

Critical Policies and Procedures for Test and Item Security

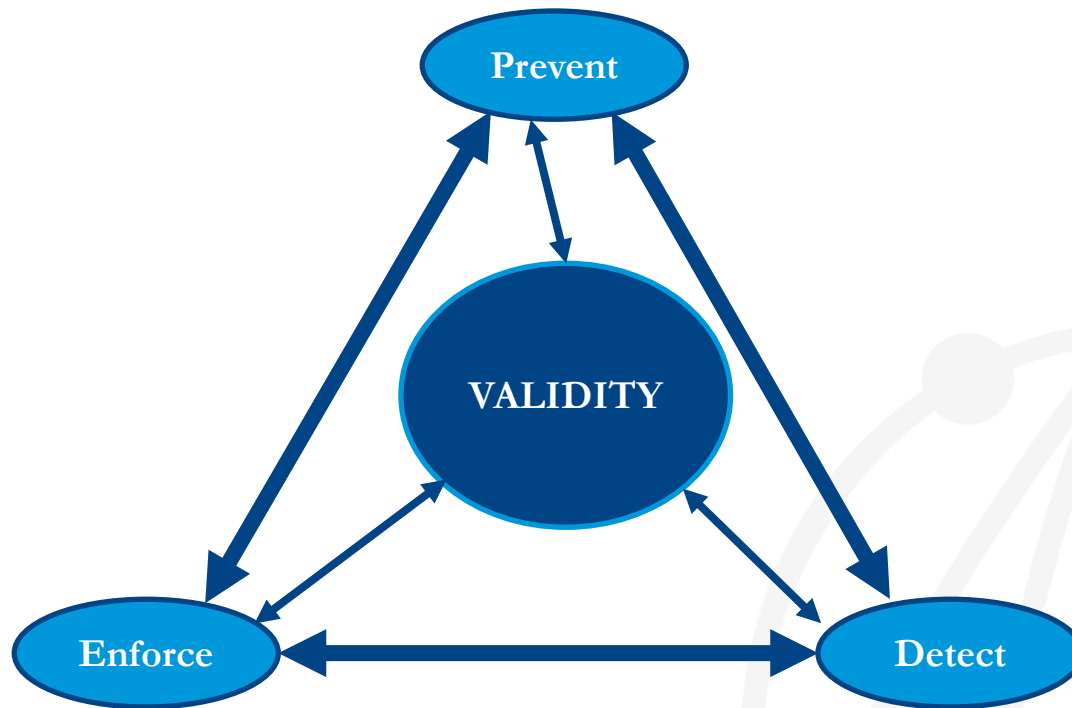
Andrew Wiley, Ph.D.



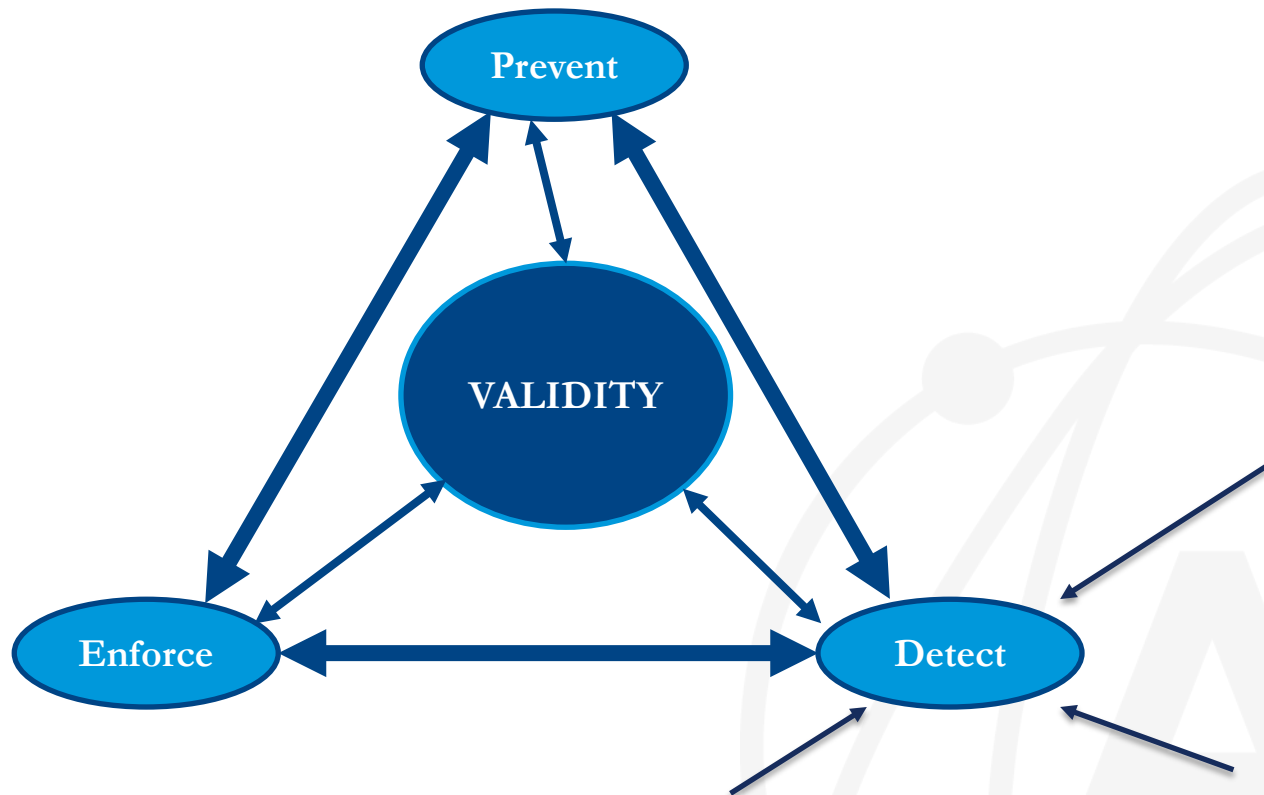
High stakes = High motivation for cheating



