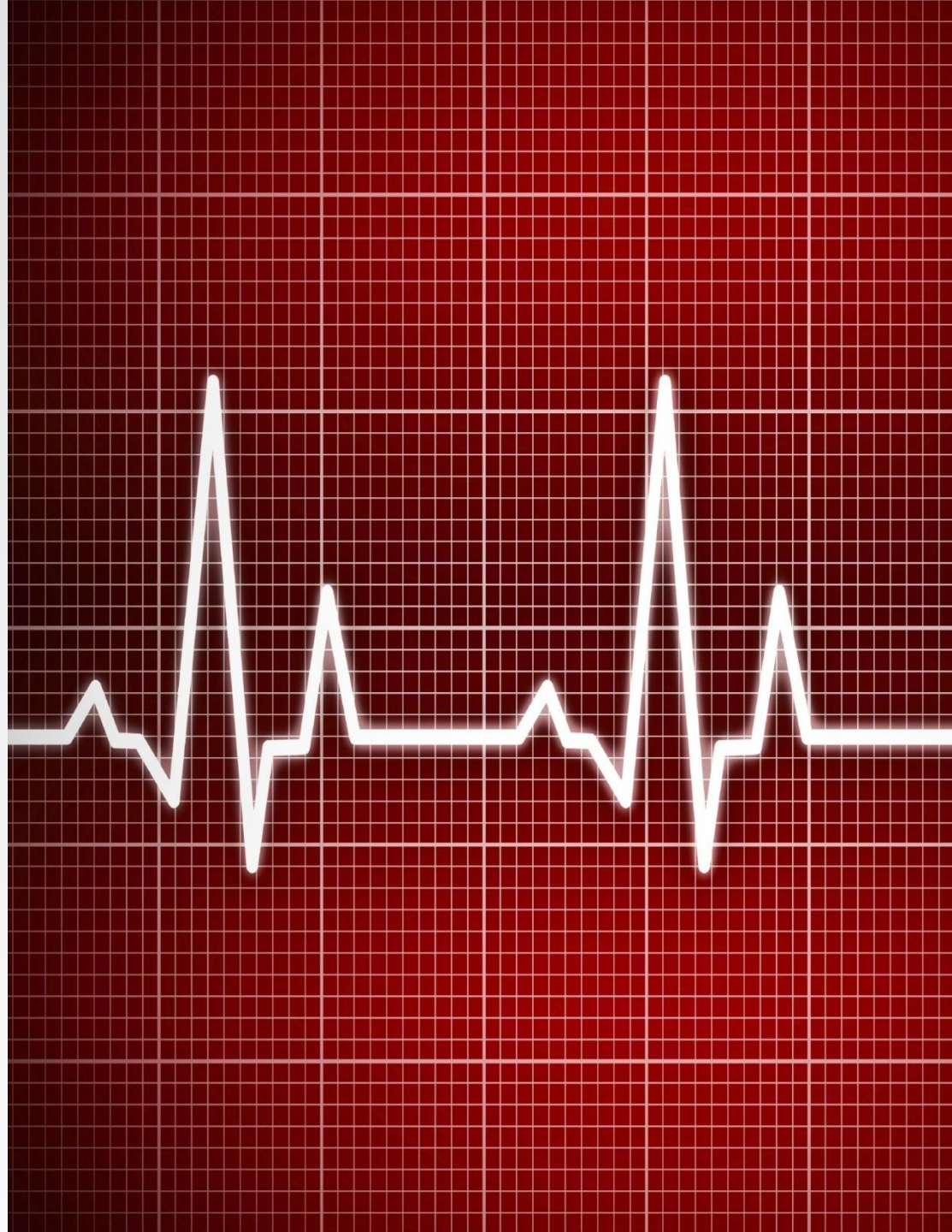


# Psychometrics 101: An Overview of Fundamental Psychometric Principles



ASSESSMENT, EDUCATION, AND  
RESEARCH EXPERTS



# Presenters



**Manny  
Straehle**



**Liberty  
Munson**

# About Your Presenter –Dr. Manny Straehle



- Inventor of the Swearing Chicken
- Ph.D. in Educational Psychology
- ABD in Counseling Psychology
- IT Certifications: GISF, Data Management Support
- Therapy Certifications: Social Therapy
- Testing Organizations Worked at:
  - Psychometrics: NBME, Prometric, USGBC
- Organizations founded:
  - International Credential Associates
  - Assessment, Education, and Research Associates (AERE)
- University Teaching Experiences: Temple, Penn State, Saint Joseph's University, Johns Hopkins, USC, and George Washington University
- Number of Organizations Consulted: 100+
- Social Responsibility: TEDx, E-ATP, ATP, ACA, ALA, Special Olympics, Spark, ESI, Habitat for America
- Number of Presentations: 70+
- Interests: Pizza Making, Presidential Libraries, Healthcare Communications, Pro Bono, Family, Friends, and Good Laughs



# About Your Presenter – Dr. Liberty Munson



- Principal Psychometrician for Microsoft's Learning & Readiness organization
- Responsible for ensuring that the skills assessments in Microsoft Technical Certification and Professional Programs are valid and reliable
- Prior to Microsoft, worked at Boeing in their Employee Selection Group, assisted with the development of their internal certification exams, and acted as a co-project manager of Boeing's Employee Survey
- BS in Psychology from Iowa State University and MA and PhD in Industrial/Organizational Psychology with minors in Quantitative Psychology and Human Resource Management from the University of Illinois at Urbana-Champaign

# Ask a Psychometrician?

What have been your experiences with psychometricians?

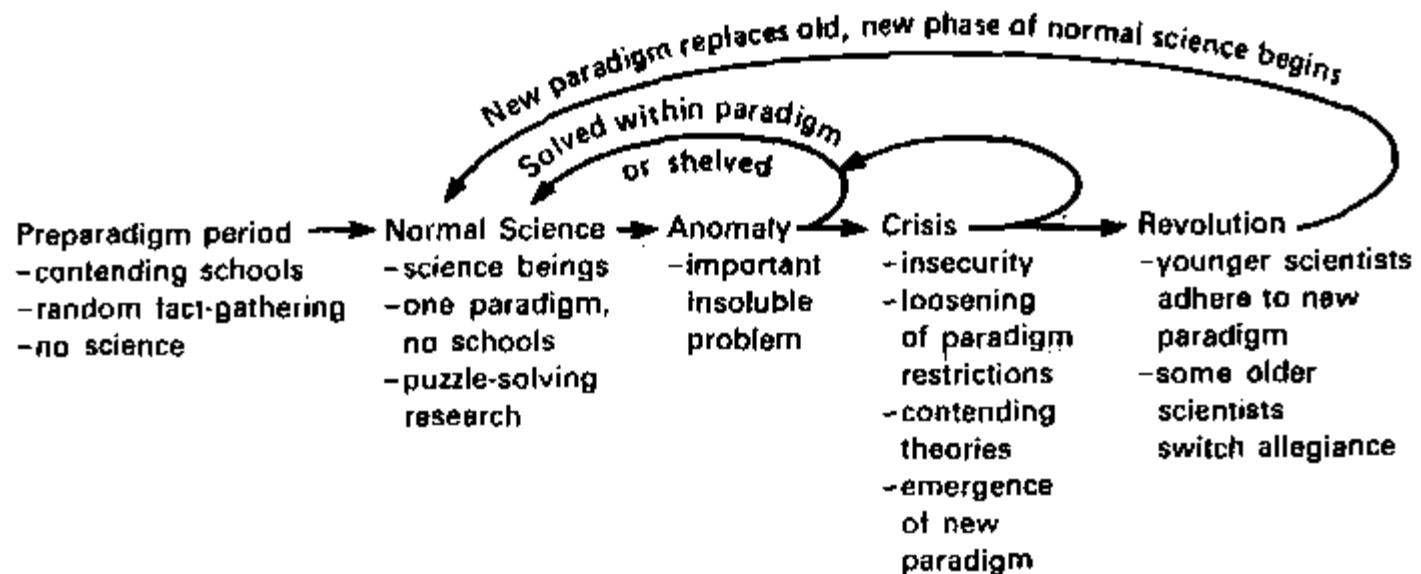
What questions have not been answered to your satisfaction?

