

# Setting a Cut Score – What's Fair and What's Not

(But I got a 60%! Why didn't I pass?)

**Jim Parry, M.Ed., CPT**

*Owner/Chief Executive Manager*

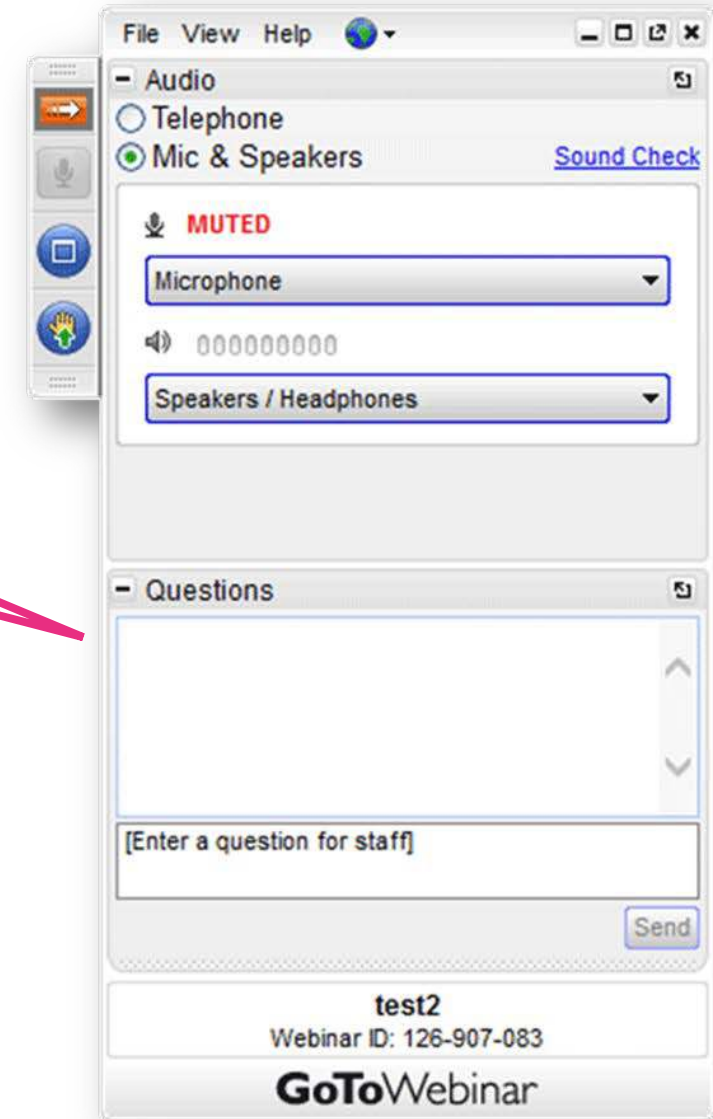
Compass Consultants, LLC

September 27, 2022

To ask questions,  
use the “Questions”  
feature

**Watch for an email after the webinar:**

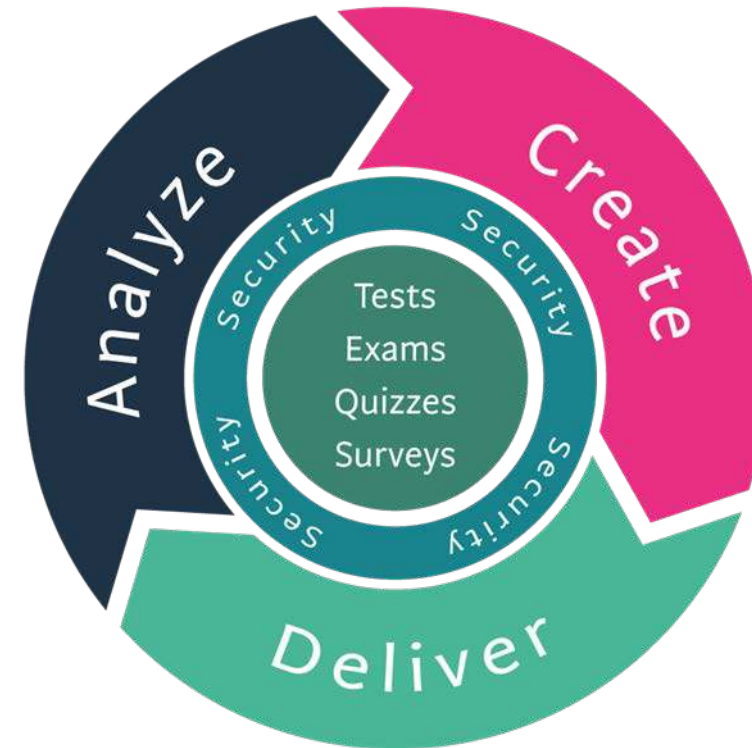
- Download slides (PDF)
- View a recording
- Answer a survey



# About Questionmark

## Background

- Founded in 1988
- Assessment solutions to measure knowledge, skills, abilities and attitudes securely for certification, regulatory compliance, workforce learning, sales-force readiness and higher education
- ISO/IEC 27001 Certified (Learn more: [www.questionmark.com/trust](http://www.questionmark.com/trust))



- *Questionmark OnDemand*
- *Questionmark OnDemand for Government*
- *Questionmark OnPremise*

# Today's Presenter

Jim Parry, M.Ed., CPT, Compass Consultants, LLC

- Owner and Chief Executive Manager of Compass Consultants, LLC
- Over 40 years experience in course design, development, and presentation and assessment design, development, and analysis
- Holds a Master of Education degree from the University of West Florida and is a Certified Performance Technologist (CPT), awarded by the International Society of Performance Improvement (ISPI)
- Has been presenter of pre-conference workshops and educational sessions at various professional conferences for many years
- Internationally recognized consultant providing services concerning test design, development, establishment of cut scores, and analysis
- Jim is a consulting partner of Questionmark



# About Compass Consultants, LLC

## Background

- Founded in 2010
- A leader in the application of Human Performance Technology (HPT), specializing in the design, development and presentation of training interventions and the psychometrics of test development and analysis.
- Learn more:  
[www.gocompassconsultants.com](http://www.gocompassconsultants.com)



# Today's Agenda

- 
- Establishing Defensible Cut Scores
  - Dangers Associated with Arbitrary Cut Scores
  - Extension of The Modified Angoff Method
  - Maintaining Difficulty & Content Across Tests
  - Designing Defensible Randomized Tests

# Legal Disclaimer

- This presentation may include information about legal issues and legal developments. Such materials are for informational and/or educational purposes only and may not reflect the most current legal developments. These informational/educational materials are not intended, and should not be taken, as legal advice on any particular set of facts or circumstances. You should contact an attorney for advice on specific legal problems or questions.
- Information is provided "as is" without any express or implied warranty of any kind including warranties of merchantability, noninfringement of intellectual property, or fitness for any particular purpose. In no event shall Compass Consultants, LLC., or its agents, officers or attorneys be liable for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information) arising out of the use of or inability to use the information, even if Compass Consultants, LLC has been advised of the possibility of such damage.

# How Can All Tests be Fair?

- Test-items must be constructed correctly
- Must be unbiased
- Must be directed to the correct population
- Cut/passing score must be defensible
- Must be valid – test the right content
- Must be reliable – repeatable results
- Parallel tests must test same content and be same difficulty



This Photo by Unknown Author is licensed under CC BY



# Quick Poll

## How does your organization set or determine a cut or passing score for an assessment?

- A. I/we use an arbitrary value such as 60% is a "D", 70% is a "C", etc.
- B. I/we set a cut or passing score using a recognized method such as the Modified Angoff Method.
- C. I/we do not set a cut or passing score – assessments are for self-check/study purposes only.



# Is 60% Correct Good Enough?



— a Learnosity company —



# The Arbitrary Cut Score

“Because I think that’s what it should be!”

