (e.g., K and K variants).

Disrupted Opportunity Analysis (DOA): A System for Detecting Unusual Similarity between a Suspected Copier and a Source

Authors: Mark A. Albanese, Cory Tracy

Detecting cheating on an examination is a complex task that depends upon a large number of factors. There are also a large number of different indexes that have been developed to identify unusual similarity between a copier and the source. One of the most difficult situations in which to detect cheating is if the suspected copier and the suspected source are taking the same form of the test and the source is a high-functioning examinee. The approach I propose to quantifying the probability of copying can be used for this situation as well as any in which an examinee is copying from some other source. The main requirement is that at some point in the examination process, the suspected copier must be separated from his/her source. It would also be helpful, but not necessary that a record of the suspected copier's answers created at that point. Several examples with actual data will be provided showing how the DOA approach would work and then some suggestions will be provided for how it could be implemented in many testing situations with reasonable and inexpensive methods. The main challenges are to create conditions in which examinees can be moved without unintended consequences and ruling out competing explanations such as speededness, or emotional upset for any change in performance. Broad implementation of the DOA method would require test organizations to take more proactive steps in addressing suspected copying. In earlier work, Albanese and Wollack argue that cheating should be considered a threat to the response validity of scores. Assuming the challenges noted for moving examinees can be met, the DOA method would be one-step testing organizations could take to be more proactive in addressing cheating at a systems level.

Closing Keynote Address

Screen is NOT Paper. Story is NOT History. Visualization, Data, Analysis, and Other Hurdles

Presenter: Nahum Gershon

Technology has enabled us to do a plethora of things we could not do before. We can represent data and information, for example, in ways unimagined before, not only in words or simple images, but also in automatically generated complex visual representations. This has enabled developers and other technically proficient people to generate by themselves visual and other representations of their data in a DIY fashion. However, making these representations effective frequently requires visual, design, representational, and analytical literacies. This poses a great challenge to many technology (and other) users. After all, the pencil inventor was not necessarily the best artist. This talk will focus on this challenge and chart some potential ways to overcome it.

Thank you for attending the 2014 Conference on Test Security

KU THE CENTER FOR
EDUCATIONAL TESTING
& EVALUATION
The University of Kansas

See you in 2015 at the University of Kansas!