ICE2012

ANNUAL EDUCATIONAL CONFERENCE

Driving Success: Setting Cut Scores for Examination Programs with Diverse Item Types

Beth Kalinowski, MBA, Prometric

Manny Straehle, PhD, GIAC

November 6 – 9, 2012 | Palm Springs, CA Westin Mission Hills Resort & Spa

Institute for Credentialing Excellence™

Objectives

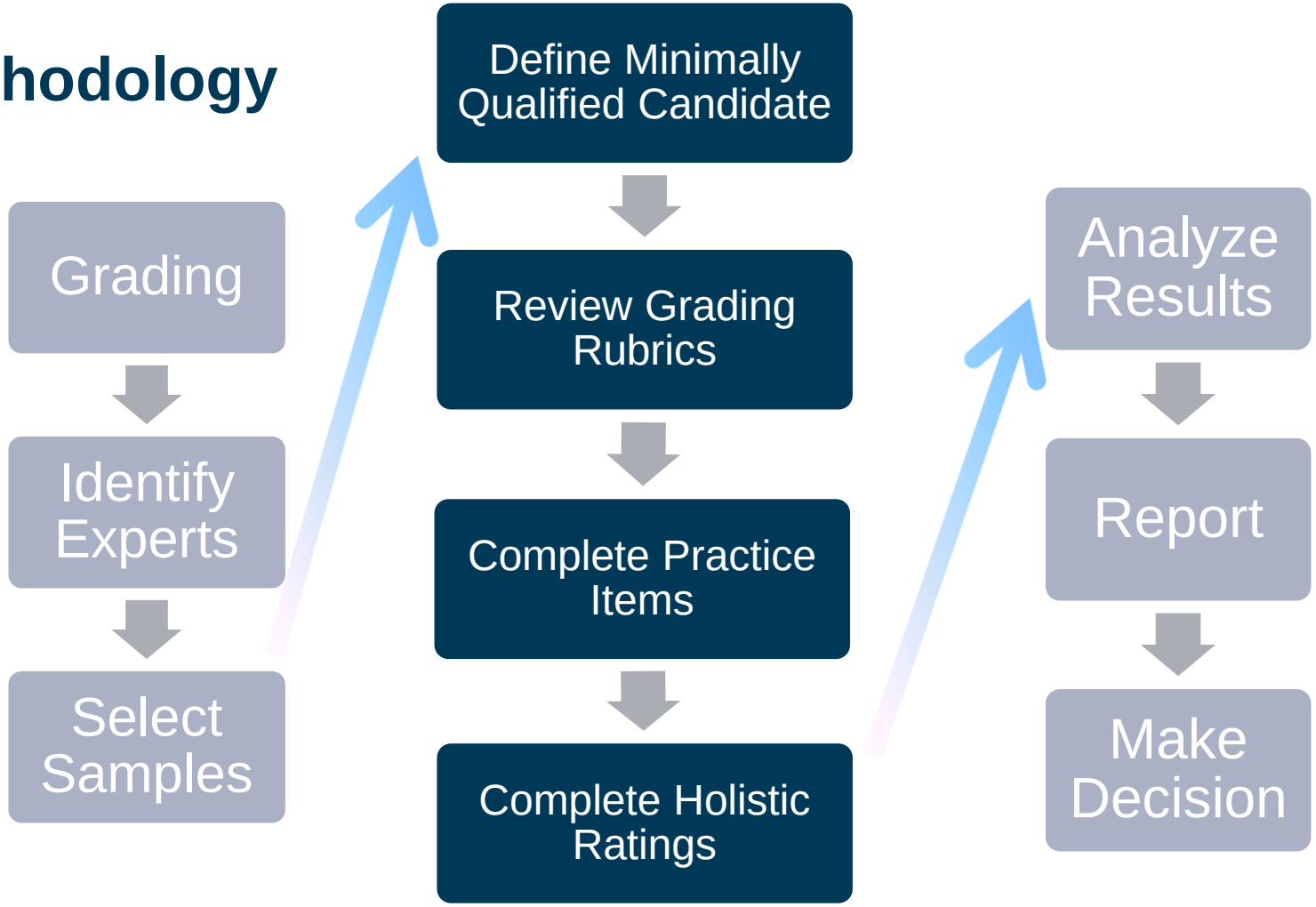
1. Describe the process for implementing a cut score methodology for a constructed response/performance-based examination
2. Compare methods for setting a cut score for different item types, such as comparing multiple choice standard setting using an Angoff with setting a cut on a performance-based examination/module
3. Identify best practices for conducting a standard setting for an examination with diverse item types, and demonstrate an understanding of how to implement cut scores for a conjunctive or a compensatory model
4. Describe results of a case study of performance based assessment (simulation with embedded questions) compared to multiple choice examination