Security Validity Triangle



Security Validity Triangle



Just a little bit about Alpine

- » Clients ranging from education sector to professional and IT credentialing
- » Assessment are delivered through all means, ranging from paper and pencil to CBT (CAT, linear, etc.)
- » Education (as a general rule) has been the last to adopt CBT delivery models
 - Challenges related to large-scale administration nationally or all students within a given state
 - Schools struggle with the requirements when it comes to infrastructure (i.e. bandwidth, physical resources, staffing)

Let's talk a little bit about prevention

- » Identification of primary and secondary security threats
- » Item banking policies and procedures
- » Candidate screening protocols
- » Test administration policies and practices
- » Training of proctors and administrators
- » Test delivery protocols
- » Audits of test administration sites
- » Test scoring protocols

A few words about DDoS attacks

Florida Not the First to Suffer Testing Cyber Attack | Sunshine State News | Florida Political News - Google Chrome

www.sunshinestatenews.com/story/florida-not-first-suffer-testing-cyber-attack

SUNSHINE STATE NEWS

[f](#) [t](#) [v](#) [r](#) About | Advertise | Contact | Sign in or Create Account

Search for something.. [DESKTOP VERSION](#)

[POLITICS](#) NANCY SMITH COLUMNS BLOG BUSINESS STATE CARTOONS NEWS ALERT SIGN-UP

Politics

[Tweet](#) 2 [ShareThis](#) 14 [Share](#) 0 [Like](#) 3

Florida Not the First to Suffer Testing Cyber Attack

By ALLISON NIELSEN

March 11, 2015 - 7:00pm

Florida is the latest victim of a cyber attack, leaving many students unable to log in and complete the writing portion of the Florida Standards Assessment, but the Sunshine State hasnt been the only place where technological issues have caused problems for standardized testing.

Just a year ago, Kansas was in the same position as Florida is this year: It was the first year the Sunflower State had administered the [Center for Educational Testing and Evaluation](#) exams under a new set of education standards, and the roll-out

A few words about DDoS attacks

- » Distributed Denial of Service attacks
- » Has caused significant disruptions in numerous statewide testing programs (Kansas, Florida, etc.)
- » Appears to be an even more significant risk in high visibility programs such as statewide assessments
- » A firewall may not provide sufficient protection against sophisticated DDoS attacks
- » The risk of a DDoS attack, particularly at the beginning of testing windows, needs to be built into the administration protocol so that threats can be quickly identified

A brief review of investigative activities

- » Essential that data collection and policies are in place that will support any anticipated investigative needs. Critical facts about the test administration (time, location, setting information)
- » Whistleblower procedures and protections
- » Establish standards that will trigger an investigation
- » Roles for individuals and protocols for data sharing must be clearly defined
- » Importance of transparency
- » Establish and identify sanctions for cheating

Detection – Content sharing

- » Can create systems for web crawling for test content
- » Google Alerts for certain materials
- » Time consuming and expensive
- » Can work with websites to have content removed
- » Also consider utilizing third party affiliates (payment vendors)

Detection – Statistical Analyses

» Advantages

- Can swiftly and efficiently analyze the behavior of all available test takers
- Can investigate for systematic patterns at both the individual and within test administration sites
- Can be built into scoring procedures before scores are released

» Disadvantages

- In education, it is difficult for statistical methods to be the sole determinant of cheating behavior
- Can yield false positives on some occasions, multiple pieces of evidence can be valuable

Detection – Statistical Analyses

» Identification of most significant threats

- Specific item content can get exposed over the test administration window
- Test content can get exposed over the test administration window
- Test candidates can attempt to copy answers from other candidates during the test administration
- Test candidates can obtain pre-knowledge of the exam content before taking the exam

Item & Form-Level Analyses


- » Evaluate statistical data regarding form and item level performance during operational administrations and across time
 - **Use:** Track exam volumes and pass rates over time
 - **Performance:** Ensure forms and items are functioning as intended in operational environment
 - **Exposure:** Track both item and form-level exposure to address security concerns

Is test content exposed?

- » Over time, items and test content can become exposed and impact behavior of test takers
- » At the item level, item statistics can be evaluated over time to determine if the item statistics have started to change over time
- » This impact could be observed by shifting performance across the test form; increased overall test scores, raising pass rates for candidates, etc.