# Who Decides Who Passes?





# Setting The Expected/Arbitrary Passing Score



This Photo by Unknown Author is licensed under [CC BY-SA](#)

# Why is the Arbitrary Score Used? Is it Fair?

- Historical precedent
  - It's always been a 70%
- State learning standards dictate
  - Common denominator
- Subject of debate
  - How is it fair?
  - Are all schools teaching to same standard?
- Sometimes recalibrated
  - Not enough pass
- Could be biased
  - Teacher omits difficult items



[This Photo](#) by Unknown Author is licensed under [CC BY](#)

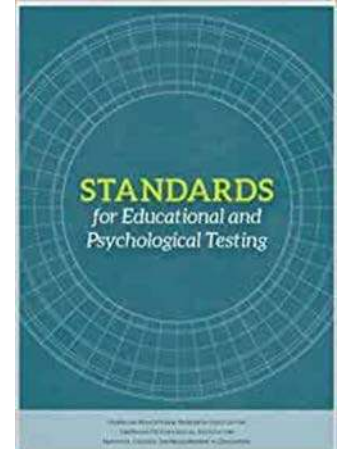
# The Importance of a Defensible Cut Score



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

# What Makes a Cut (Passing) Score Defensible?

- Based on Minimal Acceptable Competence (MAC) level
- Designed to result in a cut or pass point that represents the threshold between those candidates who can do the job and those who cannot
  - Master vs. Non-Master
- When cut scores are used they should be set as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force
- To be legally defensible and meet the Standards for Educational and Psychological Testing, a cut score cannot be arbitrarily determined, it must be empirically justified

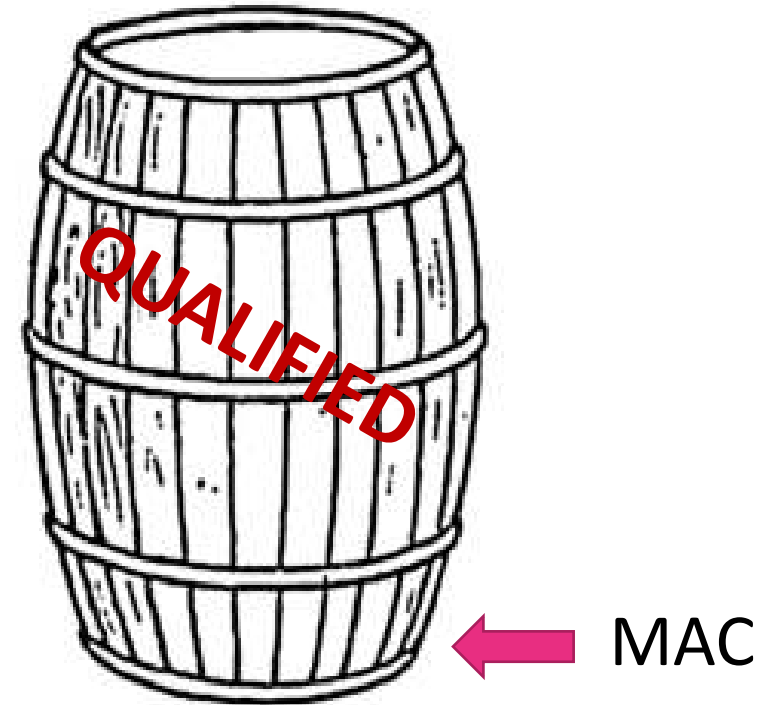




# Minimum Acceptable Competence (MAC) Level

- The level of performance on the test indicative of minimal competence
  - Bare minimum – the bottom of the **qualified** barrel
  - This is NOT the *best* or *most* qualified

Apprentice  
Journeyman  
Master



# Establishing Cut Scores

- Cut/pass score judgments must be:
  - Made by persons who are qualified to make them
  - Meaningful to the persons who are making them
  - Made in a way that takes into account the purpose of the test
- Cut scores may be set as high or as low as needed to meet organizational requirements
- Establishing cut scores involves professional judgments as well as technical and empirical considerations
- Should use a sufficiently large and representative group of judges to ensure validity
- Procedure used must be documented

# Use Caution!

- When a test is used to classify test-takers into two groups, two kinds of wrong decisions can occur:
  - A test test-taker who actually belongs in the lower group can get a score above the passing score
  - A test-taker who actually belongs in the higher group can get a score below the passing score

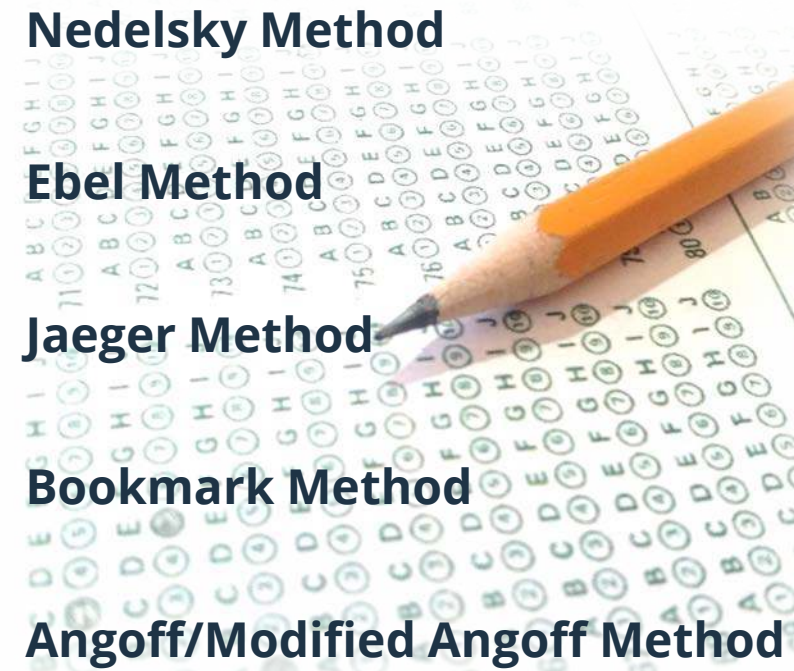


Livingston, S.A & Zieky, M.J., (1982)

# Some Recognized Methods

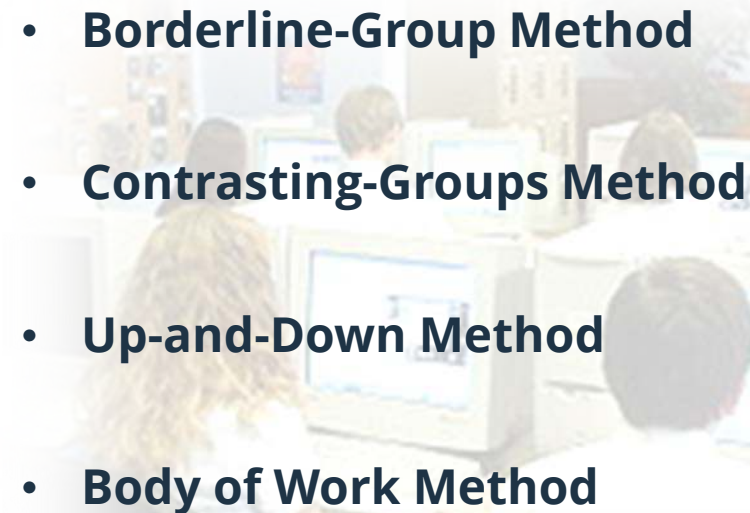
## Test Centered

Methods based on judgements about test questions

- **Nedelsky Method**
  - **Ebel Method**
  - **Jaeger Method**
  - **Bookmark Method**
  - **Angoff/Modified Angoff Method**
- 

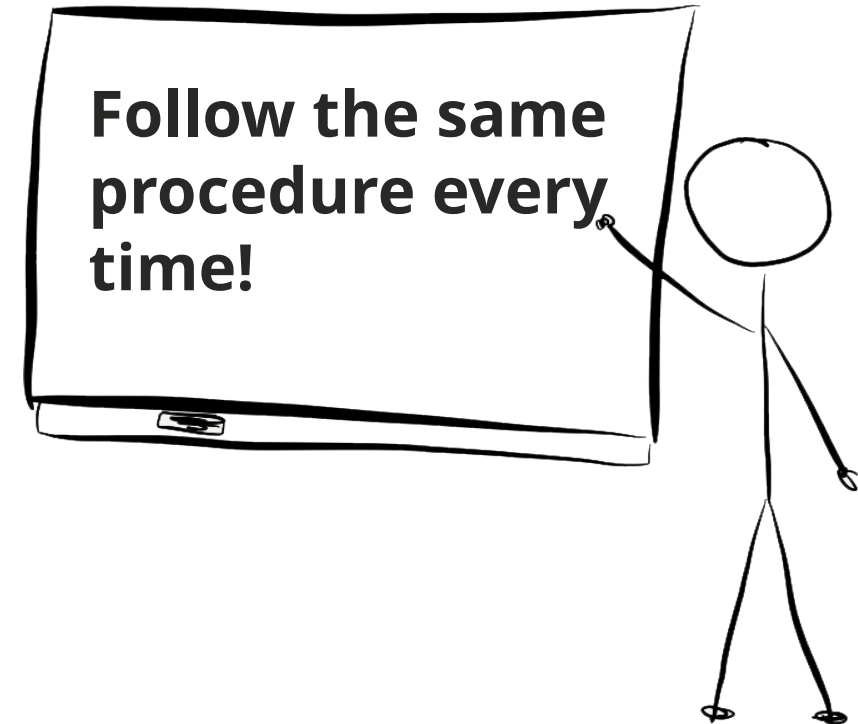
## Test-Taker Centered

Methods based on judgments about a group of test-takers

- **Borderline-Group Method**
  - **Contrasting-Groups Method**
  - **Up-and-Down Method**
  - **Body of Work Method**
- 

