ATP 2014
Innovations in Testing
Scottsdale, AZ

INSIGHT
Happens Here

Psychometric Rules of Thumb that Every Credentialing Manager Should Know

Dr. Liberty Munson Dr. Ada Woo

Dr. Manny Straehle

Overview

Manny

Sample Size

Test Fairness

Liberty

Intro & Basic Terminology

Item Analysis

Validity

Ada

Reliability

Number of Items per Form

Item Exposure

Ask a Psychometrician?

1. What have been your experiences with Psychometricians?
2. What questions have not been answered to your satisfaction?
3. Are Psychometricans contradictory from one consultant/vendor to another? Tell us how?
4. What don't you understand about psychometrics that you wish you did?

Disclaimer

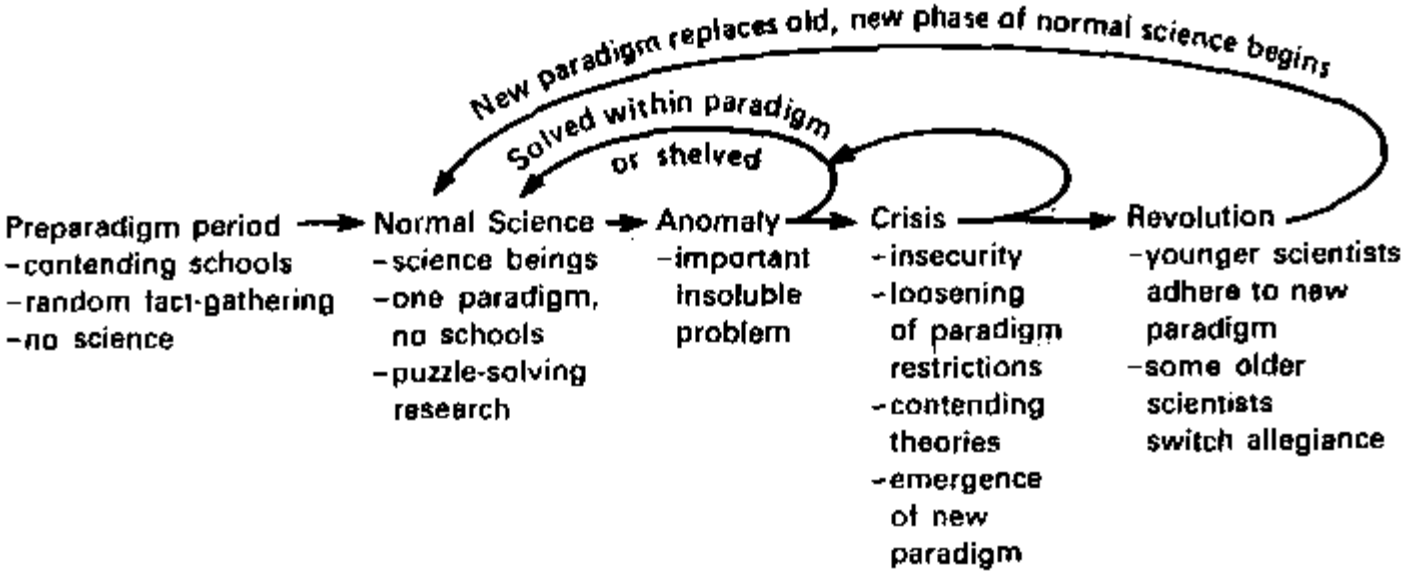
- These rules of thumb can be considered as provisional guidelines. For several of these, we provide supporting references (see handout).

- This workshop is aimed at managers and executives of credentialing programs rather than psychometricians.

Disclaimer

- We make no claim in defending these rules of thumb, especially when innovative/alternative methods may have been accepted by industry peers.

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



Basic Terminology

Some Basic Terminology

- **What is an examination?**
  - 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 items, cases or stations from which one or more test forms can be assembled by selecting a sample of items, cases or stations from this population

Some Basic Terminology

- **Examination**
  - 100 item comprehensive mathematics high-school graduation exam administered at the end of the 12th grade
- **Domain**
  - The (theoretically infinite) pool of math items (sequences and series, functions, trigonometry, polynomials, calculus, geometry, statistics, etc.) from which you selected 100 items to include in your graduation examination

Sample Size

CHOCOLATE DEMONSTRATION

How Many SMEs?

Does the sample represent the population?