PROMETRIC

- Conduct 3 to 4 constructed response standard settings per year
- Conduct 15 to 20 multiple choice standard settings per year

Objective 1: Methodology

Methodology



Grading

Scoring/Grading of Samples

- Needs to occur prior to standard setting

Independent Experts Scoring

- Those participating in scoring cannot participate in standard setting

Identify Subject Matter Experts

Representative

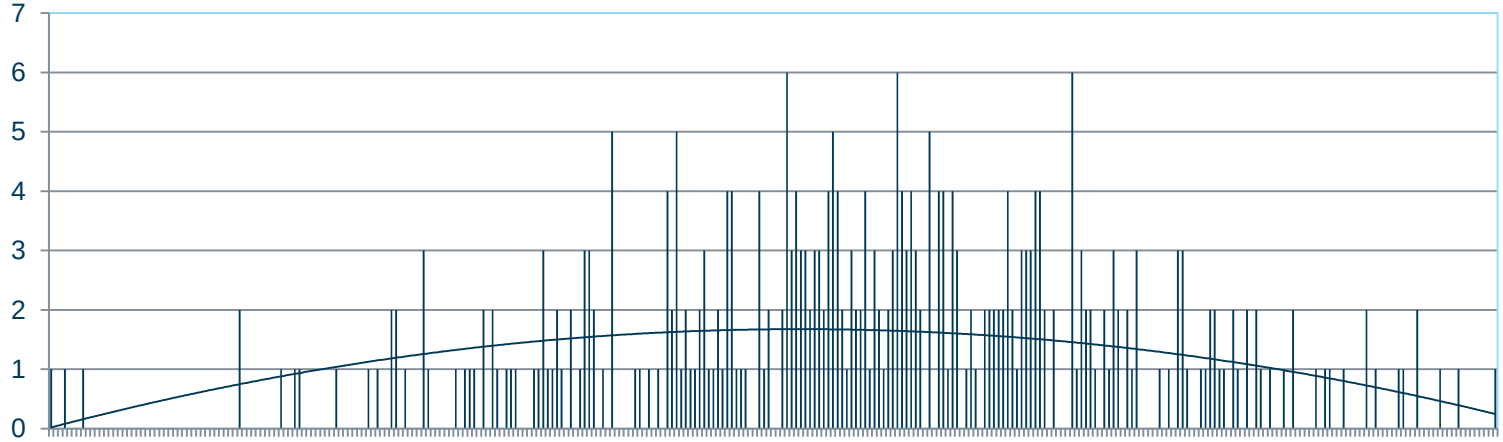
- Demographics
- Geography
- Practice-setting

Knowledgeable

- Job Requirements
- Purpose of the Test
- Candidate Group
- Preparation Materials

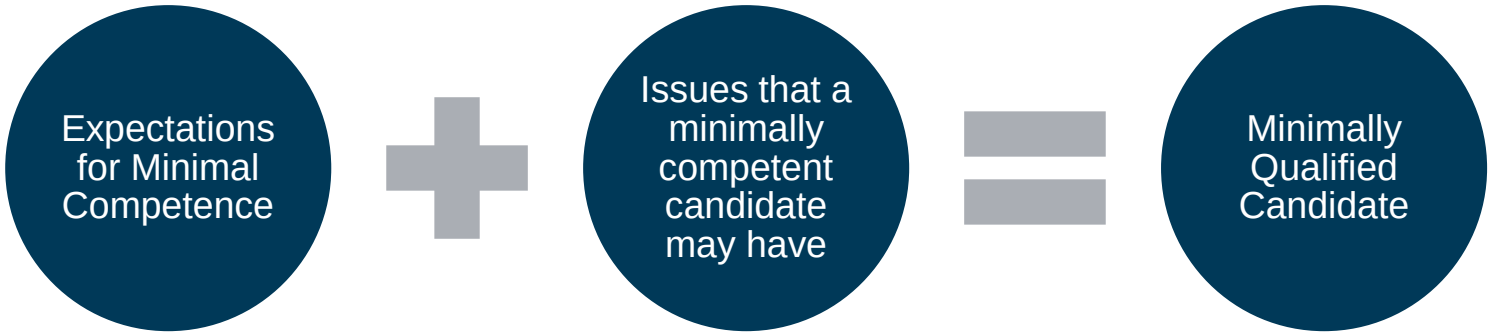
Select Samples

- Since scoring is complete, samples may be selected so that the samples' scores are spread evenly throughout the score range (of likely passing scores) or follow a normal distribution of scores
- Representative proportion