Are psychometricians contradictory from one consultant/vendor to another? Tell us how?

What don't you understand about psychometrics that you wish you did?

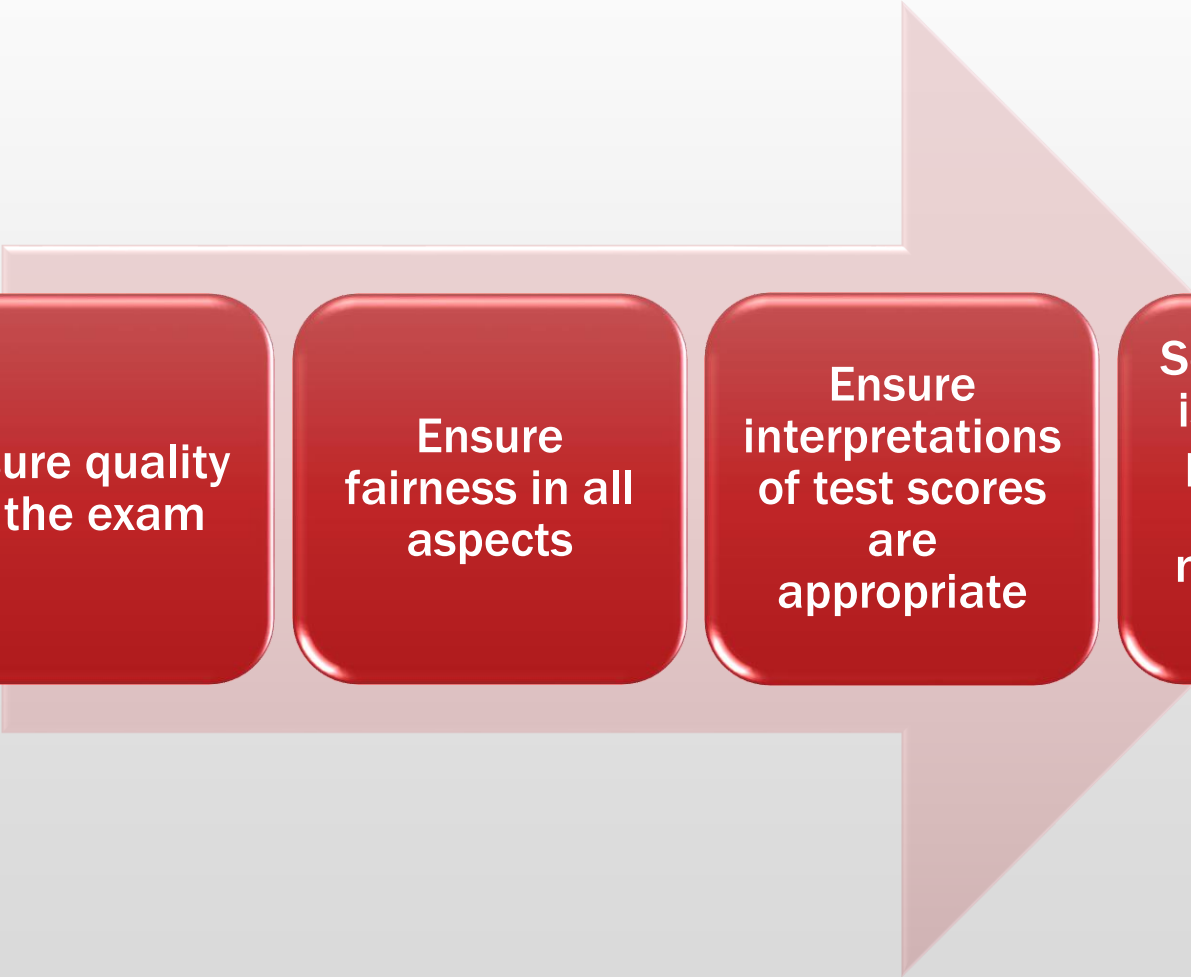
# Disclaimer

The revolutionary character of paradigm shifts, and the cyclical nature of science (a schematization of Kuhn, 1970).



- Guidelines not rules
- Intended for managers and executives of credentialing programs
- Innovations may be accepted by industry peers

# Why is Psychometrics Important?



Ensure quality  
of the exam

Ensure  
fairness in all  
aspects

Ensure  
interpretations  
of test scores  
are  
appropriate

Someone who  
is certified is  
proficient at  
skills  
measured by  
exam

# Some Basic Terminology

**What is an examination/  
assessment/test?**

- A tool that allows us to obtain a sample of an individual's behavior in one or several circumscribed domains