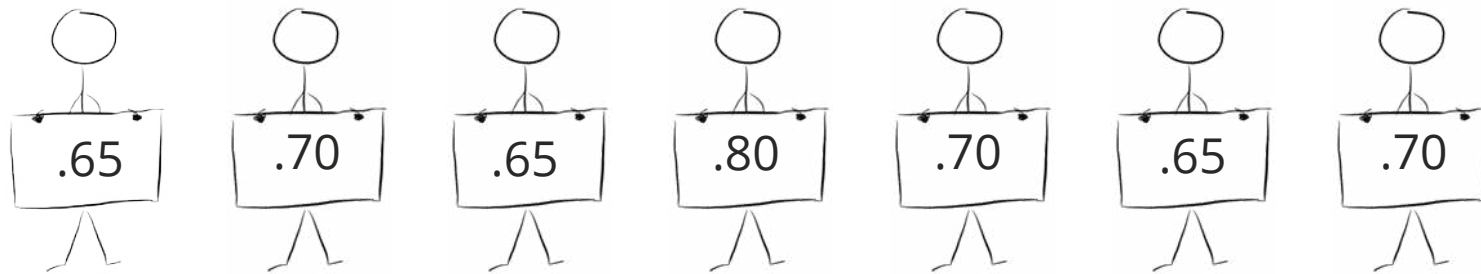
# Which Method is Best?

- It Depends!
  - Modified Angoff is most widely used
- Use whichever method or combination that suits your test format
  - Dichotomous-scored items
    - Right/wrong, true/false, etc.
  - Polytomous-scored items
    - Likert-type items, partial credit, etc.
- Important to document
- Follow same procedure every time



# Angoff

- Angoff Method
  - Item performance probability determined by panel of expert judges
    - Will MAC respond correctly? (Yes/No)
  - Item probabilities summed
- Modified Angoff Method
  - Item necessity and difficulty levels determined
  - Item performance probability determined (0.1 – 1.0)
  - Results calculated
  - Combination of Angoff and Ebel methods



“It is impossible to prove that  
a cut score is correct.”

ETS – A Primer on Setting Cut Scores on Tests of Educational Assessment

# Warning!

Calculated cut score may be modified by HR/Management requirements and set higher or lower to meet organizational needs!

# Document This!



# Alternate and Retests

Missed test day or failed first test

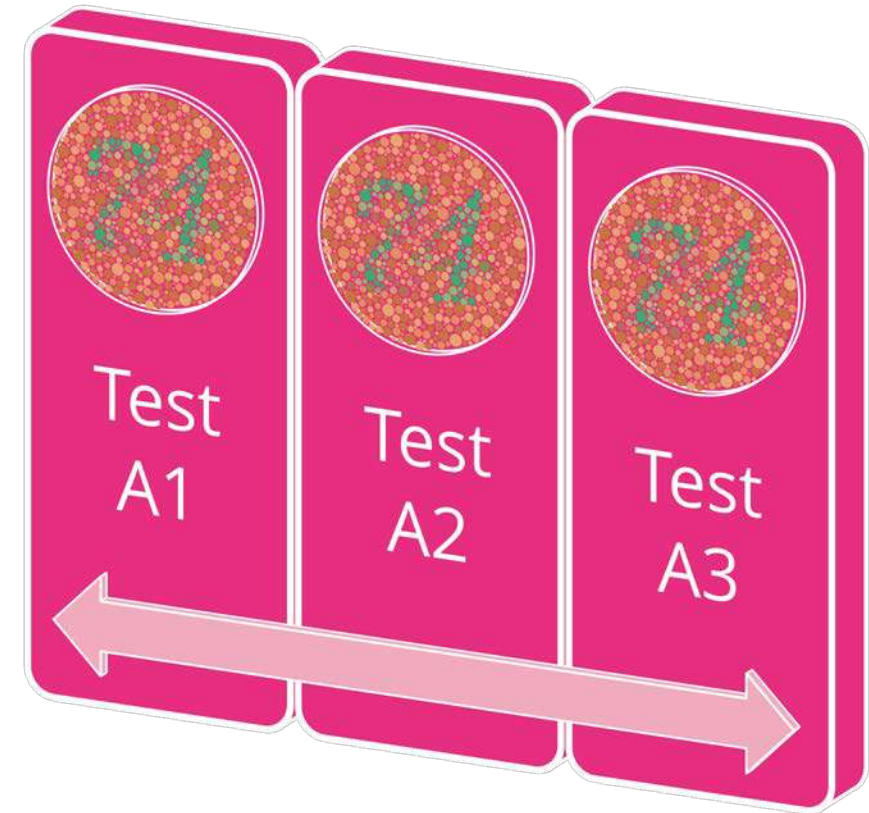
# Quick Poll

**How do you or your organization design alternate or retests?**

- A. I/we use the same test as the original – there is only one version
- B. I/we use the same test questions as the original but mix them up
- C. I/we generate a new test by randomly picking questions
- D. I/we generate a new test using stratified-randomization
- E. I/we do not offer alternate or retests

# Alternate and Retests must be parallel!

Content and difficulty must match to maintain fairness



# Successful Fair Test Design



Design  
Item  
Database

Establish  
Difficulty &  
Cut Score

Set Item  
Selection  
Criteria

# Design Test Item Database

The First and Most Important Step!



This Photo by Unknown Author is licensed under [CC BY](#)

# Topic Structure

- Repository Name
  - Objective 1.0
    - Topic 1.1
      - *Sub-Topic 1.1.1*
        - Test-Item 1.1.1/1
        - Test-Item 1.1.1/2
        - Test-Item 1.1.1/3
        - Test-Item 1.1.1/4
        - Test-Item 1.1.1/5
        - Test-Item 1.1.1/6
  - Objective 2.0
    - Topic 2.1
      - *Sub-Topic 2.1.1*
        - Test-Item 2.1.1/1
        - Test-Item 2.1.1/2
        - Test-Item 2.1.1/3
        - Test-Item 2.1.1/4
        - Test-Item 2.1.1/5
        - Test-Item 2.1.1/6

# Modified Angoff Method

Establish Item Difficulty and Set Cut Score



[This Photo](#) by Unknown Author is licensed under [CC BY](#)

# Select Raters

- Familiar with competencies/objectives covered by the test and with performance level for masters of these competencies/objectives
  - 5 is minimum, 8-10 maximum
- Diverse group (geographic location, age, gender, race, etc.)
- Proficiencies of raters:
  - Familiar with tasks the test will assess
  - Knowledge of skill sets of persons who will perform the tasks
  - Ability to pass existing test at current cut score (if any)
  - Ability to edit test-items for clarity, accuracy, spelling, and grammar



# Gather Raters/Judges

- Conduct face-to-face meeting
  - Virtual meeting is acceptable – use caution
- Raters “take” the test under same conditions as a “real” test-taker would<sup>1</sup>
  - Establishes a ceiling score – the highest score/rating each item can be assigned
    - Experts can only achieve this score so MAC can’t be expected to exceed
  - Raters provide feedback on wording, design, and accuracy of each item

<sup>1</sup> In the case of a large test item database it may not be practical for the raters to complete the entire item bank due to time constraints so this step may be omitted and noted in the test plan.

(Parry, 2017)

# Define MAC

- Judges come to consensus regarding definition of “minimally acceptable candidate” (MAC)
  - One who performs the task on the job; **not** a student
  - One who has the least amount of education and experience necessary to perform the task
  - One who meets standards, though barely
  - One whose task performance is borderline, but acceptable
  - In addition to the criteria listed above, factors specific to the job/tasks may be introduced to further identify a minimally qualified performer

Apprentice

Journeyman

Master

# Explain Process

- Estimation process explained
  - Probability estimate can never be less than .25 (25%) if there are 4 choices for a multiple-choice question
    - This is minimum value due to chance guess
    - A 3-response item would have a .33 minimum value
    - T/F & Y/N would be .50 minimum
- Establish “allowable” percentages
  - Various philosophies
  - Theoretically range from 0 to 1.00
- Widely acceptable to have “set” ranges
  - .25, .30, .35, .40, .45, .50, .55, .60, .65, .70, .75, .80, .85, .90, .95

# Process and Execution

- Estimate the difficulty of each item at the minimally competent test-taker level - NOT the level as a rater/judge (expert)
  - **Apprentice** – new staff member, entry level, may need direct supervision
  - **Journeyman** – fully effective, can work alone
  - **Master** – tasks are second nature, person has mastered their role
- Do NOT estimate the level of a typical test-taker – think of the minimally competent person who meets the minimum standard for job, competence, certification, etc.
- Set the standard at which the minimally competent performer should be able to answer
- Raters/judges do NOT discuss ratings of each item at this point
  - Read each stem, correct answer and distractor carefully
- Ratings are recorded by each rater/judge for each item