Item-Level Statistics:

Item Difficulty



Count	Item ID	Rasch Item Difficulty Measure	P-value	Item-Score Correlation	Item Reliability	Number of Responses
1	198432	0.13	0.738	0.556	0.245	56
2	198433	-0.52	0.817	0.461	0.178	56
3	198434	0.70	0.664	0.725	0.343	56
4	198436	0.70	0.664	0.560	0.265	56
5	198441	-0.60	0.826	0.495	0.188	56
6	198442	0.23	0.725	0.590	0.264	56
7	198446	-0.23	0.783	0.595	0.245	56
8	198447	-1.90	0.933	0.326	0.081	56
9	198449	0.32	0.713	0.561	0.254	56
10	198450	0.09	0.743	0.636	0.278	56
11	198451	0.05	0.748	0.572	0.249	56
12	198452	-1.53	0.910	0.264	0.076	56
13	198453	-0.45	0.810	0.484	0.190	56
14	198455	0.58	0.680	0.648	0.302	56
15	198456	-0.29	0.790	0.462	0.188	56
16	198458	0.26	0.722	0.387	0.174	56
17	198459	-2.81	0.970	0.268	0.046	56
18	198460	-0.38	0.801	0.571	0.228	56
19	198463	-0.50	0.815	0.491	0.191	56
20	198464	1.47	0.562	0.720	0.357	56

Parameters | **Item Selection** | Option Analysis | Form Level A

P-Value

- Item difficulty for dichotomous items (0,1) in CTT
- Proportion of candidates who answered the item correctly
- Ranges from 0 to 1, or 0% to 100%
- High values indicate easier items; low values indicate hard items
- Lower values indicate easier items; higher values indicate more difficult items

Average Item Score

- Item difficulty for polytomous items (0 through maximum points value) in CTT
- Average number of score points earned by candidates
- Ranges from 0 to maximum number of points
- Interpret on the scale of the maximum number of points