Define the Minimally Qualified Candidate

- Setting expectations on what a qualified candidate would know or be able to do
- Using the test blueprint is an effective way to delineate expectations



Review Grading Rubrics/Procedures

- Important to review the method / guidelines used to score candidate responses
- Critical that grades are not re-calculated during the judgment process

Complete Practice Items

Calibration

- Judges review the samples selected individually
- Assign “acceptable” or “not acceptable” according to the Definition of the Minimally Qualified Candidate
- Review the judgments
- Discuss the differences
- Reveal actual scores for the sample items

Complete Judgments

Judgments

- Holistically
- Independent
- Acceptable versus Not Acceptable
- Without knowledge of scores

Analyze Results

- Perform a monotonic increasing regression
  - Independent variable = candidate score (ordinal)
  - Dependent variable = judgment of acceptability (binary)

1 = Acceptable

0 = Not Acceptable

- The model fitted assumes the probability of a positive response on the binary dependent variable is a non-decreasing function of the level of the independent variable

Analyze Results

- The regression program provides:
  - Total number of observations (number of samples selected)
  - Mean level (average scores of samples)
  - Standard Deviation (variance of scores for samples)
  - Proportion of Positive Responses (number of “acceptable” judgments)
  - Point Biserial Correlation (for individual judges)
  - Mean rating (individual cut for each judge)
  - Standard Deviation on the distribution of judgments

Report Results

- The panel recommendation is the average of the individual panel-recommended cut scores
- Adjustments to the panel recommended cut score is provided by adding and subtracting the standard error for the panel
- The standard error is calculated as:

$$\frac{\sqrt{\sum_{i=1}^{n} SD_i^2}}{N}$$

Objective 2: Comparison of Methodologies

Case Study : Multiple Choice and Essays

- Program intended to measure the knowledge and skills of Chief Technology Officers in School Districts
- Administered IBT
- Two part exam
  - Part 1 : Multiple Choice, standard setting using modified Angoff procedure
    - Candidates must achieve a passing score in order to take the next part
  - Part 2 : Essay, standard setting using modified Contrasting Groups / Body of Work hybrid

Part 1: Multiple Choice Cut Score

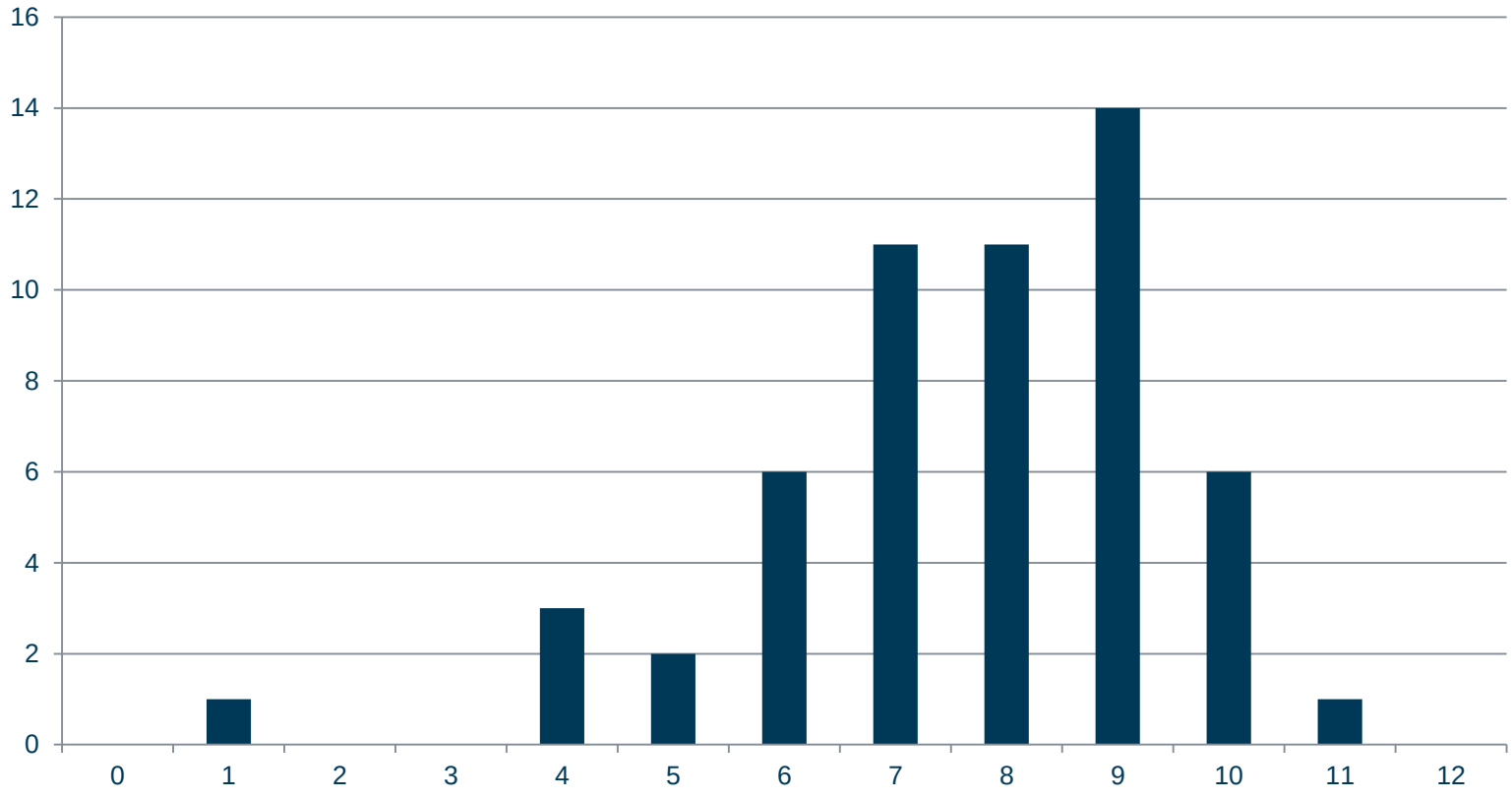
- Passing score was developed using a Modified Angoff procedure
  - 100 scored questions
  - 70 candidates in sample
  - Score range 67-90