# Record and Discuss

- After all items are rated and recorded if any vary among judges by more than Standard Deviation (SD) of 10 they should be discussed
  - Weights can be changed as a result of the discussion or original estimates can be retained

CUT SCORE CALCULATION TOOL

Course/Certification Name:		FAIRNESS RESEARCH 3			Test Name:		TEST NAME							
Facilitator Name/Date:		Facilitator Name/Date			Revision 1 Facilitator Name/Date						Date: mm/dd/yyyy			
Enter Topic/TPO/Subject ID:		Topic 1			Revision 2 Facilitator Name/Date									

This spreadsheet tool is the intellectual property of and Copyright ©2010-2019 by Compass Consultants, LLC. Use is limited to the terms of the End User License Agreement (EULA). This copy is licensed to: 30 DAY DEMO ONLY

Test Item QID	Enter * If New, R if Retired	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
1.0 M1		Moderate	48.57	60	60	40	45	40	55	40				9.45
1.0 M2		Moderate	62.14	55	70	60	75	65	60	50				8.59
1.0 M3		Moderate	57.14	50	60	70	60	60	60	40				9.51
	R			50	65	60	65	60	60	40				
1.0 M4		Moderate	62.86	70	70	70	60	60	60	50				7.56
1.0 M5		Moderate	60.00	60	70	70	50	50	70	50				10.00
1.0 E1		Easy	77.14	70	85	80	75	70	90	70				8.09
1.0 E2		Easy	75.00	70	80	90	75	70	80	60				9.57
1.0 E3		Easy	77.14	70	80	90	75	75	80	70				6.99
1.0 E4		Easy	84.29	75	90	95	90	75	90	75				8.86
1.0 M6		Moderate	62.14	55	70	50	50	70	70	70				9.94
1.0 H1		Hard	41.43	30	50	35	35	50	50	40				8.52
1.0 M7		Moderate	63.57	60	65	70	50	65	65	70				6.90
1.0 M8		Moderate	55.00	55	60	65	50	50	65	40				9.13
	R			70	80	75	80	70	90	40				
1.0 M9		Moderate	57.14	50	50	70	50	50	70	60				9.51
1.0 M10		Moderate	48.57	50	45	50	35	55	55	50				6.90
1.0 M11		Moderate	57.86	50	70	50	50	65	60	60				8.09

Compass Consultants, LLC

Topic Cut Score	58.00	Moderate Difficulty
-----------------	-------	---------------------

Approximate Difficulty Rating

25 - 48.3 : Hard

48.4 - 71.7 : Moderate

71.8 - 95 : Easy


Standard Deviation

A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

7	Easy	In this section	17%
28	Moderate	In this section	67%
7	Hard	In this section	17%
42	TOTAL		100%

Parry, J.R. (2020)

CUT SCORE CALCULATION TOOL														
Course/Certification Name:			FAIRNESS RESEARCH 3				Test Name:			TEST NAME				
Facilitator Name/Date:			Facilitator Name/Date				Revision 1 Facilitator Name/Date			Date: mm/dd/yyyy				
Enter Topic/TPO/Subject ID:			Topic 1				Revision 2 Facilitator Name/Date							
This spreadsheet tool is the intellectual property of and Copyright ©2010-2019 by Compass Consultants, LLC. Use is limited to the terms of the End User License Agreement (EULA). This copy is licensed to: 30 DAY DEMO ONLY														
Test Item QID	Enter * If New, R if Retired	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
1.0 M1		Moderate	48.57	60	60	40	45	40	55	40				9.45
1.0 M2		Moderate	62.14	55	70	60	75	65	60	50				8.59
1.0 M3		Moderate	57.14	50	60	70	60	60	60	40				9.51
	R			50	65	60	65	60	60	40				
1.0 M4		Moderate	62.86	70	70	70	60	60	60	50				7.56
1.0 M5		Moderate	60.00	60	70	70	50	50	70	50				10.00
1.0 E1		Easy	77.14	70	85	80	75	70	90	70				8.09
1.0 E2		Easy	75.00	70	80	90	75	70	80	60				9.57
1.0 E3		Easy	77.14	70	80	90	75	75	80	70				6.99
1.0 E4		Easy	84.29	75	90	95	90	75	90	75				8.86
1.0 M6		Moderate	62.14	55	70	50	50	70	70	70				9.94
1.0 H1		Hard	41.43	30	50	35	35	50	50	40				8.52
1.0 M7		Moderate	63.57	60	65	70	50	65	65	70				6.90
1.0 M8		Moderate	55.00	55	60	65	50	50	65	40				9.13
	R			70	80	75	80	70	90	40				
1.0 M9		Moderate	57.14	50	50	70	50	50	70	60				9.51
1.0 M10		Moderate	48.57	50	45	50	35	55	55	50				6.90
1.0 M11		Moderate	57.86	50	70	50	50	65	60	60				8.09



Topic Cut Score	58.00	Moderate Difficulty
Approximate Difficulty Rating		
25 - 48.3	Hard	
48.4 - 71.7	Moderate	
71.8 - 95	Easy	
Standard Deviation		
A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.		

7	Easy	In this section	17%
28	Moderate	In this section	67%
7	Hard	In this section	17%
42	TOTAL		100%



# Determine Cut Score and Design Assessment

Final Directed-Randomized Test Design Blueprint for:										TEST NAME							mm/dd/yyyy
Topic	Topic Cut Score & Difficulty	Items in Topic	% of Total Items	Available Hard	% From Topic	Available Mod	% From Topic	Available Easy	% From Topic	Total # Needed From Topic	Use Hard (Calculated)	Use Hard (Actual)	Use Mod (Calculated)	Use Mod (Actual)	Use Easy (Calculated)	Use Easy (Actual)	Topic
Topic 1	58	42	31.34%	7	17%	28	67%	7	17%	6.27	1.04	1	4.18	4	1.04	1	Topic 1
Topic 2	69	52	38.81%	1	2%	27	52%	24	46%	7.76	0.15	1	4.03	4	3.58	3	Topic 2
Topic 3	66	40	29.85%	2	5%	24	60%	14	35%	5.97	0.30	1	3.58	3	2.09	2	Topic 3
4.1		0	0.00%	0		0		0		0.00							4.1
5.1		0	0.00%	0		0		0		0.00							5.1
6.1		0	0.00%	0		0		0		0.00							6.1
7.1		0	0.00%	0		0		0		0.00							7.1
8.1		0	0.00%	0		0		0		0.00							8.1
9.1		0	0.00%	0		0		0		0.00							9.1
10.1		0	0.00%	0		0		0		0.00							10.1
11.1		0	0.00%	0		0		0		0.00							11.1
12.1		0	0.00%	0		0		0		0.00							12.1
13.1		0	0.00%	0		0		0		0.00							13.1
14.1		0	0.00%	0		0		0		0.00							14.1
15.1		0	0.00%	0		0		0		0.00							15.1
16.1		0	0.00%	0		0		0		0.00							16.1
17.1		0	0.00%	0		0		0		0.00							17.1
18.1		0	0.00%	0		0		0		0.00							18.1
19.1		0	0.00%	0		0		0		0.00							19.1
20.1		0	0.00%	0		0		0		0.00							20.1
TOTAL		134	100.00%	10		79		45		20.00	1.49	3	11.79	11	6.72	6	
										NOTE: If [shaded box] appears in the "Total # Needed From Topic" block - you do not have sufficient items in the topic indicated to design a fair test.							
										<div> <div>Compass Consultants, LLC</div> <div> <div>Test Difficulty Moderate</div> <div>Test Cut Score 64.00</div> </div> <div> <div>Set Desired Test Size 20</div> <div>Checksum 20</div> </div> <div> <div>Approximate Test Time in Minutes Based on Item Difficulty 17.43</div> </div> </div> <div>After all cut-score session data has been entered on section worksheets, set the desired test size in the block to the left. Based upon the number of available items, the quantity of Hard, Moderate and Easy from each section will populate automatically. Use these results to design the test in your test item database using established difficulty Metatags or sub-topic Approximate Difficulty Ratings . Note: Due to rounding errors in Excel, the unit/item difficulty totals may require you to round up or down manually to achieve desired test size. Set the actual number desired based upon the calculated results in the columns labeled "Actual" above. The Checksum to the left will alert you if the selected value does not match the desired test size.</div>							

Parry, J.R. (2017)

# Quick Poll

## How does your organization select which test-items appear on assessments?

- A. I/we use fixed form exams so everyone gets the same questions
- B. I/we use a fixed form but the questions and alternatives are shuffled
- C. I/we allow the testing software to select items at random each time
- D. I/we use stratified-randomization to ensure both content and difficulty are maintained