Item-Level Statistics: Correlation



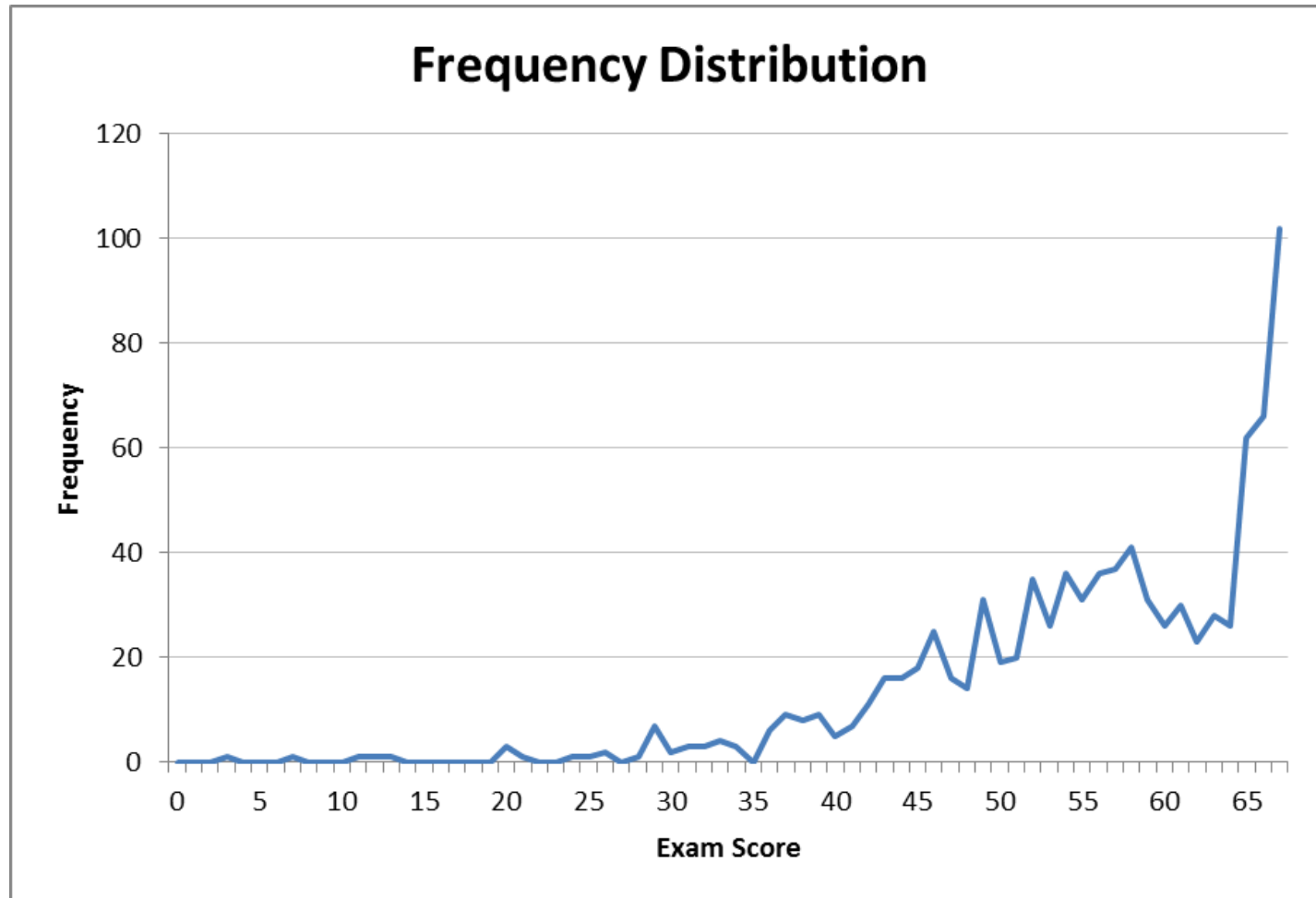
Item Selection Spreadsheet

Count	Item ID	Rasch Item Measure	P-value	Item-Score Correlation	Item Reliability	Num Res
1	198432	0.13	0.738	0.556	0.245	
2	198433	-0.52	0.817	0.461	0.178	
3	198434	0.70	0.664	0.725	0.343	
4	198436	0.70	0.664	0.560	0.265	
5	198441	-0.60	0.826	0.495	0.188	
6	198442	0.23	0.725	0.590	0.264	
7	198446	-0.23	0.783	0.595	0.245	
8	198447	-1.90	0.933	0.326	0.081	
9	198449	0.32	0.713	0.561	0.254	
10	198450	0.09	0.743	0.636	0.278	
11	198451	0.05	0.748	0.572	0.249	
12	198452	-1.53	0.910	0.264	0.076	
13	198453	-0.45	0.810	0.484	0.190	
14	198455	0.58	0.680	0.648	0.302	
15	198456	-0.29	0.790	0.462	0.188	
16	198458	0.26	0.722	0.387	0.174	
17	198459	-2.81	0.970	0.268	0.046	
18	198460	-0.38	0.801	0.571	0.228	568
19	198463	-0.50	0.815	0.491	0.191	568
20	198464	1.47	0.562	0.720	0.357	568

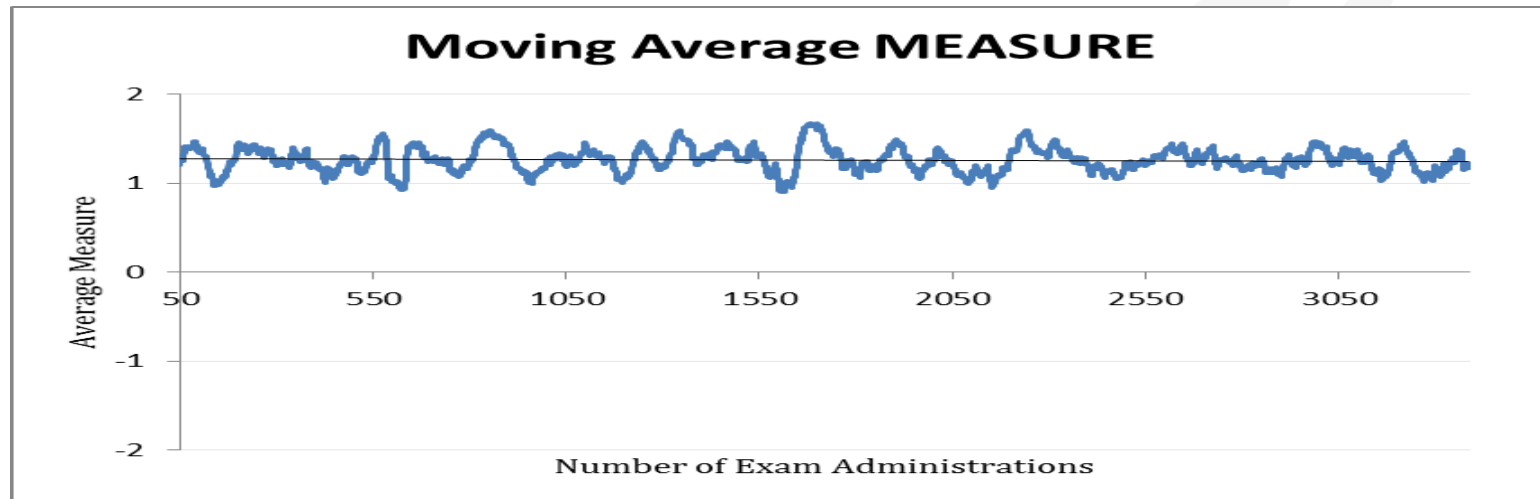
Item Score Correlation

- Point Biserial Correlation for dichotomous items
- How well an item differentiates between high and low ability candidates
- Estimated by performance on the exam---typically relationship between performance on the item and total score (although other values can be used)
- Range from -1 to 1
- Strong + correlations = item discriminates well between candidates; high ability candidates answer item correctly/low ability answer incorrectly
- Low + or - correlations = item does not discriminate between candidates; high ability candidates answer item incorrectly or low ability answer correctly