**What is a domain?**

- Defined population of what could be measured by assessment process

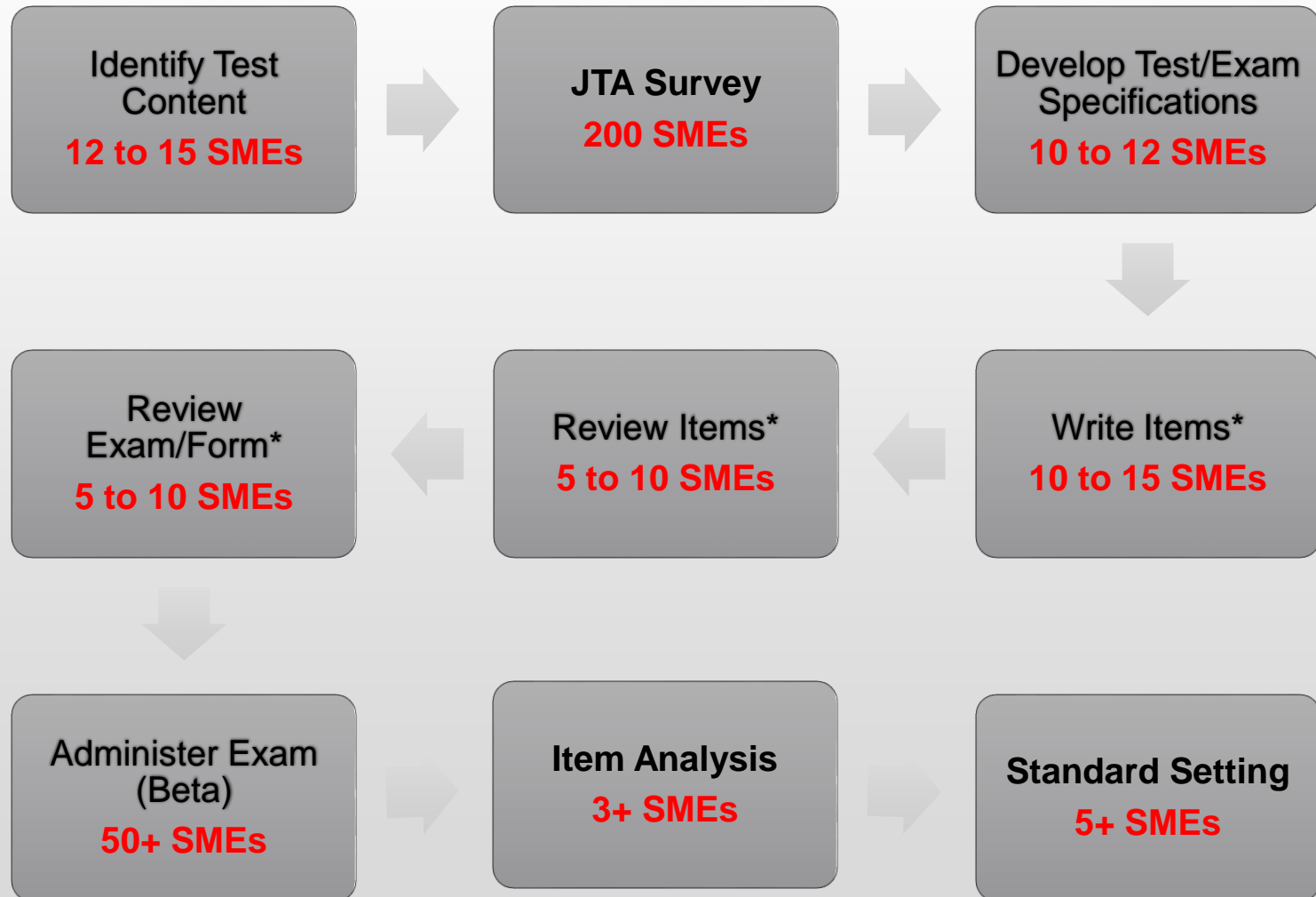
**What are forms?**

- Set of items seen by a set of candidates
- Fixed or dynamic



# Test Development

# Test Development Lifecycle



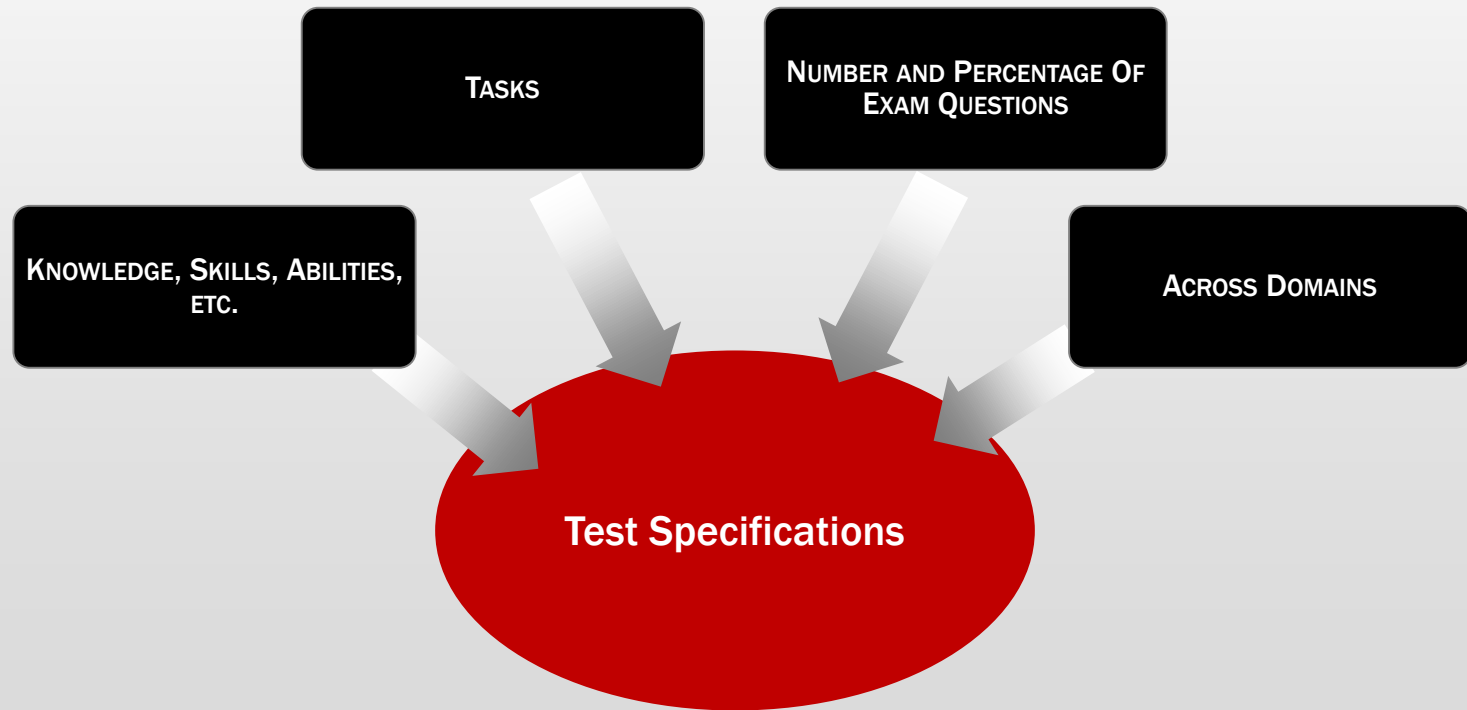
# **Job/Practice Analysis**

# Job Analysis

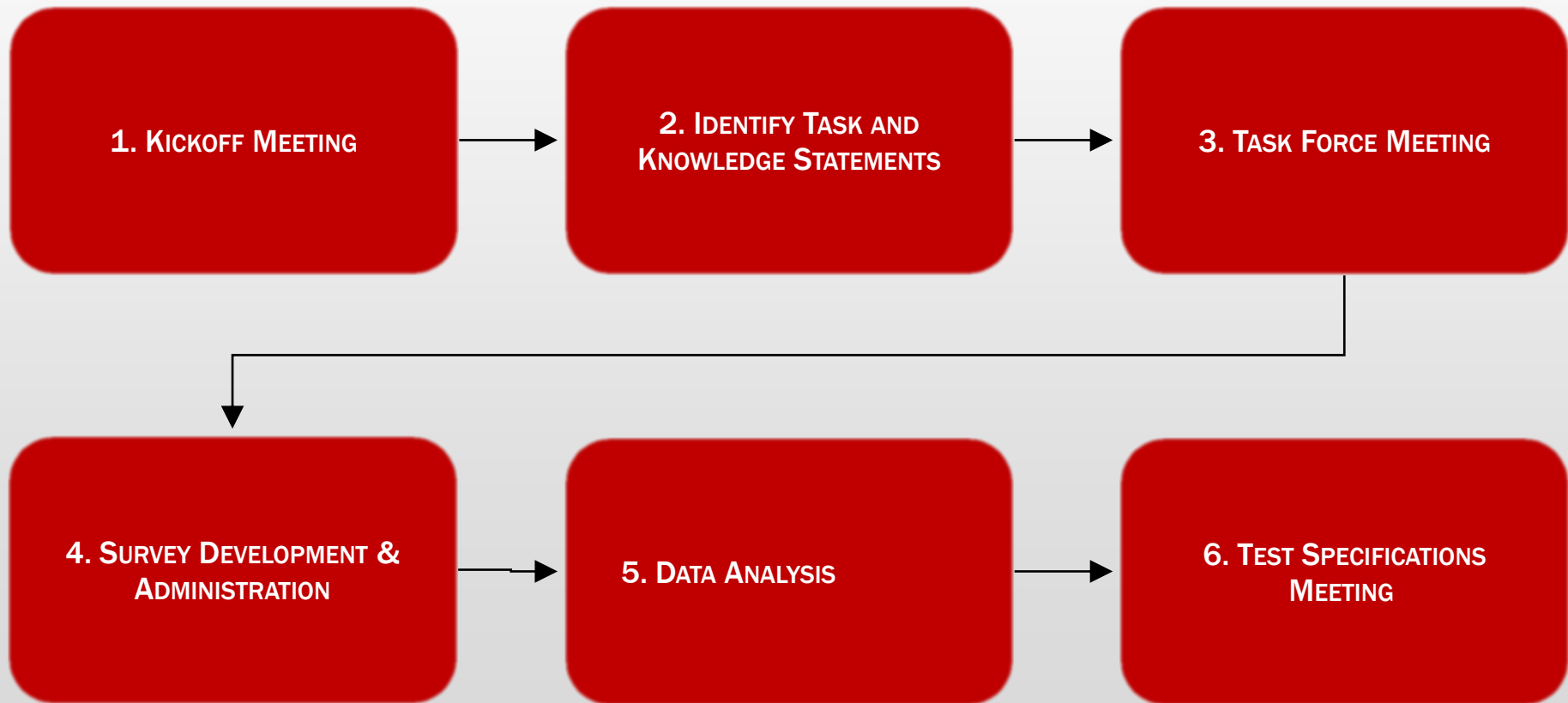
“...the systematic process of discovery of the nature of a job by dividing it into smaller units, where the process results in one or more written products with the goal of describing what is done in the job or what capabilities are needed to effectively perform the job” (p. 8).

-Michael Brannick (2006)

# Common Purpose for a Job Analysis



# Job Analysis Lifecycle





# Job Analysis



**Research method** using inputs, SMEs, focus groups, interviews, and surveys to **identify**:

- Tasks
- Knowledge, skills, abilities, etc.

**Results in:**

- Exam content domain that is the foundation of the exam
- Drives test specification/exam blueprint that will be used to develop items and build examination forms

# Test Specifications/Exam Blueprint



## EXAM OBJECTIVES (DOMAINS)

The table below lists the domains measured by this examination and the extent to which they are represented:

DOMAIN	PERCENTAGE OF EXAMINATION
1.0 Network Architecture	22%
2.0 Network Operations	20%
3.0 Network Security	18%
4.0 Troubleshooting	24%
5.0 Industrial Standards, Practices and Network Theory	16%
<b>Total</b>	<b>100%</b>

# Item Writing

# Use Evidence-Based Item Writing Guidelines

APPLIED MEASUREMENT IN EDUCATION, 2(1), 37-50  
Copyright © 1989, Lawrence Erlbaum Associates, Inc.