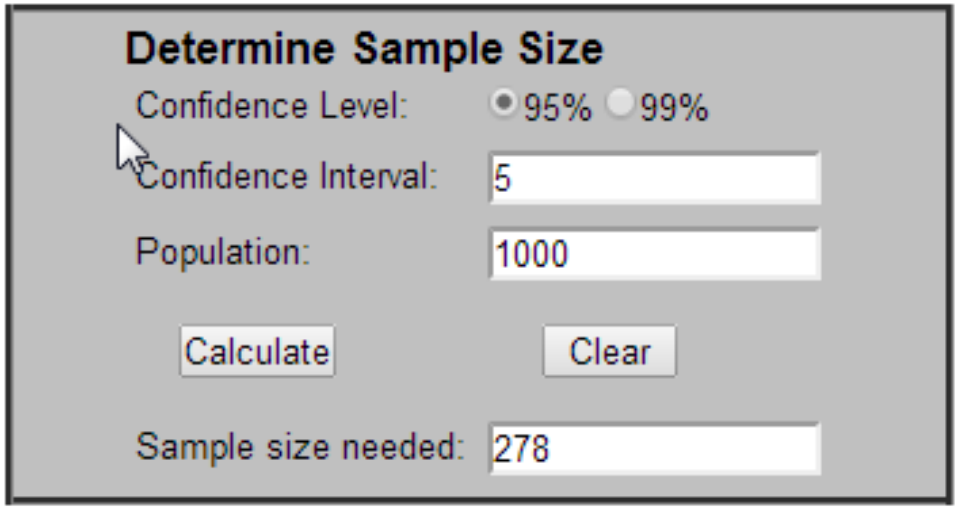
- Sample of Sommeliers (n=100)
  - Experience
  - Education/Training
  - Industry
  - Geography
  - Gender
  - Ethnicity
- Population of Sommeliers (N=1000)
  - Experience
  - Education/Training
  - Industry
  - Geography
  - Gender
  - Ethnicity

Two Sampling Methods

- Random Sampling
  - Obtaining a certain percentage of a sample at random will lead to greater confidence that your sample represents the population
- Stratified Sampling
  - Sample represents percentage of population
    - 5% Male
    - 5% White
    - 5% 20+ Year of Experience
    - 5% from US Southeast

Sampling Calculators/Tables

- Definitions
  - Confidence Level
    - If you sampled100 different times, your results would be the same 95% of the time
  - Confidence Intervals (Margin of Error)
    - Your results would be in a range of +/- 5%



Sampling calculators/tables available at http://measurementstatistics.wikispaces.com/Sample+Size

Sampling Example

Poll Example

- 90% of the survey takers believed that Mr. T would be the next US president.
  - Sample size was 1000
  - Can conclude that this result is
    - +/-3% margin of error
      - 87 to 93%
    - 99% confidence that this result would occur 99 out of 100 times