# Randomization vs. Stratified- Randomization

Item Selection Criteria is Important to Maintain Fairness



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

# Randomized Item Selection

- Experiments by Jim Parry:
  - Test-items selected at random from entire item database (n=30)

Question selections
20 random question(s) from topic 'FAIRNESS RESEARCH' including subtopics (Avoid previously delivered)

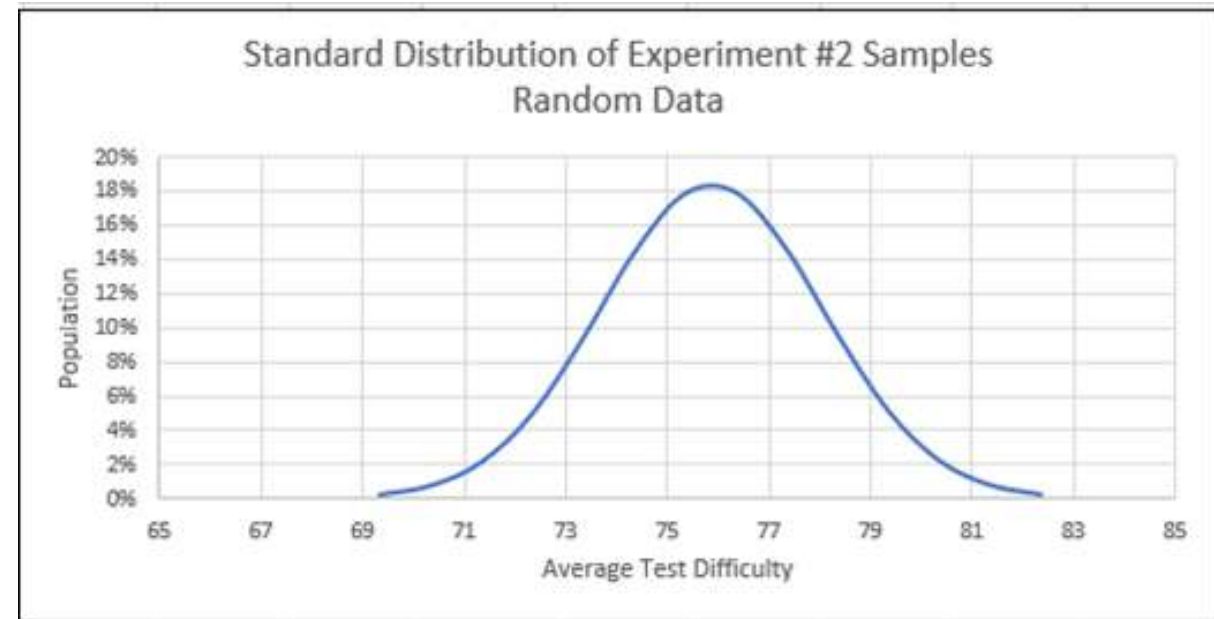
- Produced unpredictable results in topic coverage although average difficulty was acceptable
  - Number of hard, moderate, and easy items varied significantly

Experiment #2 - Random Selection of 20 items from all 3 topics. Real Client Data. Desired target difficulty is 76.13.																			
Attempt 1		Attempt 2		Attempt 3		Attempt 4		Attempt 5		Attempt 6		Attempt 7		Attempt 8		Attempt 9		Attempt 10	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E10	77.00	1.0 E1	80.00	1.0 E10	77.00	1.0 E1	80.00	1.0 E12	75.00	1.0 E10	77.00	1.0 E12	75.00	1.0 E12	75.00	1.0 E2	75.00	1.0 E1	80.00
1.0 E13	76.00	1.0 E10	77.00	1.0 E11	78.00	1.0 E11	78.00	1.0 E2	75.00	1.0 E14	79.00	1.0 E14	79.00	1.0 E13	76.00	1.0 E4	76.00	1.0 E10	77.00
1.0 E3	77.00	1.0 E11	78.00	1.0 E2	75.00	1.0 E7	78.00	1.0 E7	78.00	1.0 E5	94.00	1.0 E9	83.00	1.0 E2	75.00	1.0 E7	78.00	1.0 E14	79.00
1.0 E4	76.00	1.0 E5	94.00	1.0 E6	89.00	1.0 E8	91.00	1.0 E9	83.00	1.0 E6	89.00	1.0 M1	63.00	1.0 E3	77.00	1.0 M3	69.00	1.0 E4	76.00
1.0 E8	91.00	1.0 E9	83.00	2.0 E1	83.00	1.0 M1	63.00	2.0 E10	83.00	1.0 E9	83.00	2.0 E10	83.00	1.0 E7	78.00	2.0 E16	83.00	1.0 M1	63.00
2.0 E1	83.00	1.0 M1	63.00	2.0 E13	79.00	1.0 M4	71.00	2.0 E16	83.00	1.0 M3	69.00	2.0 E11	82.50	2.0 E1	83.00	2.0 E2	92.00	1.0 M3	69.00
2.0 E11	82.50	2.0 E14	90.00	2.0 E2	92.00	2.0 E14	90.00	2.0 E20	80.00	2.0 E1	83.00	2.0 E13	79.00	2.0 E10	83.00	2.0 E3	76.00	2.0 E1	83.00
2.0 E12	90.00	2.0 E15	82.00	2.0 E20	80.00	2.0 E16	83.00	2.0 E4	75.00	2.0 E10	83.00	2.0 E14	90.00	2.0 E13	79.00	2.0 E4	75.00	2.0 E17	79.00
2.0 E15	82.00	2.0 E16	83.00	2.0 E3	76.00	2.0 E19	86.00	2.0 E5	74.00	2.0 E12	90.00	2.0 E17	79.00	2.0 E3	76.00	2.0 E8	81.00	2.0 E18	81.00
2.0 E2	92.00	2.0 E17	79.00	2.0 E4	75.00	2.0 E2	92.00	2.0 E8	81.00	2.0 E17	79.00	2.0 E21	78.00	2.0 E5	74.00	2.0 E9	89.00	2.0 E5	74.00
2.0 E4	75.00	2.0 E21	78.00	2.0 E7	75.00	2.0 E3	76.00	2.0 M10	56.25	2.0 E20	80.00	2.0 E5	74.00	2.0 E8	81.00	2.0 M1	63.00	2.0 E6	80.00
2.0 E5	74.00	2.0 E4	75.00	2.0 E9	80.00	2.0 E4	75.00	2.0 M1	63.00	2.0 E5	74.00	2.0 E6	80.00	2.0 H1	46.25	2.0 M3	67.00	2.0 E8	81.00
2.0 E7	75.00	2.0 E6	80.00	2.0 H1	46.25	2.0 E5	74.00	2.0 M3	67.00	2.0 E8	81.00	2.0 H1	46.25	2.0 M8	52.50	2.0 M5	68.00	2.0 M1	63.00
2.0 E9	89.00	2.0 H1	46.25	2.0 M3	67.00	2.0 E7	75.00	2.0 M6	53.75	2.0 M3	67.00	2.0 M3	67.00	3.0 E1	90.00	3.0 E10	85.00	2.0 M3	67.00
2.0 M9	70.00	2.0 M4	53.00	2.0 M8	52.50	2.0 M4	53.00	2.0 M7	66.25	2.0 M6	53.75	2.0 M4	53.00	3.0 E10	85.00	3.0 E2	87.00	2.0 M4	53.00
3.0 E17	72.00	2.0 M8	52.50	3.0 E10	85.00	2.0 M9	70.00	3.0 E12	85.00	3.0 E15	73.00	2.0 M6	53.75	3.0 E14	83.00	3.0 E3	84.00	2.0 M8	52.50
3.0 E2	87.00	3.0 E12	85.00	3.0 E1	90.00	3.0 E13	72.00	3.0 E14	83.00	3.0 E15	90.00	2.0 M8	52.50	3.0 E16	75.00	3.0 E4	74.00	3.0 E13	72.00
3.0 E9	89.00	3.0 E14	83.00	3.0 E12	85.00	3.0 E14	83.00	3.0 E4	74.00	3.0 E6	73.00	3.0 E12	85.00	3.0 E17	72.00	3.0 E7	83.00	3.0 E15	73.00
3.0 M1	57.50	3.0 E15	73.00	3.0 E7	83.00	3.0 E16	75.00	3.0 E5	79.00	3.0 E8	72.00	3.0 E15	73.00	3.0 E5	79.00	3.0 E8	72.00	3.0 E2	87.00
3.0 M3	57.50	3.0 M3	57.50	3.0 M2	61.00	3.0 E6	73.00	3.0 M3	57.50	3.0 M2	61.00	3.0 E7	83.00	3.0 M1	57.50	3.0 M3	57.50	3.0 E3	84.00
Difficulty	78.63	Difficulty	74.61	Difficulty	76.44	Difficulty	76.90	Difficulty	73.59	Difficulty	77.54	Difficulty	72.95	Difficulty	74.86	Difficulty	76.73	Difficulty	73.68
Easy	17	Easy	15	Easy	16	Easy	16	Easy	14	Easy	16	Easy	14	Easy	17	Easy	15	Easy	14
Moderate	3	Moderate	4	Moderate	3	Moderate	4	Moderate	6	Moderate	4	Moderate	5	Moderate	2	Moderate	5	Moderate	6
Hard	0	Hard	1	Hard	1	Hard	0	Hard	0	Hard	0	Hard	1	Hard	1	Hard	0	Hard	0
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	6	Topic 1	4	Topic 1	6	Topic 1	4	Topic 1	6	Topic 1	4	Topic 1	5	Topic 1	4	Topic 1	6
Topic 2	10	Topic 2	10	Topic 2	11	Topic 2	10	Topic 2	11	Topic 2	9	Topic 2	13	Topic 2	8	Topic 2	9	Topic 2	10
Topic 3	5	Topic 3	4	Topic 3	5	Topic 3	4	Topic 3	5	Topic 3	5	Topic 3	3	Topic 3	7	Topic 3	7	Topic 3	4