## A Taxonomy of Multiple-Choice Item-Writing Rules

Thomas M. Haladyna  
*Arizona State University  
West Campus*

Steven M. Downing  
*National Board of Medical Examiners*

A taxonomy of 43 multiple-choice item-writing rules is presented and discussed. The taxonomy derives from an analysis of 46 authoritative textbooks and other sources in the educational measurement literature. The analysis also leads to a “validity by consensus” for each rule. The taxonomy is viewed as a complete and authoritative set of guidelines for writing multiple-choice items.

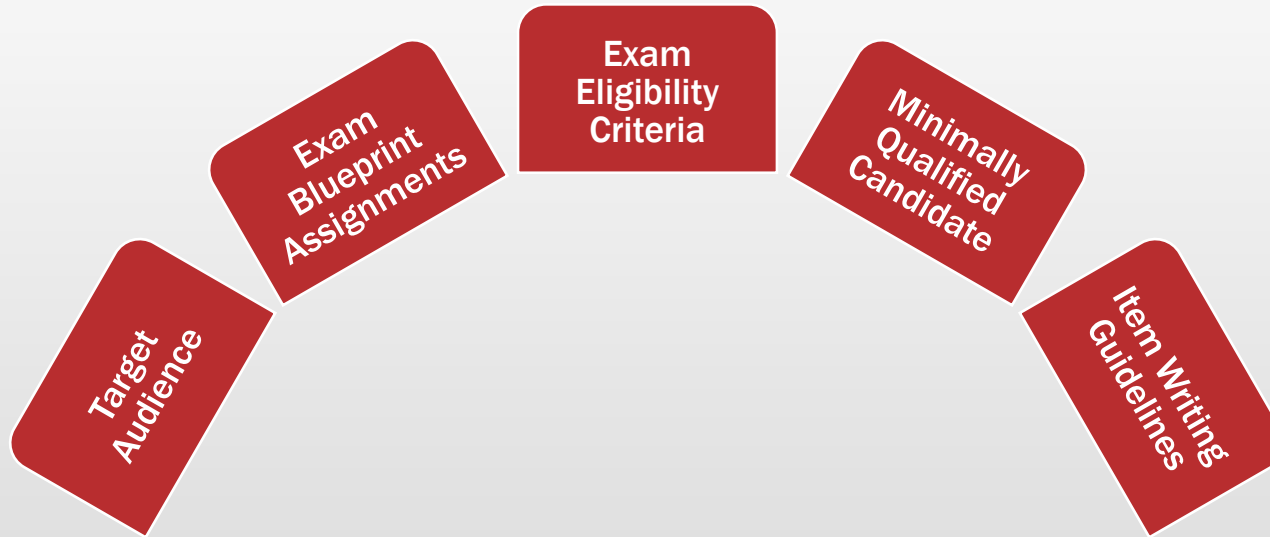
In the development of any achievement test, one of the most fundamental steps is writing the test item. Each of the three editions of the prestigious *Educational Measurement* devotes at least a chapter to item writing, and in most textbooks on educational measurement, a chapter or significant passage is devoted to multiple-choice (MC) item writing. Measurement experts as well as testing organizations prefer the MC format for many reasons:

1. Sampling of content is generally superior when compared to other formats; the use of MC formats generally leads to more content-valid test-score interpretations.
2. Reliability of test scores can be very high with sufficient numbers of high-quality MC items.
3. MC items can be easily pretested, stored, used, and reused, particu-

---

Requests for reprints should be sent to Thomas M. Haladyna, Department of Education and Human Services, West Campus, Arizona State University, P.O. Box 37100, Phoenix, AZ 85069-7100.

# General Considerations



# Item Anatomy

Who was the first president of the United States under the **Continental Congress**?