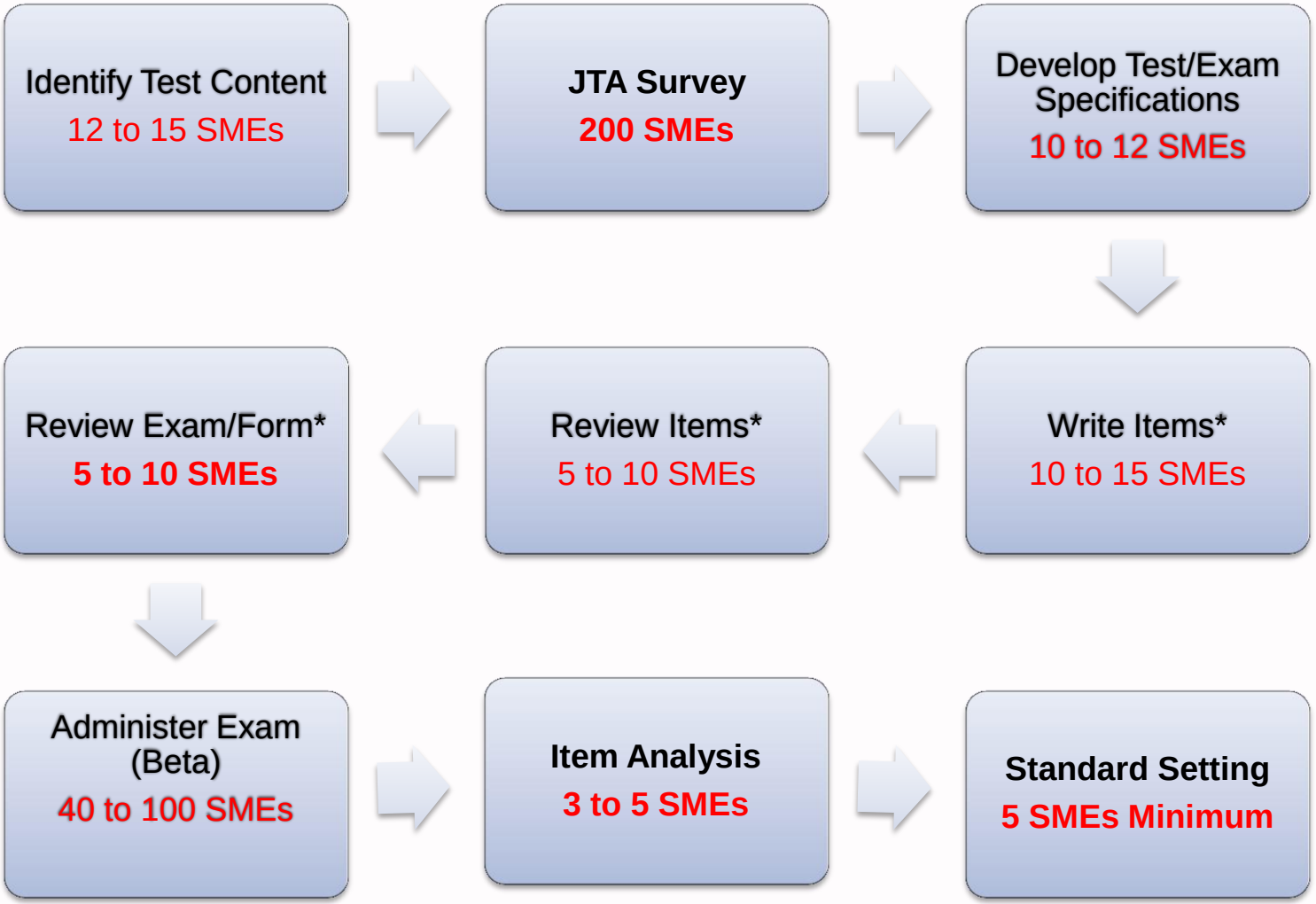
Sampling Exercise

- Total of 500 credential holders (population). So, how many survey participants do I need to perform a Job Task Analysis?
- Confidence Interval is set at 95%
- Margin of Error 5%
- **ANSWER: 278**

Required Sample Size†

Population Size	Confidence = 95% Margin of Error				Confidence = 99% Margin of Error			
	5.0%	3.5%	2.5%	1.0%	5.0%	3.5%	2.5%	1.0%
10	10	10	10	10	10	10	10	10
20	19	20	20	20	19	20	20	20
30	28	29	29	30	29	29	30	30
50	44	47	48	50	47	48	49	50
75	63	69	72	74	67	71	73	75
100	80	89	94	99	87	93	96	99
150	108	126	137	148	122	135	142	149
200	132	160	177	196	154	174	186	198
250	152	190	215	244	182	211	229	246
300	169	217	251	291	207	246	270	295
400	196	265	318	384	250	309	348	391
500	217	306	377	475	285	365	421	485
600	234	340	432	565	315	416	490	579
700	248	370	481	653	341	462	554	672
800	260	396	526	739	363	503	615	763
1,000	278	440	606	906	399	575	727	943

Test Development Lifecycle



Item Analysis

A First Look at Our Scores: Item Analysis (mostly for MCQs)

Basic item-level psychometric analyses

- *Item difficulty* (*p*-value)
- *Item discrimination*
  - Discrimination index (*D*)
  - Biserial/Point-biserial correlation coefficients
- *Distractor analysis*
  - Quintiles table

Item Difficulty Index: p-value

- ***p*-value**
  - Proportion of candidates who correctly answer a test item
  - Ranges from 0 – 1 (when dichotomously scored)
    - Polytomously scored = p-value is average score and ranges from min to max score possible
  - Low values are indicative of “difficult” items
  - High values are indicative of “easy” items

How "Difficult" Should Items Be?