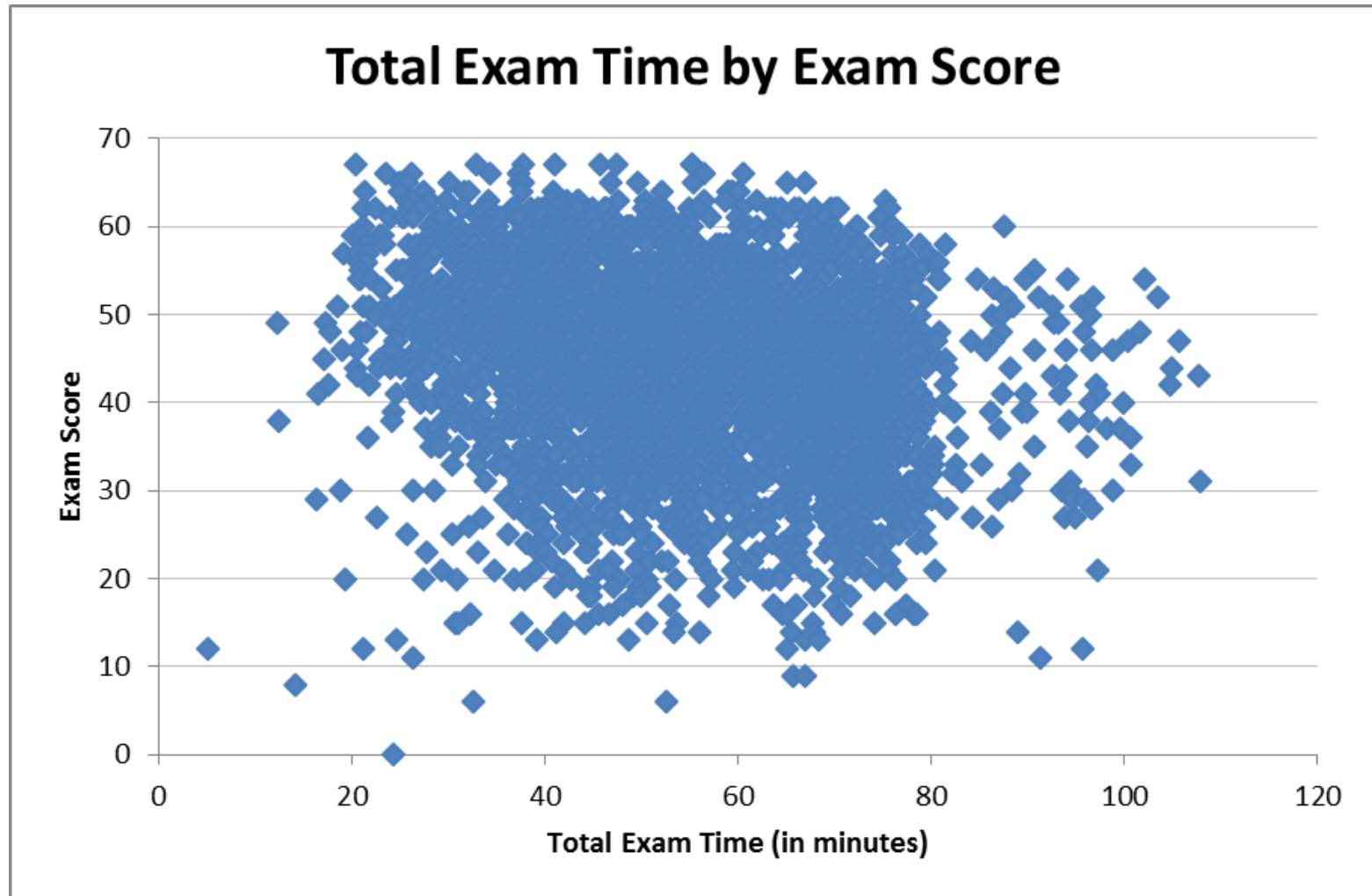
Is test content exposed? Test Scores



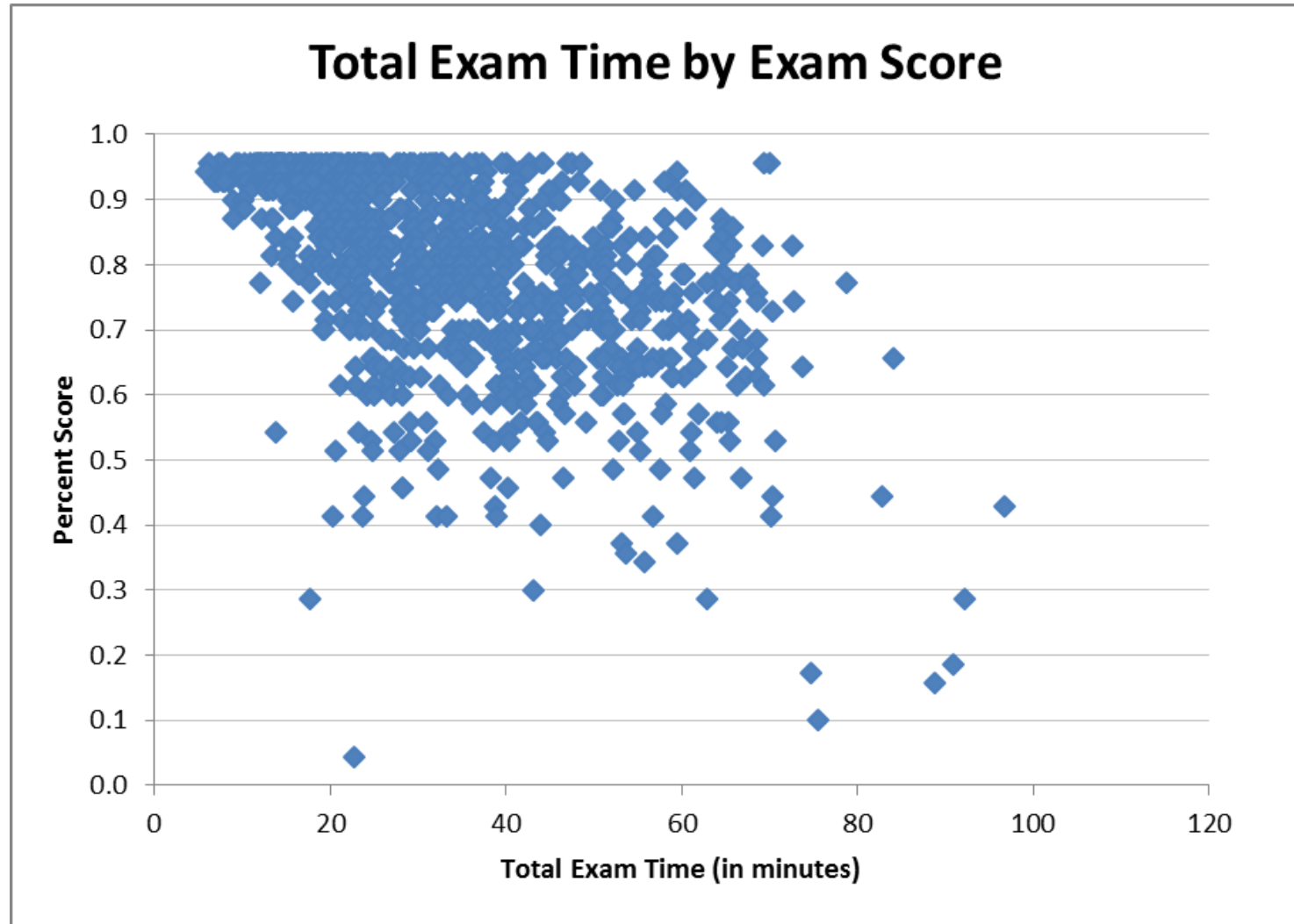
Is test content exposed? Test Scores



Is test content exposed? Test score by time



Is test content exposed? Test score by time



Indications that candidates have cheated?

- » Candidates can cheat during the course of taking the exam by trying to copy the answers of other test takers
- » Candidates can also cheat by obtaining access to the test content prior to sitting for the test administration

Candidates copying answers

- » Evaluate test taker response patterns across candidates in the same test center
- » Necessary to have data and information on the time and location of test administration procedures
- » Critical to evaluate similar of response patterns for correct and incorrect items
- » CAUTION: Consistent incorrect response patterns could reflect consistency of curriculum or teaching behavior if students or training has occurred together