Stem = Question

Key = Correct  
Answer

- A. John Hanson
- B. John Adams
- C. Thomas Jefferson
- D. George Washington

Distracters =  
Wrong Answers

Options = Key +  
Distracters



# Stem

## Formats

- Open
  - The year that John Adams was elected President:
- Closed (Preferred)
  - In what year was John Adams elected President?

## Best Practices

- Succinct –remove unwanted language
- Relevant and important
- Non-trivial
- Stem is **NOT** teaching
- Avoid using “Not” and “Except”
- Avoid using definitions in stem
- Avoid two questions at once

## Quality Check

- Can you cover the options and answer the question?
- What is the capital of France?
  - A. Lyon
  - B. Paris
  - C. Normandy
  - D. Orlean

# Key: Best Practices

**Should NOT be systematically different from distractors**

- Longest
- Contains technical jargon

**Don't use words that are in the stem**

- Known as cueing

# Distractors

## Best Practices

- **Incorrect**
- **Plausible**
- **Common misconceptions**
- **No overlap**

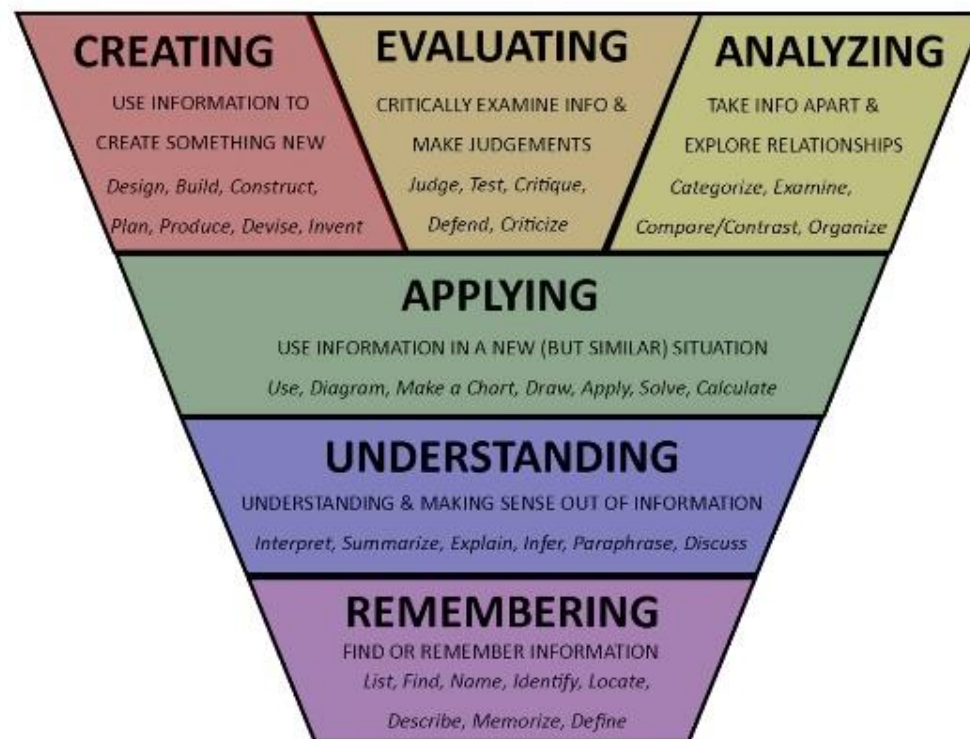
# Reference and Rationale

## Best Practices

- Use the approved reference list only
  - If unavailable, use the more common and often cited references
- When providing a rationale, consider that a **majority** of other SMEs would agree with your rationale

# Determine Cognitive Level to Evaluate with Items

In general, consider writing “**application**” level questions rather than “**remembering**” or “**understanding**” level questions to evaluate a deeper level



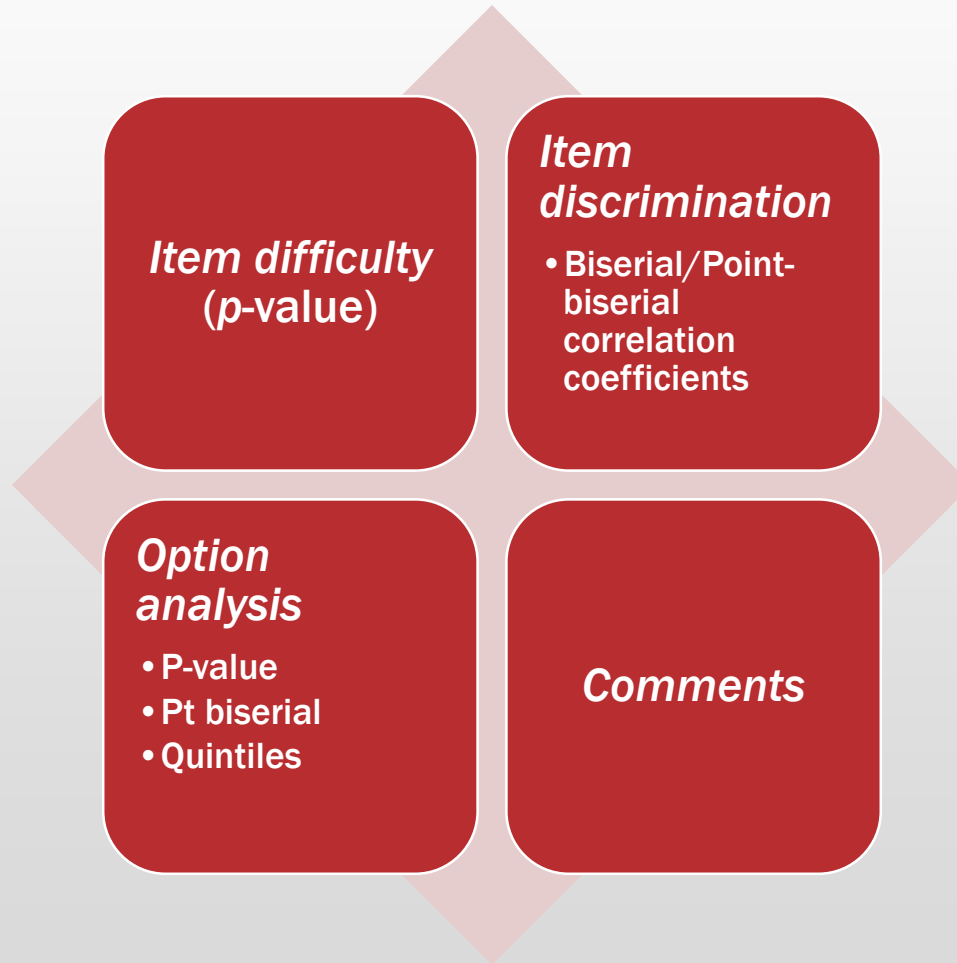
# How long does it take to write an item?