- General rule of thumb: **0.3-0.7** since items maximize information exam provides about differences between candidates
  - Item *p*-value of .5 provides max. information
    - $Var_{item\ score} = p_i(1-p_i)$
    - If *p*-value=0.5 then Var = 0.5(1-0.5) = 0.25
- **Try to avoid items with *p*-values near 0 or 1**
  - No information provided unless needed for content validity reasons

How "Difficult" Should Items Be?

- Tests that employ a cut-score (passing standard)
  - Select items that maximize information near the cut-score
  - More easily accomplished using IRT

Item Discrimination

- To what extent does an item "discriminate" between candidates of low and high ability levels?
- **What do we expect to see?**
  - Candidates who are more proficient on the exam should correctly answer an item in a higher proportion than those who are less able
  - If not, item is unrelated to constructs targeted by examination!

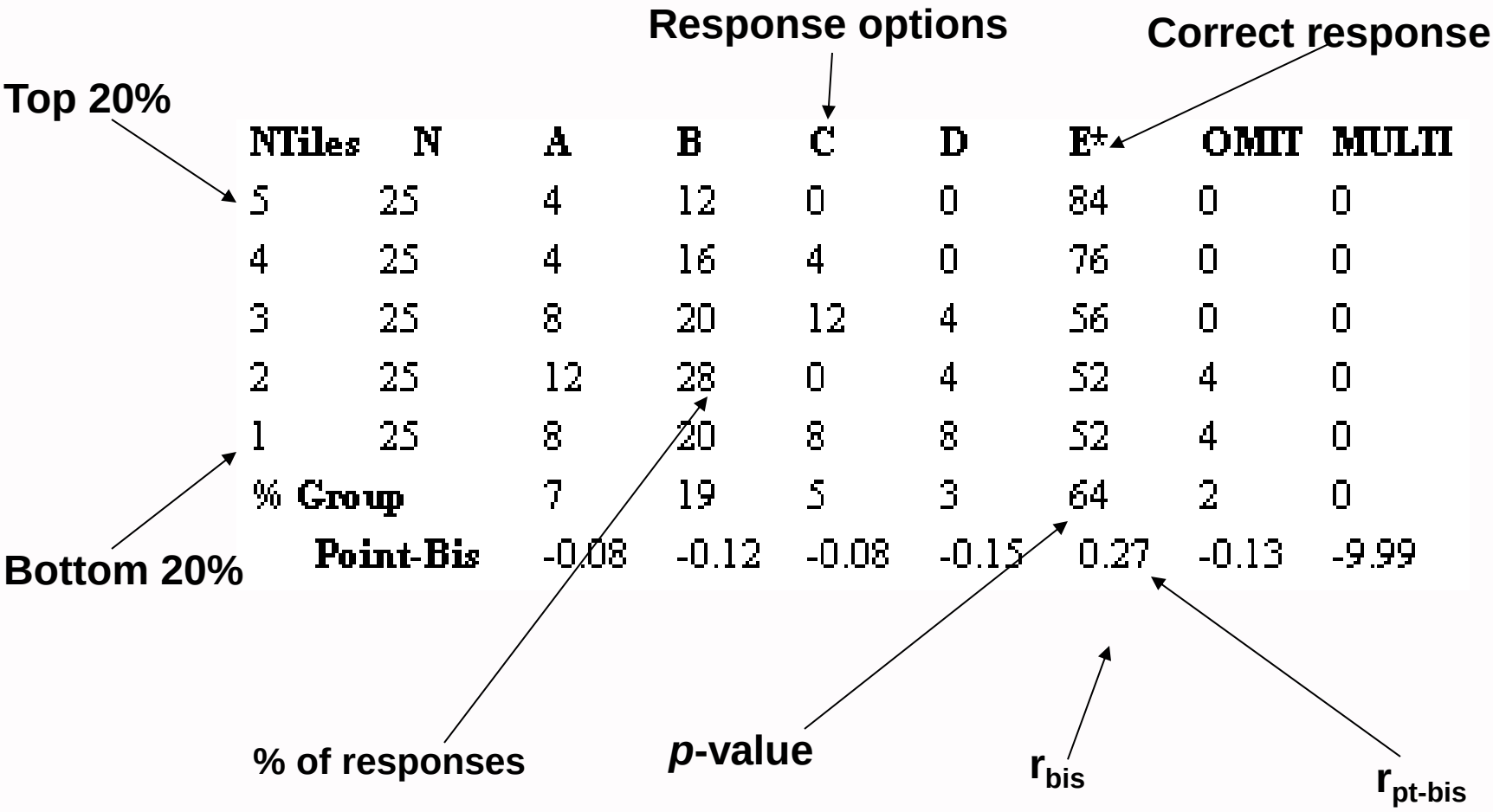
Item Discrimination

- **Point-biserial correlation coefficient r_{pbis}**
  - An index indicating the degree of relationship between the score on an item (0 or 1) and a criterion score (e.g. total or section score)
  - Negative values = high performers not answering item correctly
  - 0 = No relationship
  - Positive correlation = high performers answering correctly