Candidate pre-knowledge

- » If candidates have pre-knowledge of test content, they can receive an unfair advantage and perform significantly better than would be expected based upon their “true” ability
- » Can happen with large scale knowledge of test content being made available
- » Can also happen in a more “local” fashion with specific cheating behavior or information about specific portions of the test

Candidate pre-knowledge

O'Leary and Smith (in press)

» Differential Person Functioning (DPF)

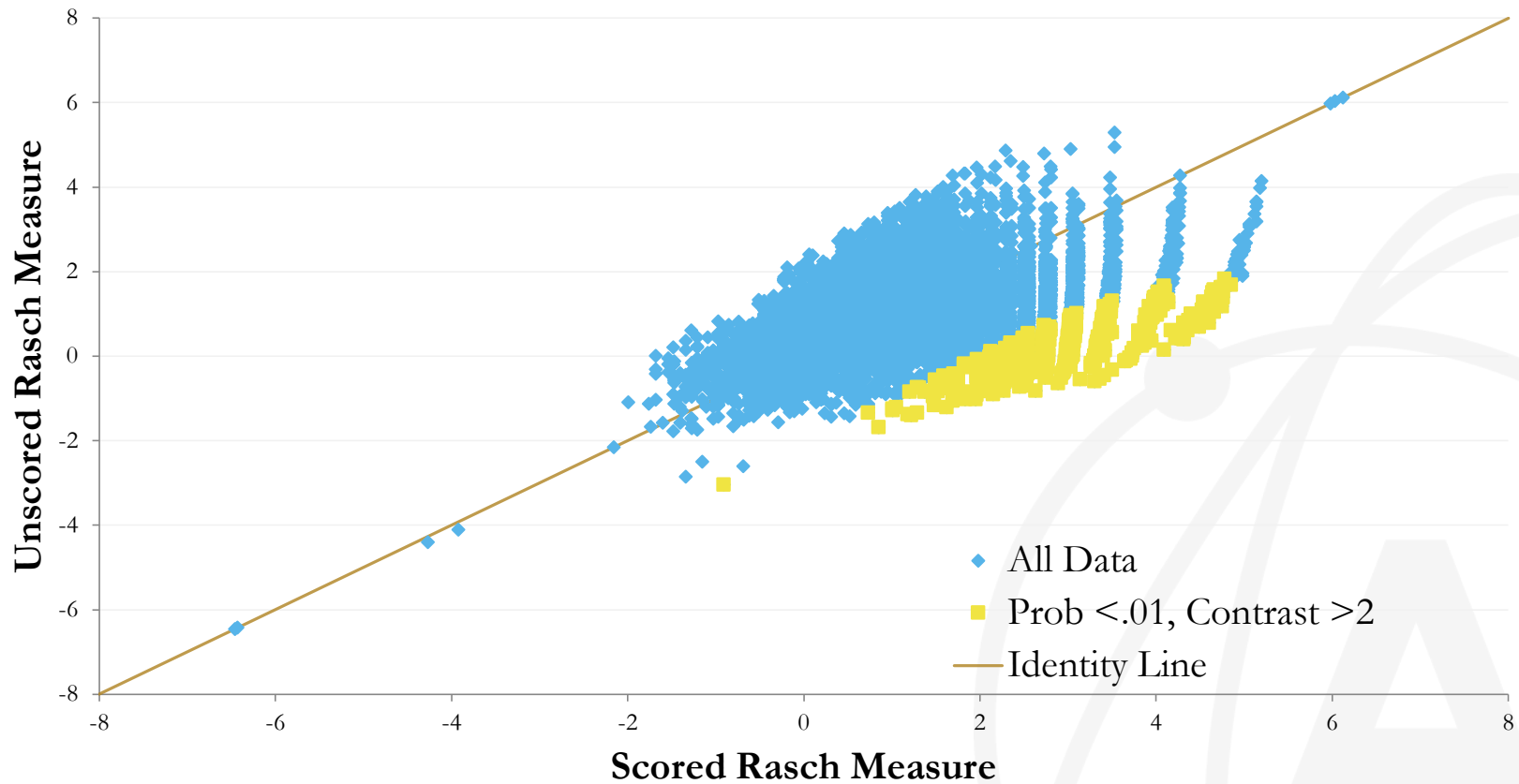
- Compare candidate performance on items that have already been used (exposed) to a new set of items (unexposed), while holding item and candidate characteristics constant
- Candidates who perform unusually well on exposed items, as compared to unexposed items are identified as possibly having pre-knowledge