<b>Cognitive Level of Item</b>	<b># of Items Written in a Day</b>	<b>Estimated Time to Complete One Item</b>
Understanding	12-15	30 minutes
Application	10-12	30-45 minutes
Problem Based Items	6-8	45-60 minutes



# Item Analysis

# A First Look at Item Performance: Basic Item Analysis



# Item Difficulty Index: p-value

Proportion of candidates who correctly answer a test item

Ranges from 0 – 1

- Polytomously scored = Recommend dividing average score by total number of points possible so on same scale regardless of number of points

Low values = “Difficult” items

High values = “Easy” items

# How “Difficult” Should Items Be?

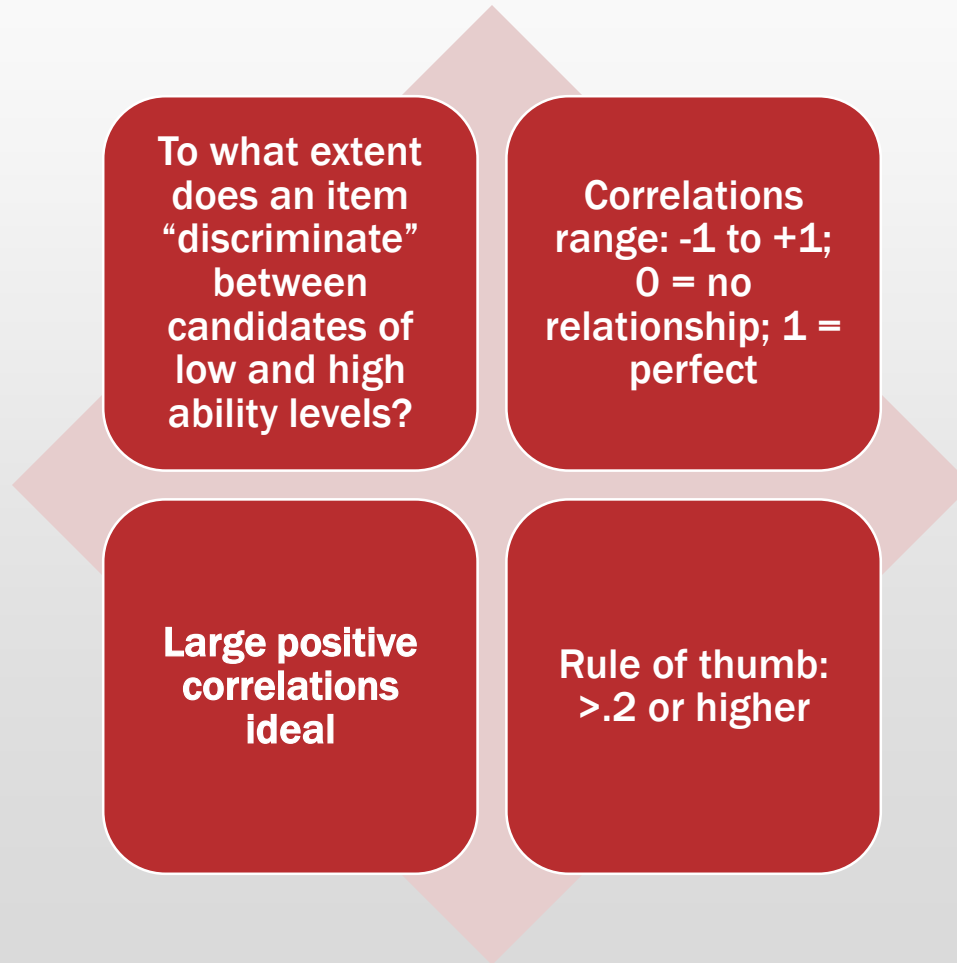
**General rule of thumb: 0.3 to 0.7**

- $p$ -value = .5 provides max information about candidates
  - If  $p$ -value = 0.5 then variance =  $0.5 * (1 - 0.5) = 0.25$  (max variance)

**Avoid items with  $p$ -values near 0 or 1**

- No information provided unless needed for content validity reasons

# Item Discrimination



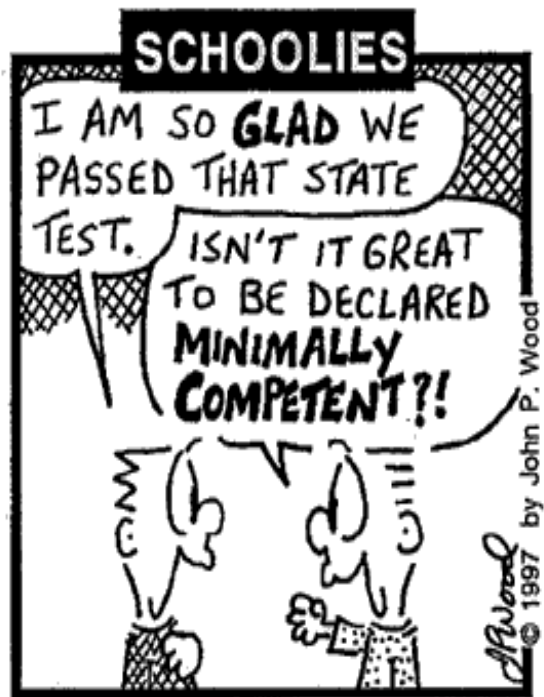
# Item Discrimination Interpretation

<u><math>R_{pbis/bis}</math> range</u>	<u>Interpretation</u>
If $r_{pbis/bis} \geq 0.30$	Item is functioning very well
If $r_{pbis/bis} [0.20 - 0.29]$	Little or no revision required
If $r_{pbis/bis} [0.10 - 0.19]$	Item is marginal and needs to be revised
If $r_{pbis/bis} < 0.10$	Item requires serious revision or should be eliminated