Item Discrimination - Ranges

$R_{pbis/bis}$ range	Interpretation
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

Distractor Analysis



NTiles	N	A	B	C	D	E*	OMIT	MULTI
5	25	4	12	0	0	84	0	0
4	25	4	16	4	0	76	0	0
3	25	8	20	12	4	56	0	0
2	25	12	28	0	4	52	4	0
1	25	8	20	8	8	52	4	0
% Group		7	19	5	3	64	2	0
Point-Bis		-0.08	-0.12	-0.08	-0.15	0.27	-0.13	-9.99

Example

- An adult female rabbit has a sudden onset of paralysis of the hindquarters and urinary incontinence. Which of the following is the most likely diagnosis?

  (A) Idiopathic demyelination

  (B) Metastatic uterine adenocarcinoma

  (C) Mycotoxicosis

  (D) Saddle thrombus

  (E) Spinal injury from inappropriate handling

Example 1

NTiles	N	A	B	C	D	E*	OMIT	MULTI
5	25	0	0	0	0	100	0	0
4	25	0	0	0	0	100	0	0
3	25	0	4	0	0	96	0	0
2	25	4	0	4	0	92	0	0
1	25	0	16	0	0	84	0	0
% Group		1	4	1	0	94	0	0
Point-Bis		-0.08	-0.23	-0.05	-9.99	0.24	-9.99	-9.99

An Example of a Psychometrically Poor Item

You are creating an ASP.NET application for an online store that sells movies on videocassette. Each user is assigned a profile based on the user's previous purchases.

You write a procedure named DisplayRecommendations that calls the LoadUserProfile function and displays a list of movie recommendations when a user logs on. The LoadUserProfile function throws a FileNotFoundException if the user profile cannot be found.

When the FileNotFoundException is thrown, you want to throw a more descriptive error. The text of the error message is stored in a variable named descriptionString. You also want the FileNotFoundException error to be accessible programmatically for debugging purposes.

You need to program the catch block for the exception. Which code should you use?

A.
```
catch (ApplicationException ex)
{
  throw (new App
    ex));
\}
```

B.
```
catch (FileNotFo
{
  throw (new Appli
    ex));
\}
```

C.
```
catch (Application
{
  throw (new Appli
    ex.InnerException));
\}
```

D.
```
catch (FileNotFoundException ex)
{
  throw (new ApplicationException(descriptionString,
    ex.InnerException));
\}\}
```

Name	N	P-Value	Mean Time	Median Time	Point Biserial	Item Reliability	Rasch b (Difficulty)	Infit MNSQ	Outfit MNSQ	Status
6.5.b	2433	0.30	110.32	94	-0.13	-0.06	3.10	2.10	3.99	Delete

Item	Response	Key	N	Percent	Pt Bis	Q1	Q2	Q3	Q4	Q5
6.5.b	A		66	0.027127	-0.11729		0.00349	0.007052	0.009331	0.016525
6.5.b	B	Key	725	0.297986	-0.13495	0.06015	0.054337	0.126939	0.138414	0.106744
6.5.b	C		121	0.049733	-0.14304		0.004985	0.01481	0.018663	0.029924
6.5.b	D		1502	0.617345	0.23267		0.291127	0.222849	0.188958	0.164806

Reliability

Reliability

- The reliability of a measure is its degree of consistency.
- A perfectly reliable measure gives the same result every time it is applied to the same person or object, barring changes in the variable being measured.

Reliability (cont.)

An observed test score (X) is a function of two sources: a “true” score (T) and “error” (E)

$$X = T + E$$

X = Observed test score

T = True source

E = Random error

Forms of Reliability

1. Consistency across time

- Examinee scores preserve the same general rank order from first to second administrations
- e.g., class room tests
- Test-retest reliability

Forms of Reliability (cont.)