# Distribution of Random Selection Results

Sample Difficulty Statistics	
Target Cut Score	76.13
Mean difficulty	75.87
Median	75.34
Minimum	73.00
Maximum	79.95
Variance Target vs. Mean	0.03
Standard Deviation all Averages	2.17
95% Confidence Score	0.777910235
Kurtosis	-0.52653425
Skewness	0.613319589



Most often, kurtosis is measured against the normal distribution. If the kurtosis is close to 0, then a normal distribution is often assumed. A low kurtosis indicates a lack of significant outliers. A high kurtosis indicates significant outliers. (-2,2 is acceptable)

Skewness is usually described as a measure of a dataset's symmetry – or lack of symmetry. A perfectly symmetrical data set will have a skewness of 0 which is referred to as “normal” distribution. Negative skew indicates data is skewed left and positive indicates data is skewed right when referring to the “tail”. (-1,1 is acceptable)

# Preparing Item Database for Stratified-Random Selection

## Using Sub-Topics

- Repository Name
  - Objective 1.0
    - *Topic 1.1*
      - Sub-Topic 1.1.1
        - 1.1.1 HARD
          - Test-Item 1.1.1/1
          - Test-Item 1.1.1/2
        - 1.1.1 MODERATE
          - Test-Item 1.1.1/3
          - Test-Item 1.1.1/4
        - 1.1.1 EASY
          - Test-Item 1.1.1/5
          - Test-Item 1.1.1/6

## Using Metatags

- Repository Name
  - Objective 1.0
    - *Topic 1.1*
      - Sub-Topic 1.1.1
        - Test-Item 1.1.1/1
          - <tag> Expert-Easy
          - <tag> Journey-Mod
          - <tag> Appren-Hard
        - Test-Item 1.1.1/2
          - <tag> Expert-Mod
          - <tag> Journey-Hard
          - <tag> Appren-Hard

# Stratified-Random Item Selection Criteria

- Test-items selected by both topic and difficulty (n=30)

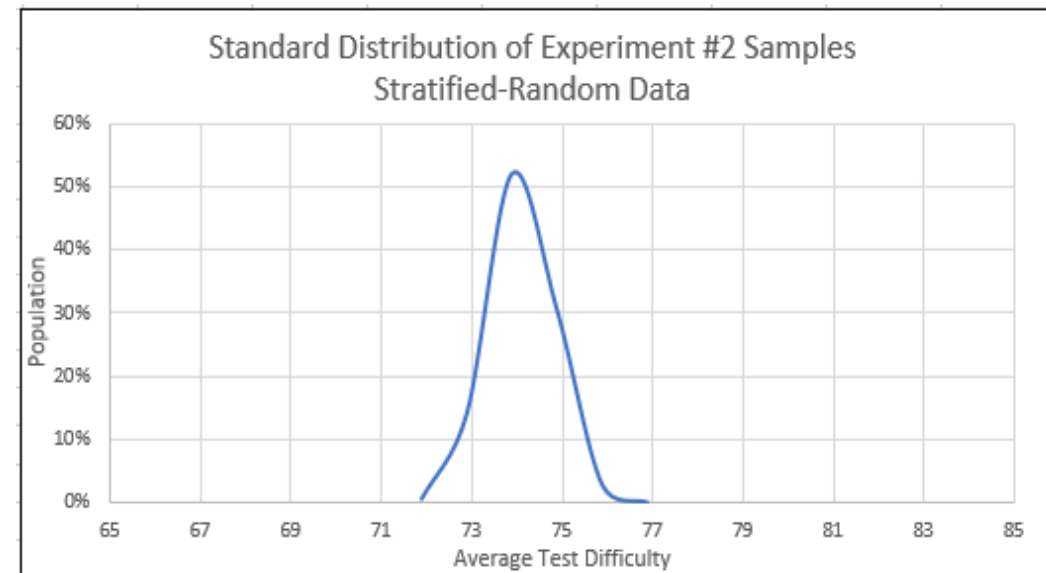
Question selections
4 random question(s) from topic 'FAIRNESS RESEARCH 2/1.0 TOPIC 1/1.0 EASY' excluding subtopics (Avoid previously delivered)
1 random question(s) from topic 'FAIRNESS RESEARCH 2/1.0 TOPIC 1/1.0 MODERATE' excluding subtopics (Avoid previously delivered)
6 random question(s) from topic 'FAIRNESS RESEARCH 2/2.0 TOPIC 2/2.0 EASY' excluding subtopics (Avoid previously delivered)
3 random question(s) from topic 'FAIRNESS RESEARCH 2/2.0 TOPIC 2/2.0 MODERATE' excluding subtopics (Avoid previously delivered)
1 random question(s) from topic 'FAIRNESS RESEARCH 2/2.0 TOPIC 2/2.0 HARD' excluding subtopics (Avoid previously delivered)
4 random question(s) from topic 'FAIRNESS RESEARCH 2/3.0 TOPIC 3/3.0 EASY' excluding subtopics (Avoid previously delivered)
1 random question(s) from topic 'FAIRNESS RESEARCH 2/3.0 TOPIC 3/3.0 MODERATE' excluding subtopics (Avoid previously delivered)

- Produced same topic coverage and acceptable difficulty each iteration
  - Number of hard, moderate, and easy items from each topic remained constant