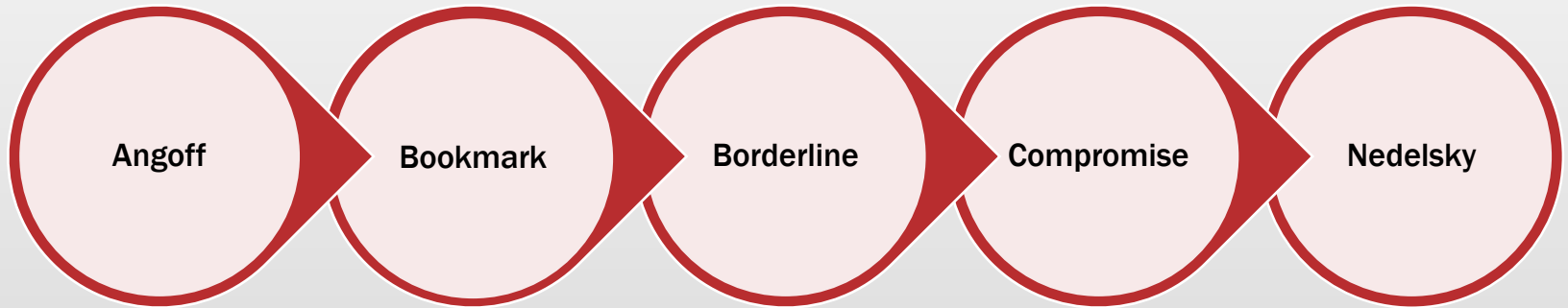


# Standard Setting

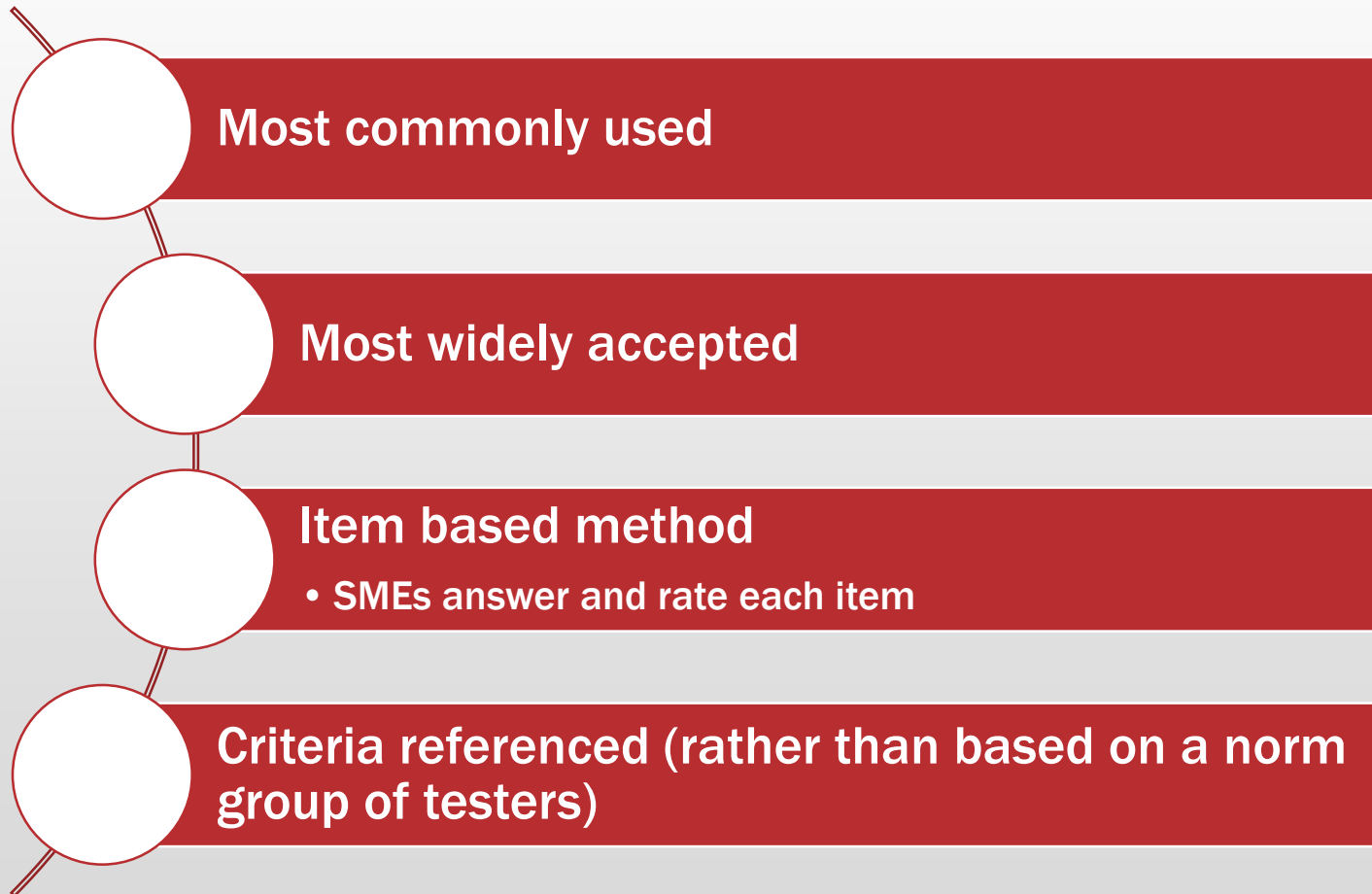
# Definition



# Many Methods but Only a Few Are Used



# Angoff Method



# Reliability

# Forms of Reliability

**Test-retest  
Reliability:  
Consistency across  
time**

- Same general rank order between administrations

**Alternate forms  
reliability; interrater  
reliability:  
Consistency across  
forms**

- Each form is designed to measure same content areas in the same manner

**Internal consistency;  
split-half reliability:  
Consistency among  
items**

- Cronbach's alpha
- Determined by interrelatedness of the items and test length

# Rules of Thumb for Internal Consistent Reliability

- $\alpha \geq 0.9$  = Excellent
- $0.7 \leq \alpha < 0.9$  = Good
- $0.6 \leq \alpha < 0.7$  = Acceptable
- $0.5 \leq \alpha < 0.6$  = Poor
- $\alpha < 0.5$  = Unacceptable

# Factors to Consider

- In general, longer exams are more reliable
- How many skills, abilities, etc. do you need to cover?
- How many items are in your item pool?
- How many test forms will you create?
- How much overlap is acceptable?





# Rules of Thumb

<b>Reliability</b>	To meet reliability the number Items on a form should often be 21 or greater. There are exceptions but for most credentialing exam the SMEs will often believe there are more items necessary.	Cortina, J.M., (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. <i>Journal of Applied Psychology</i> , 78(1), 98–104.
<b>Items to be Developed (Item Banking)</b>	“For selected response, a rule of thumb is that the item bank should be 2.5 times the size of a test.”	Haladyna, T.M., & Rodriguez, M.C. (2013). <i>Developing and validating test items</i> . New York, NY: Routledge. Page 17
<b>Testing Time</b>	<p>“Clear majority of examinees should have reached and attempted 90% or more of the items in a test.”</p> <p>Characteristics of the testing sample also plays a factor in how long (e.g., items in German have greater reading loads than most other languages).</p>	Schmeiser, C.B., & Welch, C.J. (2006). <i>Test development</i> . In Brennan, R.L. (Ed.), <i>Educational Measurement</i> (4 <sup>th</sup> ed.). Westport, CT: Praeger. – Page 338
<b>Distribution of Cognitive Items (optional)</b>	“...should be based on empirical data collected in a systematic way,” such as a Job Task Analysis/Practice Analysis Results	Schmeiser, C.B., & Welch, C.J. (2006). <i>Test development</i> . In Brennan, R.L. (Ed.), <i>Educational Measurement</i> (4 <sup>th</sup> ed.). Westport, CT: Praeger. – Page 316
<b>Content</b>	“In many cases the test domain must be prioritized to measure knowledge and skills judged to be most important by the relevant test audiences. The emphasis gathered through empirical survey data can serve as the basis for distributing items across these domains.”	Schmeiser, C.B., & Welch, C.J. (2006). <i>Test development</i> . In Brennan, R.L. (Ed.), <i>Educational Measurement</i> (4 <sup>th</sup> ed.). Westport, CT: Praeger. – Page 319

# Validity

# Overview

## Defined

- How well an exam measures what it is meant to measure
- A property of how the exam is used (scores are interpreted) rather than of the exam itself

## Ensuring Validity

- Exam objectives must be derived from job role requirements and skills needed to use the product
- Exam must include items that cover all functional groups and major objectives
- Exam content must be representative of the appropriate domain of knowledge
- Subject matter experts (SMEs) should review the objectives and items; revisions should be incorporated as necessary

## •What this REALLY Means

- Appropriateness of **inferences or judgments** based on test scores, given supporting empirical evidence