Orientation → Definition of the Minimally Qualified Candidate → Practice Ratings → Independent Ratings → Discussion of First Ratings → Submission of Second Ratings

Part 2: Constructed Responses

- Four essays covering three domains of practice
- Scored by members of the governance board
  - Finalized the rubrics
  - Crafted scoring metrics
    - 0 = Blank or inappropriate response
    - 1 = Not qualified
    - 2 = Minimally met
    - 3 = Outstanding
- Each essay scored individually

Score Distribution



Standard Setting Procedure

- 25% of candidate responses used as samples
  - 5 judges, considered experts
- Definition of Minimally Qualified Candidate developed
- Calibration
  - Four candidate responses used
  - Scores shared
- Independent evaluation of responses

Objective 3: Scoring Models

Determining Scoring

- Mix of compensatory and conjunctive
- Must pass Part 1
- Must achieve combined scores for Part 2

Objective 4: Case Study


GLOBAL INFORMATION ASSURANCE CERTIFICATION
GIAC
www.giac.org

GIAC Certified Intrusion Analyst (GCIA)

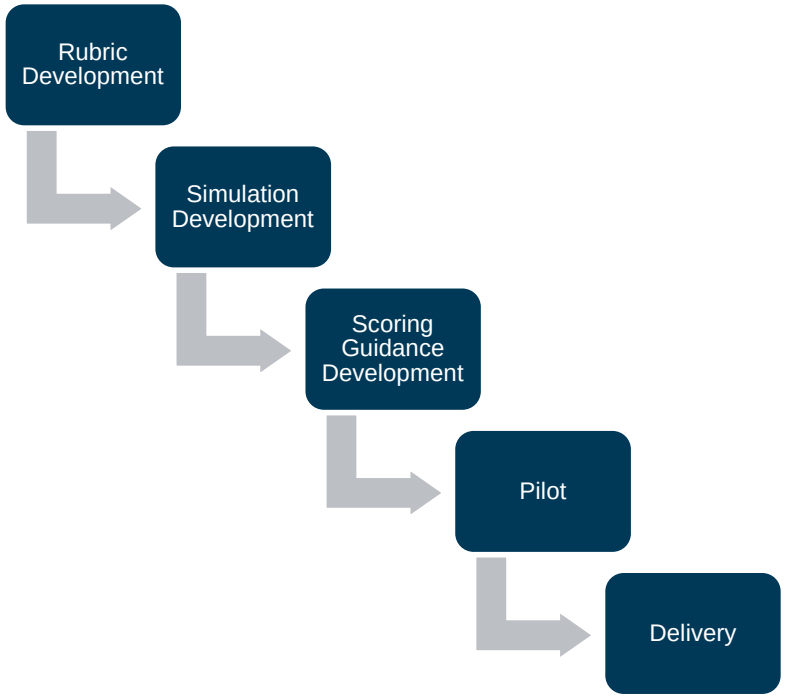
- -One of over 20 GIAC exams
- -150 MCQ exam
- -Low volume

GIAC Leadership's Question

Does the GCIA exam measure minimal competency or beyond?

Can we use a simulation that represents the workplace to determine minimal competency or beyond?