Experiment #2 - Directed Random Selection of 20 items from all 3 topics. Real Client Data. Desired target difficulty is 76.13.																			
Attempt 1		Attempt 2		Attempt 3		Attempt 4		Attempt 5		Attempt 6		Attempt 7		Attempt 8		Attempt 9		Attempt 10	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E13	76.00	1.0 E13	76.00	1.0 E2	75.00	1.0 E10	77.00	1.0 E8	91.00	1.0 E1	80.00	1.0 E9	83.00	1.0 E12	75.00	1.0 E7	78.00	1.0 E13	76.00
1.0 E1	80.00	1.0 E9	83.00	1.0 E12	75.00	1.0 E1	80.00	1.0 E5	94.00	1.0 E11	78.00	1.0 E12	75.00	1.0 E1	80.00	1.0 E4	76.00	1.0 E1	80.00
1.0 E8	91.00	1.0 E14	79.00	1.0 E8	91.00	1.0 E11	78.00	1.0 E14	79.00	1.0 E7	78.00	1.0 E7	78.00	1.0 E7	78.00	1.0 E5	94.00	1.0 E8	91.00
1.0 E4	76.00	1.0 E1	80.00	1.0 E3	77.00	1.0 E3	77.00	1.0 E4	76.00	1.0 E9	83.00	1.0 E11	78.00	1.0 E8	91.00	1.0 E13	76.00	1.0 E3	77.00
1.0 M1	63.00	1.0 M1	63.00	1.0 M2	67.00	1.0 M2	67.00	1.0 M2	67.00	1.0 M3	69.00	1.0 M4	71.00	1.0 M3	69.00	1.0 M4	71.00	1.0 M4	71.00
2.0 E15	82.00	2.0 E22	76.00	2.0 E3	76.00	2.0 E9	89.00	2.0 E11	82.50	2.0 E15	82.00	2.0 E20	80.00	2.0 E7	75.00	2.0 E15	82.00	2.0 E9	89.00
2.0 E14	90.00	2.0 E17	79.00	2.0 E7	75.00	2.0 E13	79.00	2.0 E16	83.00	2.0 E7	75.00	2.0 E6	80.00	2.0 E18	81.00	2.0 E13	79.00	2.0 E4	75.00
2.0 E8	81.00	2.0 E13	79.00	2.0 E21	78.00	2.0 E18	81.00	2.0 E10	83.00	2.0 E9	89.00	2.0 E2	92.00	2.0 E14	90.00	2.0 E17	79.00	2.0 E1	83.00
2.0 E2	92.00	2.0 E9	89.00	2.0 E17	79.00	2.0 E12	90.00	2.0 E17	79.00	2.0 E14	90.00	2.0 E9	89.00	2.0 E20	80.00	2.0 E5	74.00	2.0 E5	74.00
2.0 E18	81.00	2.0 E7	75.00	2.0 E10	83.00	2.0 E8	81.00	2.0 E13	79.00	2.0 E19	86.00	2.0 E15	82.00	2.0 E15	82.00	2.0 E7	75.00	2.0 E15	82.00
2.0 E17	79.00	2.0 E15	82.00	2.0 E14	90.00	2.0 E5	74.00	2.0 E22	76.00	2.0 E4	75.00	2.0 E13	79.00	2.0 E4	75.00	2.0 E4	75.00	2.0 E2	92.00
2.0 M5	68.00	2.0 M9	70.00	2.0 M4	53.00	2.0 M7	66.25	2.0 M7	66.25	2.0 M4	53.00	2.0 M9	70.00	2.0 M9	70.00	2.0 M9	70.00	2.0 M10	56.25
2.0 M10	56.25	2.0 M8	52.50	2.0 M2	48.75	2.0 M2	48.75	2.0 M10	56.25	2.0 M9	70.00	2.0 M7	66.25	2.0 M10	56.25	2.0 M10	56.25	2.0 M9	70.00
2.0 M1	63.00	2.0 M6	53.75	2.0 M6	53.75	2.0 M3	67.00	2.0 M9	70.00	2.0 M10	56.25	2.0 M2	48.75	2.0 M6	53.75	2.0 M1	63.00	2.0 M6	53.75
2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25
3.0 E9	89.00	3.0 E1	90.00	3.0 E10	85.00	3.0 E6	73.00	3.0 E10	85.00	3.0 E13	72.00	3.0 E10	85.00	3.0 E11	84.00	3.0 E17	72.00	3.0 E16	75.00
3.0 E17	72.00	3.0 E10	85.00	3.0 E12	85.00	3.0 E4	74.00	3.0 E4	74.00	3.0 E14	83.00	3.0 E7	83.00	3.0 E5	79.00	3.0 E3	84.00	3.0 E7	83.00
3.0 E11	84.00	3.0 E14	83.00	3.0 E11	84.00	3.0 E17	72.00	3.0 E11	84.00	3.0 E9	89.00	3.0 E17	72.00	3.0 E13	72.00	3.0 E12	85.00	3.0 E5	79.00
3.0 E8	72.00	3.0 E2	87.00	3.0 E1	90.00	3.0 E5	79.00	3.0 E8	72.00	3.0 E3	84.00	3.0 E3	84.00	3.0 E6	73.00	3.0 E10	85.00	3.0 E3	84.00
3.0 M3	57.50	3.0 M1	57.50	3.0 M2	61.00	3.0 M1	57.50	3.0 M2	61.00	3.0 M3	57.50	3.0 M1	57.50	3.0 M2	61.00	3.0 M3	57.50	3.0 M1	57.50
Difficulty	74.95	Difficulty	74.30	Difficulty	73.64	Difficulty	72.84	Difficulty	75.21	Difficulty	74.80	Difficulty	74.99	Difficulty	73.56	Difficulty	73.90	Difficulty	74.74
Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14
Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5
Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5
Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10
Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5

# Distribution of Stratified-Random Selection Results

Sample Difficulty Statistics	
Target Cut Score	76.13
Mean difficulty	74.11
Median	73.98
Minimum	73.00
Maximum	75.76
Variance Target vs. Mean	2.04
Standard Deviation all Averages	0.74
95% Confidence Score	0.263545877
Kurtosis	0.117166773
Skewness	0.579229905




Most often, kurtosis is measured against the normal distribution. If the kurtosis is close to 0, then a normal distribution is often assumed. A low kurtosis indicates a lack of significant outliers. A high kurtosis indicates significant outliers. (-2,2 is acceptable)

Skewness is usually described as a measure of a dataset's symmetry – or lack of symmetry. A perfectly symmetrical data set will have a skewness of 0 which is referred to as “normal” distribution. Negative skew indicates data is skewed left and positive indicates data is skewed right when referring to the “tail”. (-1,1 is acceptable)



# Determine Stratification Criteria

Final Pseudo-Randomized Test Design Blueprint for:											TEST NAME						mm/dd/yyyy
Topic	Topic Cut Score & Difficulty	Items in Topic	% of Total Items	Available Hard	% From Topic	Available Mod	% From Topic	Available Easy	% From Topic	Total # Needed From Topic	Use Hard (Calculated)	Use Hard (Actual)	Use Mod (Calculated)	Use Mod (Actual)	Use Easy (Calculated)	Use Easy (Actual)	Topic
Topic 1	78	18	25.35%	0	0%	4	22%	14	78%	5.07	0.00	0	1.13	1	3.94	4	Topic 1
Topic 2	74	33	46.48%	1	3%	10	30%	22	67%	9.30	0.28	1	2.82	3	6.20	6	Topic 2
Topic 3	77	20	28.17%	0	0%	3	15%	17	85%	5.63	0.00	0	0.85	1	4.79	4	Topic 3
4.1		0	0.00%	0		0		0		0.00							4.1
5.1		0	0.00%	0		0		0		0.00							5.1
6.1		0	0.00%	0		0		0		0.00							6.1
7.1		0	0.00%	0		0		0		0.00							7.1
8.1		0	0.00%	0		0		0		0.00							8.1
9.1		0	0.00%	0		0		0		0.00							9.1
10.1		0	0.00%	0		0		0		0.00							10.1
11.1		0	0.00%	0		0		0		0.00							11.1
12.1		0	0.00%	0		0		0		0.00							12.1
13.1		0	0.00%	0		0		0		0.00							13.1
14.1		0	0.00%	0		0		0		0.00							14.1
15.1		0	0.00%	0		0		0		0.00							15.1
16.1		0	0.00%	0		0		0		0.00							16.1
17.1		0	0.00%	0		0		0		0.00							17.1
18.1		0	0.00%	0		0		0		0.00							18.1
19.1		0	0.00%	0		0		0		0.00							19.1
20.1		0	0.00%	0		0		0		0.00							20.1
TOTAL		71	100.00%	1		17		53		20.00	0.28	1	4.79	5	14.93	14	
										NOTE: If <div></div> appears in the "Total # Needed From Section" block - you do not have sufficient items in the section indicated to design a fair test.							
		<div>Test Difficulty Range</div>		<div>Test Cut Score</div>		<div>Set Desired Test Size</div>		<div>After all cut-score session data has been entered on section worksheets, set the desired test size in the block to the left. Based upon the number of available items, the quantity of Hard, Moderate and Easy from each section will populate automatically. Use these results to design the test in your test item database using established difficulty Metatags or sub-topic Approximate Difficulty Ratings . Note: Due to rounding errors in Excel, the unit/item difficulty totals may require you to round up or down manually to achieve desired test size. Set the actual number desired based upon the calculated results in the columns labeled "Actual" above. The Checksum to the left will alert you if the selected value does not match the desired test size.</div>									
		76.00		20													
				Checksum													
				20													