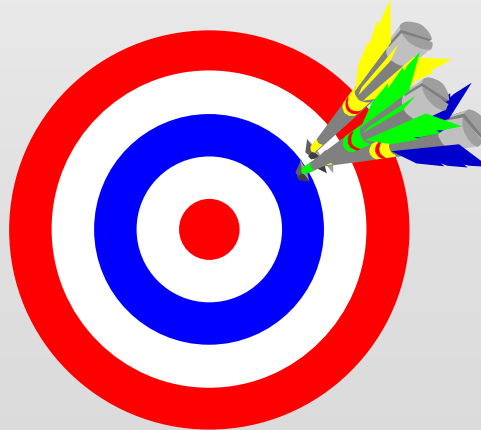
# Relationship between Validity and Reliability

An exam can be reliable without being valid, but a test cannot be valid without being reliable

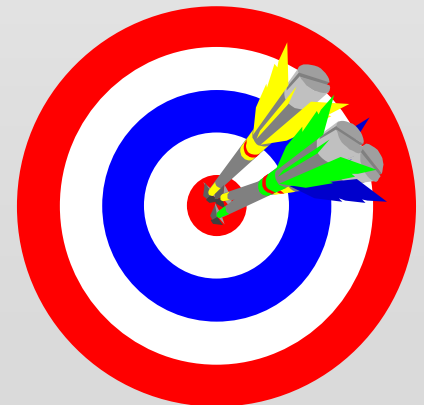
**Unreliable and Invalid**



**Reliable but Invalid**



**Reliable and Valid**



# What are critical steps in validating exams?

**Clearly lay out the claim that you'd like to make based on the candidate test scores**

- Is it clear and coherent?
- Is it plausible given the empirical evidence at hand?
- What claims would your test NOT support?

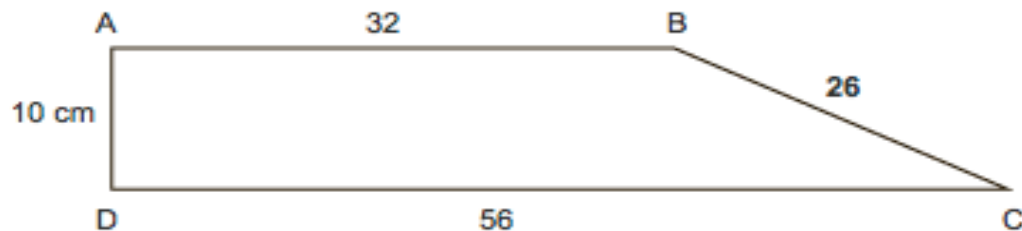
**Don't claim more than what is supported by evidence**

# Test Fairness

# Example #1

## The Mistake\*

According to the State Education Department, the question should have said to double AB, CD, and AD. BC would then be doubled because of the proportionality of similar triangles, a rule fifth graders learn.



$$\text{Perimeter} = 10 + 32 + 26 + 56 = 124 \text{ cm}$$

# Example #2

Premise 1: Females in patriarchal societies cannot make healthcare decisions for themselves.

Premise 2: The United States is a patriarchal society.

Conclusion: The United States should not allow women to make healthcare decisions for themselves.

If the first two premises are true, the conclusion is:

- A. true.
- B. false.
- C. uncertain.



# What is Fairness?

***“Although fairness has been a concern of test developers and test users for many years, we have no widely accepted definition”***

**p. 25 Haladyna and Rodriguez (2013)**



# What is Test Fairness?

## SIOP

- Equal group outcomes
  - Passing scores are relatively equal for subgroups (males and females)
- Equal treatment
- Comparable opportunity to learn material

## ETS

- **Items:**
  - Are not offensive or controversial
  - Do not reinforce stereotypical views of any group
  - Are free of racial, ethnic, gender, socioeconomic and other forms of bias
  - Are free of content believed to be inappropriate or derogatory toward any group

**Unfairness is anything that adds construct irrelevant variance to assessment process**

# AVOID UNNECESSARY DIFFICULTY IN LANGUAGE OF QUESTION

- Some groups are familiar with aspects of question while others are not based on life experiences
- Topics to be avoided:
  - Military
  - Regionalism
  - Religion
  - Specialized tools
  - Sports
  - US centric
  - Etc.

## WHAT IS THE ISSUE?

A family decides to put in a swimming pool. The pool will be 8 feet long and be 48 square feet.

How wide will the pool be?

- A. 6 feet
- B. 7 feet
- C. 8 feet
- D. 9 feet

**Potential fairness issue:**  
Those from lower economic statuses may not realize that swimming pools are square or rectangular.

# Other Fairness Considerations

**Selection of subject matter experts  
(sampling)**

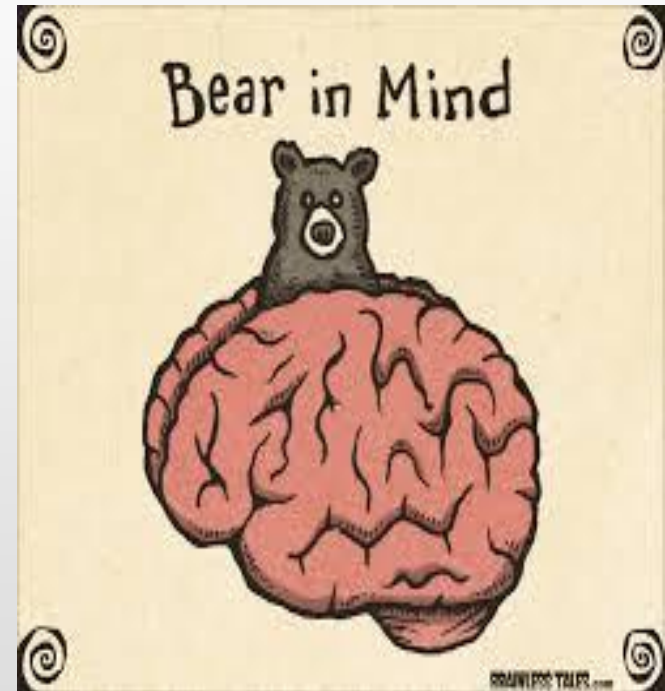
**Minimizing external influences on testing  
process**

**Statistical methods - differential item  
functioning (DIF) - uncovers bias towards  
one group**

# Summary

*“...any characteristics of items that affect test scores and are unrelated to what is being measured is unfair”*

*p. 25 Haladyna and Rodriguez (2013)*



# Questions



**Manny Straehle, Ph.D., GISF**

Founder and President

[manny@aerexperts.com](mailto:manny@aerexperts.com)

**Liberty J. Munson, Ph.D.**

Chief Psychometrics Officer

[liberty@aerexperts.com](mailto:liberty@aerexperts.com)

Assessment, Education, and Research Experts (AERE)

[www.aerexperts.com](http://www.aerexperts.com)

Client-Centered, Solution-Focused, and Practicing the Practical

# References

AERA, APA, NCME. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Westport, CT: Praeger.

Francis, G. (Ed.). (2007). *Behavior Research Methods*. New York: Springer.

Linn, R. L. (Ed.). (1989). *Educational Measurement*. New York: Macmillan.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Whitley, B. E. (1996). *Principles of Research in Behavioral Science*. Mountain View, CA, Mayfield.