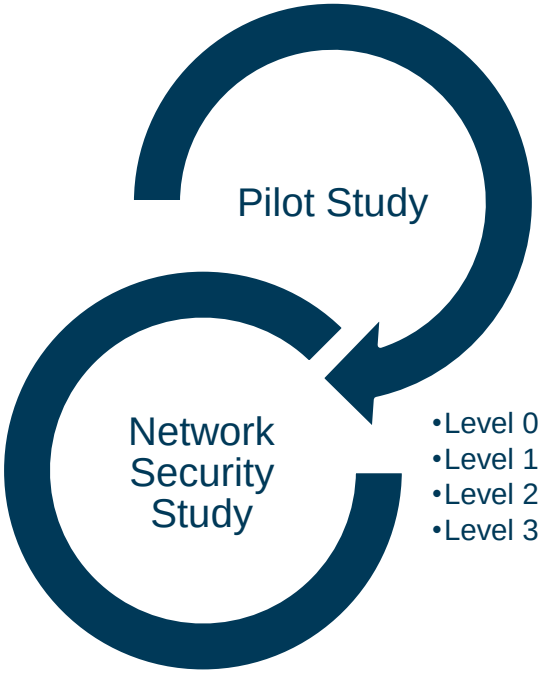
Methodology



Rubric Development

- Analytical rubric rather than holistic
- Used internal and external SMEs
- Based on GCIA certification objectives
- Began with three levels but expanded to four levels
- Resulted in 14 rubric areas

Rubric Levels



- Entry/Beginner Level
- Intermediate
- Expert

- Level 0 - Below Minimally Competency
- Level 1 - Minimal Competency
- Level 2 - Above Minimal Competency
- Level 3 - Expert

Rubric Example: Final Version

Exam Certification Objectives	Certification Objective Outcome Statement	Level 0	Level 1		Level 2		Level 3	
		---	Does Know How To:	Does Not Know How To:	Does Know How To:	Does Not Know How To:	Does Know How To:	Does Not Know How To:
E. Transport Layer Essentials	**5. The candidate will demonstrate an understanding of TCP and UDP headers and usage.**	Cannot demonstrate any tasks listed in Level 1, 2 or 3	1. Purpose of the most commonly used fields (e.g., source and destination port) 2. Purpose of the most commonly used flags (e.g., SYN, ACK, FIN, RST) 3. Analyze and identify which packets are TCP and UDP 4. Select the TCP and UDP packets using a protocol analyzer 5. Basic properties of TCP and UDP 6. Use Protocol analyzer	1. Interpret tcpdump output thoroughly and quickly 2. Use commonly used flags, window, options, and checksum (e.g., URG, PSH, ECN, SACK) 3.Understand TCP options (compute check sum) 4. Easily identify the packet artifacts (source port, destination port, and flags, TCP options) without a Protocol Analyzer	1. Determine transport layer header **without** using a protocol analyzer 2. Purpose and common use of all the fields in the TCP header 3. Use of each of the TCP fields in the header and their common use 4. Read a tcpdump output of a packet capture and determine the major artifacts (e.g., header size, embedded protocol) 5. Interpret the embedded protocol of a TCP packet and the payload using tcpdump.	1. Easily decode a TCP header using the hex output 2. Easily interpret the meaning of TCP header options 3. Easily determine the nature of malicious or unusual TCP packet headers 4. Modify cap packet captures using a hex editor	1. Decode the TCP headers using only the hex code 2. Understand malicious and abnormal use of the TCP options, checksum, flags, and fragmentation 3. Interpret the embedded protocol of a TCP packet and the payload using hex dump.	1. Rewrite the TCP stack of an operating system 2. Create better transport layer protocols than TCP and UDP

Simulation & Embedded Questions

External SMEs developed simulated packet traffic work place scenarios based on rubric

Developed and embedded questions into simulations based on

- MCQ
- Fill-in-the-blank

Provided scoring guidance

- Minimally qualified if they answer 3 of 4 correct for this simulation exercise

Delivery: Pilot

- July 2012
- 3-hour pilot study
- 5 test subjects

- Two evaluators
- Rated each test subject
- Using the scoring guidance and rubric
- Calibrate - discussions after each test subject to obtain consensus often by revising exercises and rubrics

Interrater Reliability

- Average ICC .512 (moderate agreement)

Improvements Prior to Official Launch

Simulation	• Included **MORE** scenarios and items aimed at minimally competency rather than intermediate and expert
Evaluators	• Used **FOUR** evaluators instead of **TWO** • Training and discussions occurred prior to official launch
Participants	• Better incentives to attract more candidates
Background information	• Online survey collected information on **how often** participants used the current rubric areas in their workplace
Delivery	• Over two-day period was able to evaluate more candidates and extended the delivery from 3 to 4 hours