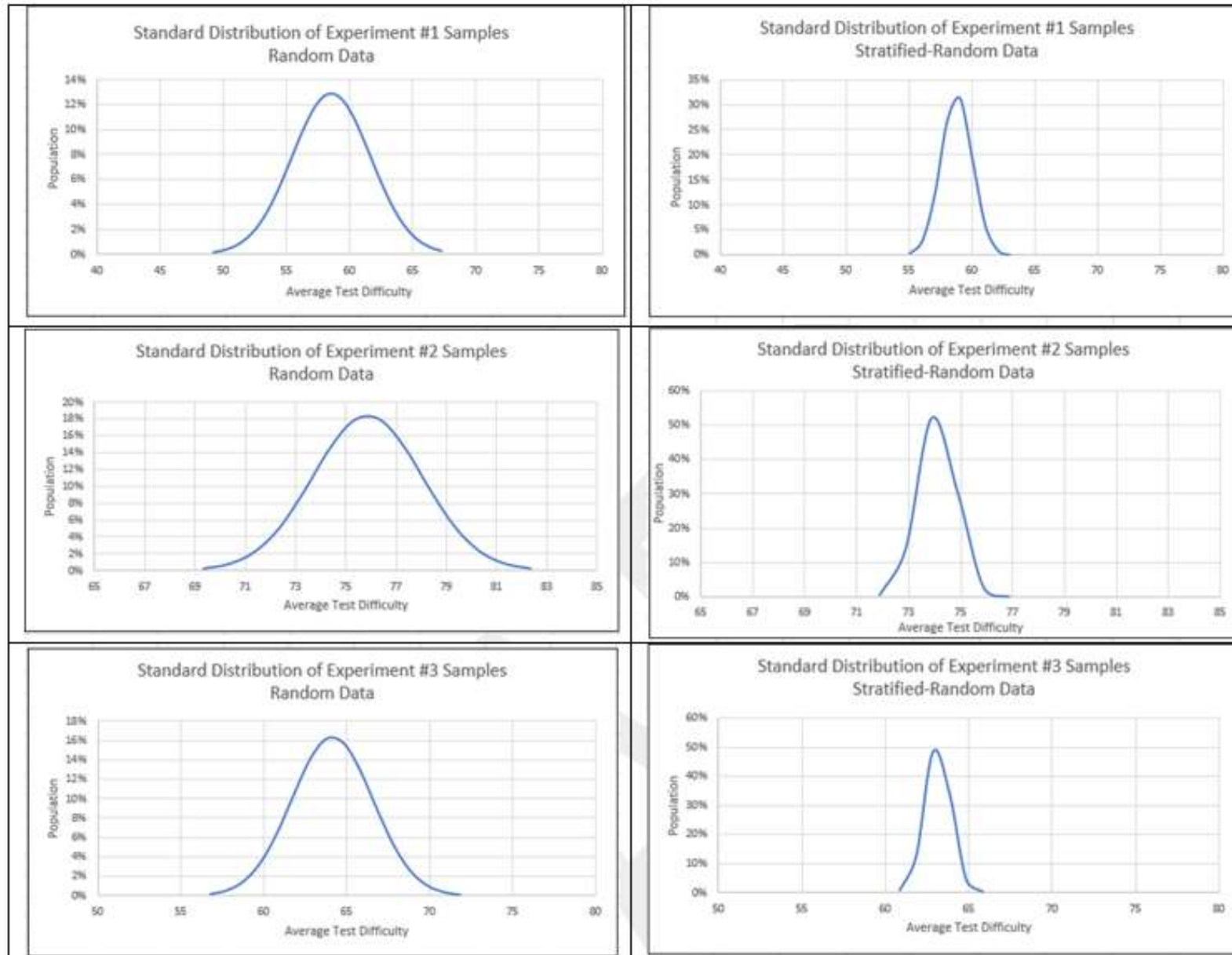
Total test-items available by topic at each difficulty level.

Topic	Topic Cut Score & Difficulty	Items in Topic	% of Total Items	Available Hard	% From Topic	Available Mod	% From Topic	Available Easy	% From Topic
Topic 1	78	18	25.35%	0	0%	4	22%	14	78%
Topic 2	74	33	46.48%	1	3%	10	30%	22	67%
Topic 3	77	20	28.17%	0	0%	3	15%	17	85%

Total # Needed From Topic	Use Hard (Calculated)	Use Hard (Actual)	Use Mod (Calculated)	Use Mod (Actual)	Use Easy (Calculated)	Use Easy (Actual)	Topic
5.07	0.00	0	1.13	1	3.94	4	Topic 1
9.30	0.28	1	2.82	3	6.20	6	Topic 2
5.63	0.00	0	0.85	1	4.79	4	Topic 3

Recommended test design based on number of items available at each difficulty level to maintain difficulty and topic coverage.

R  
a  
n  
d  
o  
m



S  
t  
r  
a  
t  
i  
f  
i  
e  
d  
R  
a  
n  
d  
o  
m

# Conclusions

- Random selection produces unpredictable results
  - Content coverage is erratic
  - Number of items at each difficulty level in each topic is erratic
  - Average difficulty remains within acceptable range from the desired (calculated) cut score but SD of 30 attempts was high
- Stratified random selection produces predictable results
  - Content coverage is always equal
  - Number of items at each difficulty level in each topic is constant
  - Average difficulty remains within acceptable range from the desired (calculated) cut score and SD of 30 attempts was significantly lower than random selection

# Recommendations to Maintain Fairness

- Test-items must be constructed using universally recognized standards
- Cut scores should be established using a recognized test-centered method or, if appropriate, a test-taker centered method, because arbitrary methods are not defensible
- Each item in a test-item database should be evaluated by a panel of expert raters/judges and a difficulty score or rating established based upon the agreed upon MAC level of the target test-taker
- Tests should not be generated in a pure random fashion from a test-item database without regard to content and item difficulty because content coverage and item difficulty among tests will be erratic
- Regular monitoring of the statistical Item Response Theory (IRT) and/or Classical Test Theory (CTT) performance of tests and test-items is necessary to ensure validity and reliability





# Defensibility!

# References

- Beauchamp, C. (2017). *Setting a cut score for a performance-based assessment: the Ebel method*. Yardstick Testing and Training Experts. Retrieved June 9, 2017 from [http://getyardstick.com/new2017/wp-content/uploads/2016-April\\_Backgrounder\\_Setting-a-cut-score-for-a-PBA-with-the-Ebel-Method.pdf](http://getyardstick.com/new2017/wp-content/uploads/2016-April_Backgrounder_Setting-a-cut-score-for-a-PBA-with-the-Ebel-Method.pdf)
- Cizek, G. J., (2012). *Setting performance standards – foundations, methods, and innovations*. (2<sup>nd</sup>)., Routledge Taylor & Francis Group, New York, NY
- Coscarelli, W., Barrett, A., Kleeman, J., & Shrock, S. (2005). The problem of saltatory cut-score: some issues and recommendations for. Proceedings of the 9th CAA Conference. Loughborough: Loughborough University. Retrieved from <https://hdl.handle.net/2134/1984>.
- Livingston, S.A. & Kastrinos, W., (1982). *A study of the reliability of Nedelsky's method for choosing a passing score*. Educational Testing Service.: Princeton, NJ. Retrieved June 9, 2017 from <http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1982.tb01292.x/epdf>
- Livingston, S.A & Zieky, M.J., (1982). *Passing scores – A manual for setting standards on educational and occupational tests*. Educational Testing Service.: Princeton, NJ
- Parry, J.R., (2017, March). *Setting fair, defensible cut (passing) scores*. Presentation at 2017 Questionmark Users Conference, Santa Fe, NM
- Parry, J.R. (2020, April 30). *Ensuring fairness in difficulty and content among parallel assessments generated from a test-item database*. Online Submission. Retrieved July 18, 2022, from <https://eric.ed.gov/?q=fairness%2Bin%2Bperformance%2Bassessment&id=ED605523>
- Pitoniak, Mary J., "Investigation of two standard setting methods for a licensure examination." (2002). Masters Theses 1911 – February 2014. 2386. Retrieved December 20, 2018 from <https://scholarworks.umass.edu/theses/2386>
- United States Coast Guard. (2015). *Training system standard operation procedure 10 – Testing (SOP-10)*. United States Coast Guard Forces Command, Washington, DC
- Yousef, Mohammed & Alshawwa, Lana & Tekian, Ara & Park, Yoon Soo. (2017). *Challenging the arbitrary cutoff score of 60%: Standard setting evidence from preclinical Operative Dentistry course*. Medical Teacher. 39. 10.1080/0142159X.2016.1254752.

# Questions?





# Upcoming Webinars

## Introduction to Questionmark's Assessment Platform

◆ October 6, 2022 - 12:00 pm to 1:00 pm (EDT)

Learn the basics of authoring, delivering and reporting on surveys, quizzes, tests and exams using Questionmark's assessment platform.

[Click to Register](#)

## Workplace Exams 101: How to Prevent Cheating

◆ October 4, 2022 - 11:00 am to 12:00 pm (EDT)

Tests and exams given in the workplace serve a purpose – they are used to make important decisions. When employees cheat, they devalue that purpose and the integrity of your business suffers.

[Click to Register](#)

## Tuesday Training with the Techs: Advancing Your Knowledge of Advanced Editor

◆ November 15, 2022 - 11:00 am to 11:45 am (EDT)

Advanced Editor gives you broader control over your Questionmark Assessments by unlocking certain features that may not be accessible or are not considered applicable to the selected question type in the Standard Editor.

[Click to Register](#)



# Additional Reading



## READ NOW:

- [Defensibility and Legal Certainty for Tests and Exams – A Best Practice Guide](#)
- [Ensuring Fairness in Difficulty and Content Among Parallel Assessments Generated From a Test-Item Database](#)



— a Learnosity company —



# Thank you for your attention!

*Reach out to Questionmark at [sales@questionmark.com](mailto:sales@questionmark.com)  
or request a demo at <https://www.questionmark.com/request-demo>*

*If you would like to reach out to Jim Parry – [james.parry@gocompassconsultants.com](mailto:james.parry@gocompassconsultants.com)  
[www.gocompassconsultants.com](http://www.gocompassconsultants.com)*