» Differential Item Functioning (DIF)

- Compare candidate performance on items between flagged candidates and un-flagged candidates

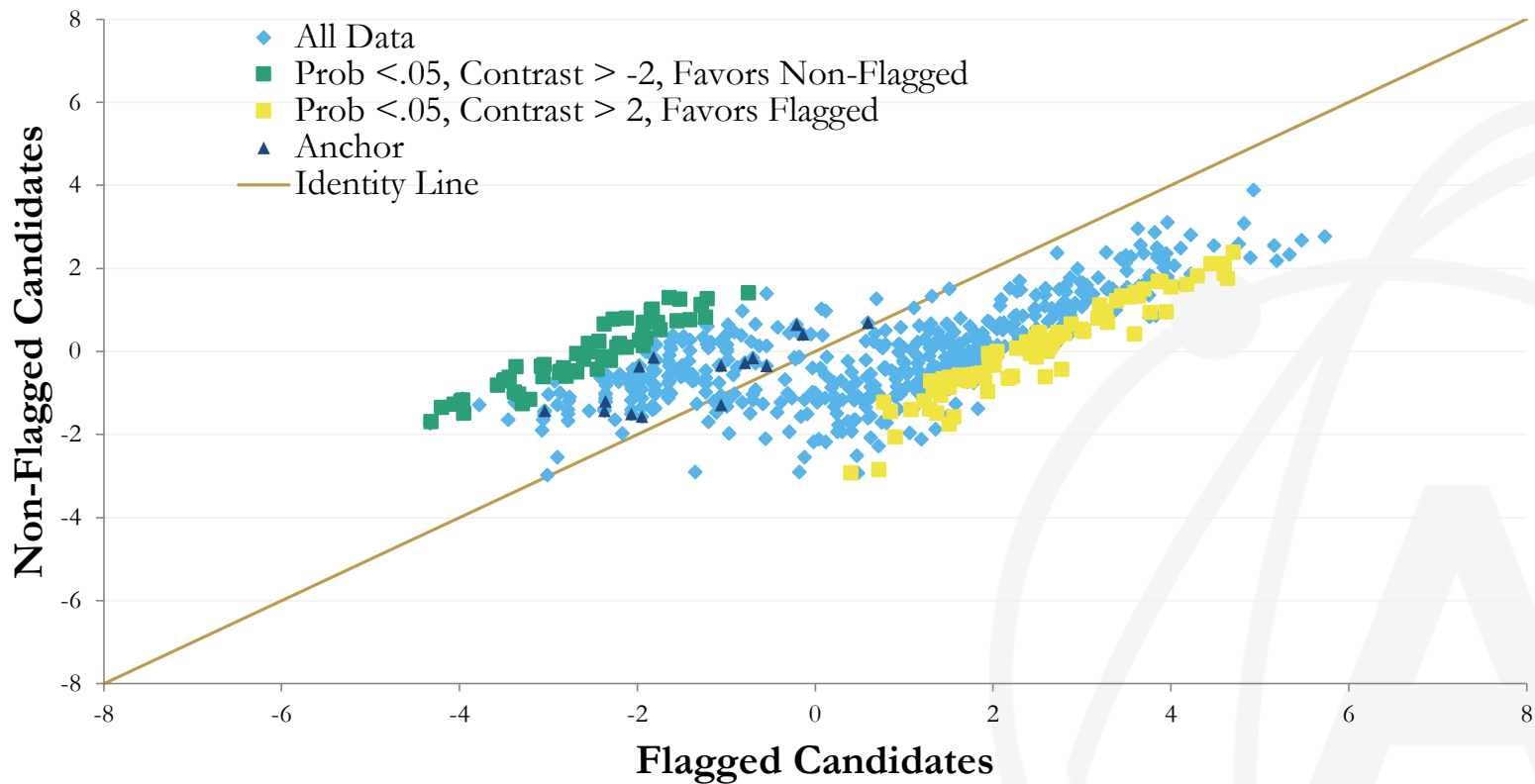
Differential Person Functioning

Differential Person Functioning



Differential Item Functioning

Differential Item Functioning



Example: Automated Candidate Flagging Criteria Included in Scoring

Security Flag	Suspect Action
Exam Retakes	Candidate takes same exam X times within Y period
Rapid Exam Completion	X% of items completed in less than Y seconds
Retake After Pass	Candidate takes the same exam after already passing
Large Score Differential	Score increase by more than X%
High Score/Low time	Exam score above X%, time spent on test less than Y
Too Little Exam Time	Candidate took less than X minutes to complete the exam
Possible Collusion	Candidates at the same test center on the same date and scored within Y% of each other on the same exam
Security Items	Candidate correctly answered X security items out of Y total security items
Differential Item Performance	X% of items correct or above on 1 st item type and Y% or below correct on 2 nd item type
Watch List	Candidate is on watch list
Banned List	Candidate is on banned list

Example: Automated Candidate Flagging Criteria Included in Scoring

Security Flag	Suspect Action
Exam Retakes	Candidate takes same exam X times within Y period
Rapid Exam Completion	X% of items completed in less than Y seconds
Retake After Pass	Candidate takes the same exam after already passing
Large Score Differential	Score increase by more than X%
High Score/Low time	Exam score above X%, time spent on test less than Y
Too Little Exam Time	Candidate took less than X minutes to complete the exam
Possible Collusion	Candidates at the same test center on the same date and scored within Y% of each other on the same exam
Security Items	Candidate correctly answered X security items out of Y total security items
Differential Item Performance	X% of items correct or above on 1 st item type and Y% or below correct on 2 nd item type
Watch List	Candidate is on watch list
Banned List	Candidate is on banned list

Example: Automated Candidate Flagging Criteria Included in Scoring

Security Flag	Suspect Action
Exam Retakes	Candidate takes same exam X times within Y period
Rapid Exam Completion	X% of items completed in less than Y seconds
Retake After Pass	Candidate takes the same exam after already passing
Large Score Differential	Score increase by more than X%
High Score/Low time	Exam score above X%, time spent on test less than Y
Too Little Exam Time	Candidate took less than X minutes to complete the exam
Possible Collusion	Candidates at the same test center on the same date and scored within Y% of each other on the same exam
Security Items	Candidate correctly answered X security items out of Y total security items