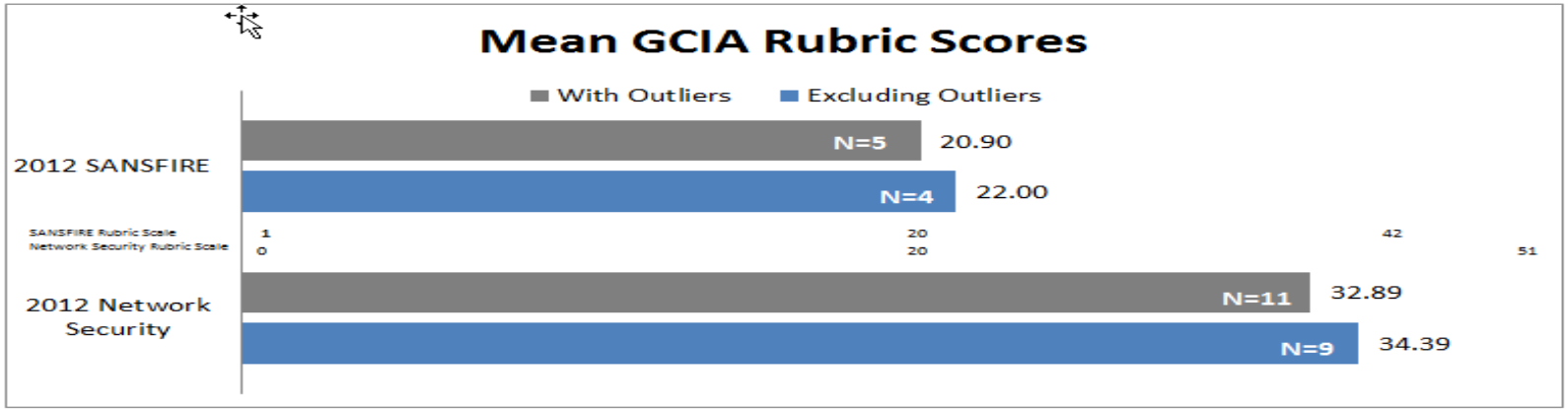
Delivery: September 2012 Study

- September 2012
- 4 -hour pilot study
- 11 test subjects

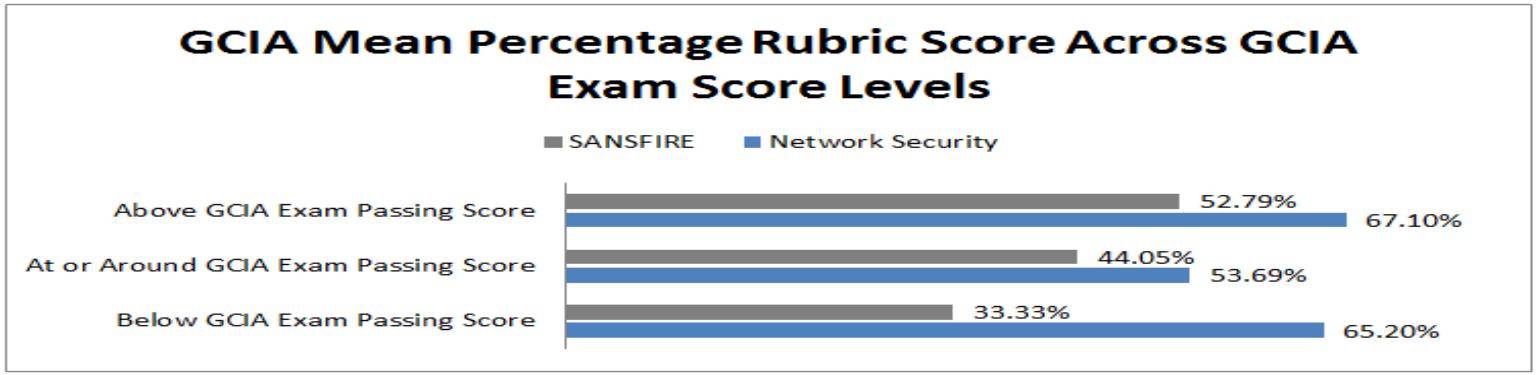
- Four evaluators
- Rated each test subject
- Used the scoring guidance and rubric

Interrater Reliability

- .753 (strong agreement range)



*2012 SANSFIRE rubric scale had a range from 1 to 42; 2012 Network Security rubric scale had range from 0 to 51.



*GCIA is investigating the one candidate who scored low on the GCIA exam (failed) and contributed to the high GCIA mean percentage rubric score of 65.20%. GIAC staff is confirming some possibilities such as this candidate may have been recently studying for the GCIA exam, has gained intrusion analyst worked experience, or other factors.

Data Transformations

Two different rubrics

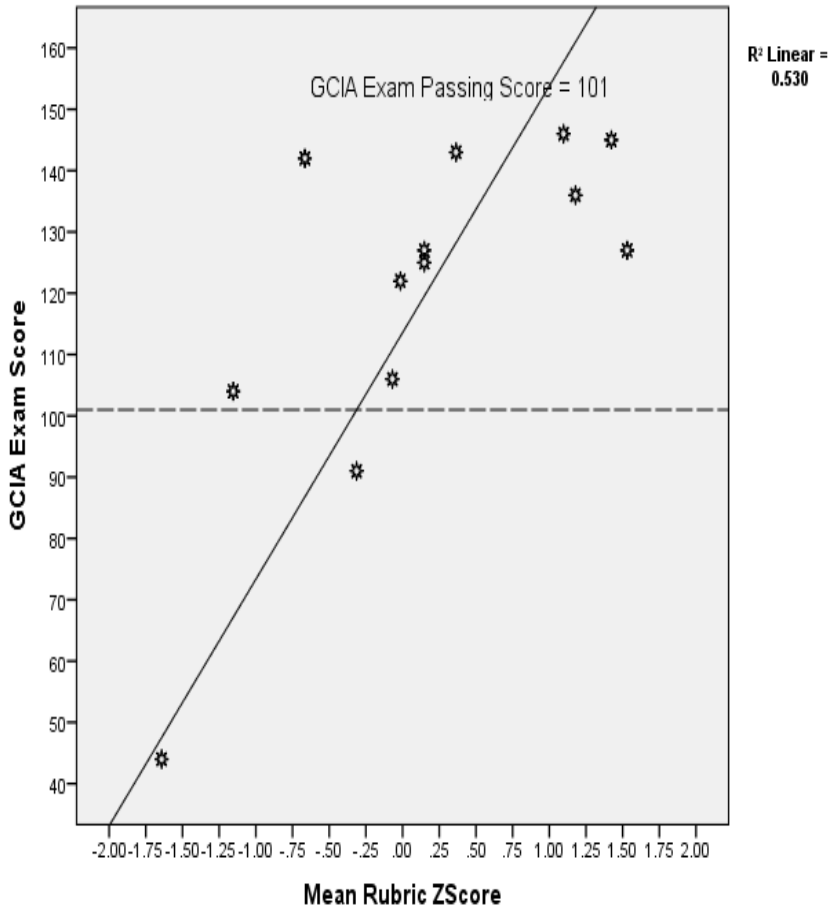
- Converted Mean Rubric Scores to Z-Scores

Excluded three outliers

Key Preliminary Results:

↑ GCIA Exam Scores = ↑ Demonstrated Intrusion Analyst Competency

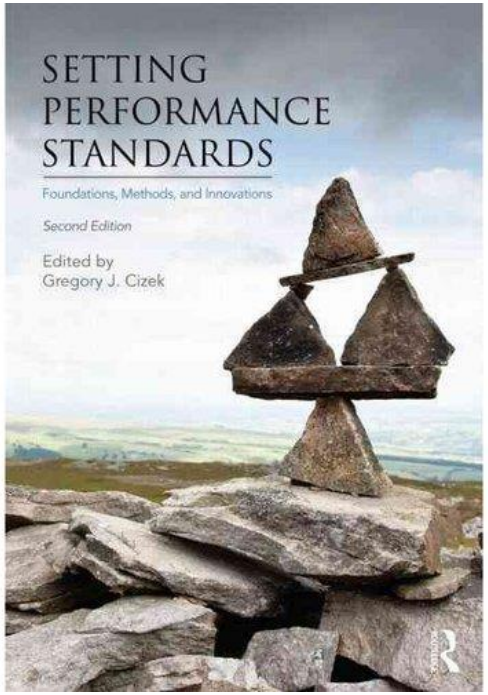
r = .728 (p = .005)