Differential Item Performance	X% of items correct or above on 1 st item type and Y% or below correct on 2 nd item type
Watch List	Candidate is on watch list
Banned List	Candidate is on banned list

Questions?

Please feel free to contact me at:

Andrew.Wiley@alpinetesting.com

Some useful references

- Association of Test Publishers (ATP) Security Committee (2013, January). Assessment Security Options: Considerations by Delivery Channel and Assessment Model. Retrieved April 11, 2013 from http://www.testpublishers.org/assets/assessment_security_options-considerations_by_delivery_channel_and_assessment_model_1-23-13.pdf.
- Carson, J. D. (2013). Certification/licensure testing case studies. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 261-283). New York, NY: Routledge.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cohen, A. S. & Wollack, J. A. (2006). Test administration, security, scoring and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355-386). New York, NY: Macmillan.
- Foster, D. (2013). Security issues in technology-based testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 39-83). New York, NY: Routledge.
- Han, N. & Hambleton, R. K. (2008). Detecting exposed items in computer-based testing. In C. L. Wild & R. Ramaswamy (Eds.), *Improving testing: Applying process tools and techniques to assure quality* (pp. 423-448). New York, NY: Lawrence Erlbaum Associates.
- Lu, Y., & Hambleton, R. K. (2003). Statistics for detecting disclosed item in a CAT environment. *Center for Education Assessment Research Report No. 498*. Amherst, MA: University of Massachusetts, School of Education.

Some useful references

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, NY: Routledge.

O'Leary, L. S. & Smith, R. W. (2013, April). *Extending Differential Person and Item Functioning to aid in maintenance of exposed exams*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

O'Leary, L.S. & Smith, R.W. (in press). Detecting Candidate Pre-knowledge and Compromised Content using Differential Person and Item Functioning. In J.A. Wollack & Cizek, G.J. (Eds.), *Handbook on detecting cheating on tests*. New York, NY: Taylor & Francis.

Smith, R. W. & Davis-Becker, S. (2011, April). *Detecting suspect candidates: An application of differential person functioning analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Smith, R.W. (2004, April). *The impact of braindump sites on item exposure and item parameter drift*. Paper presented at annual meeting of the American Educational Research Association, San Diego, CA.

Veerkamp, W. J. J. & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, 25, 373-389.

Wollack, J. A. & Fremer, J. J. (2013). Introduction: The test security threat. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 1-13). New York, NY: Routledge.