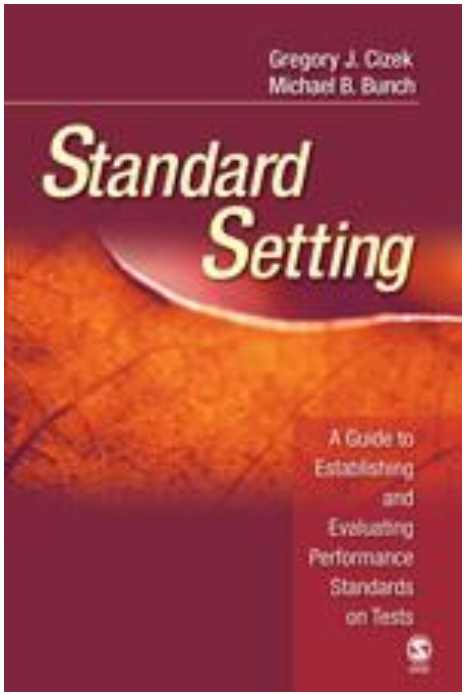


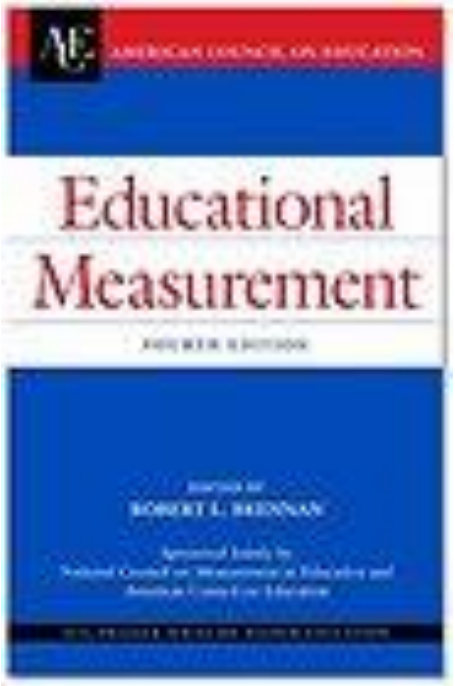
Next Steps

- 2013 – Run Study Again
- Recruitment Intelligence
- Increase Sample Size

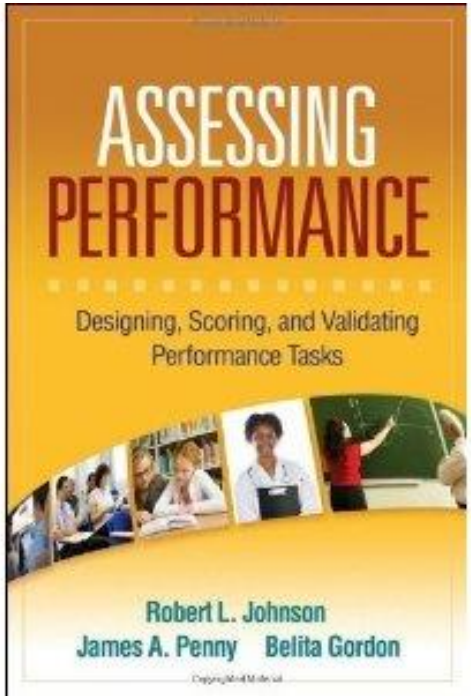
References

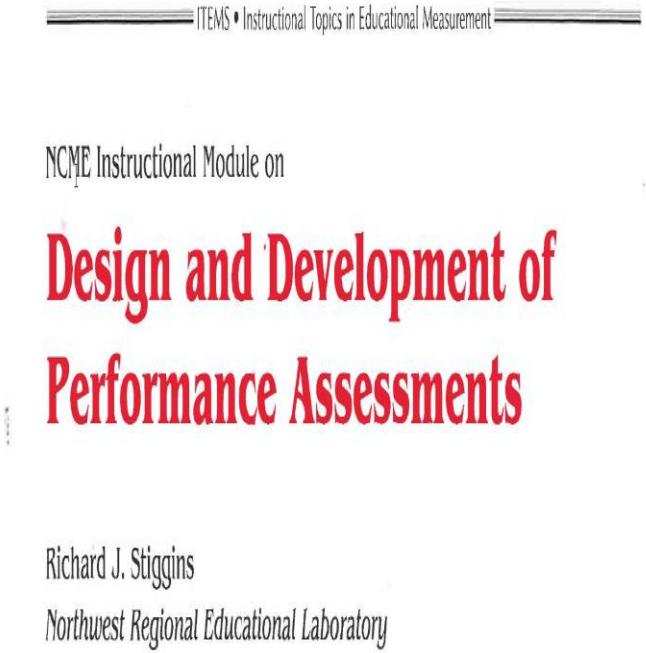

SETTING
PERFORMANCE
STANDARDS
Foundations, Methods, and Innovations
Second Edition
Edited by
Gregory J. Cizek


Gregory J. Cizek
Michael B. Bunch
Standard
Setting
A Guide to
Establishing
and
Evaluating
Performance
Standards
on Tests


Educational
Measurement

References




ITEMS • Instructional Topics in Educational Measurement

NCME Instructional Module on

Design and Development of Performance Assessments

Richard J. Stiggins

Northwest Regional Educational Laboratory