2. Consistency across forms

- Equivalency in alternate forms
- Each form is designed to measure the same construct (e.g. content areas) in the same manner.
- e.g., Paper-and-pencil tests, fixed form computer-based tests
- Alternate forms reliability, interrater reliability

Forms of Reliability (cont.)

3. Consistency among items

- Tests that use multiple items to assess a trait, with the sum of a person's scores on the items being the total score of the measure.
- Internal consistency, split-half reliability

Cronbach's Alpha (α)

- Estimate of internal consistency
- Ranges from 0 to 1 (correlation)
- Determined by interrelatedness of the items and test length
- **Rule of thumb:**
  - $\alpha \geq 0.9$ – 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

Reliability Indices

- Correlation indices
- Range from 0 to 1
- Estimate ratio of true score variance to observed score variance
- Generic formula:

$$\rho_{xx} = \frac{\sigma^2_T}{\sigma^2_X}$$

Number of Items per Form

Number of Items per Forms

- In Classical Test Theory, an easy way to make tests more reliable is to make them longer.
- Several factors to consider:
  - Internal reliability of items
  - Number of test plan categories
  - Available usable items in the inventory
  - Number of test forms and allowable overlap among them

Spearman-Brown Prophecy Formula

- If the number of items on a test increased by *X*, by how much would reliability increase?
  - $$r_{kk} = \frac{kr_{11}}{1+(k-1)r_{11}}$$

r_{11} = original reliability; r_{kk} = new reliability

k = factors by which the number of items were increased

e.g., if 40 items are added to a 20-item test, the test would be lengthened by a factor of 3.

i.e., 60 ÷ 20 = 3

Spearman-Brown Calculation Example

- e.g., the reliability of a 20-item test is 0.70, you'd like to add 40 items from the same domain to the test. What would the reliability of the 60-item test form be?

$$r_{kk} = \frac{3 \times 0.7}{1 + (3 - 1) \times 0.7}$$

$$r_{kk} = \frac{2.1}{1 + 1.4}$$

$$r_{kk} = 0.88$$

Rules of Thumb

Consideration	Rule of Thumb	Supporting Document
Reliability	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. Journal of Applied Psychology, 78(1), 98–104.
Items to be Developed (Item Banking)	"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). Developing and validating test items. New York, NY: Routledge. Page 17
Testing Time	"Clear majority of examinees should have reached and attempted 90% or more of the items in a test" 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.	Schmeiser, C.B., & Welch, C.J. (2006). Test development. In Brennan, R.L. (Ed.), Educational Measurement (4th ed.). Westport, CT: Praeger. – Page 338
Distribution of Cognitive Items (optional)	"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). Test development. In Brennan, R.L. (Ed.), Educational Measurement (4th ed.). Westport, CT: Praeger. – Page 316
Content	"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). Test development. In Brennan, R.L. (Ed.), Educational Measurement (4th ed.). Westport, CT: Praeger. – Page 319

Validity

Validity

Validity 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

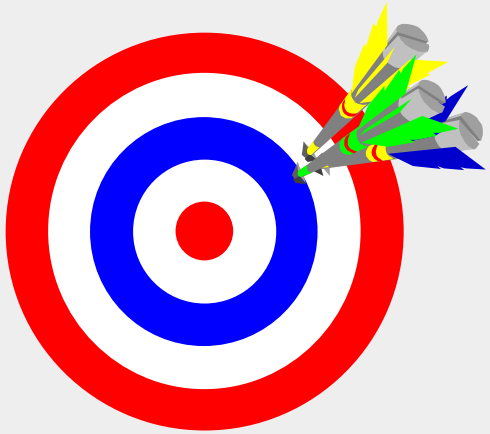
Validity – What It's Not

- **There is no such thing as a valid or invalid test**
  - Statements such as "my test shows construct validity" are completely devoid of meaning
  - Validity refers to the appropriateness of **inferences or judgments** based on test scores, given supporting empirical evidence
    - Standards for Educational and Psychological Testing (AERA, APA, NCME)

Relationship between Validity and Reliability

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

Reliable but Invalid



Reliable and Valid

What this Means When Establishing Validity

- Clearly lay out the intended use of the test
  - **What do I want to infer based on my test scores? What judgment or argument do I want to make?**
- Gather as much (empirical) evidence as possible to support the (intended) score-based inferences

Approaches to Validity

- Five basic arguments

1. **Evaluation**: Evaluate the candidates' performances on the exam
2. **Generalization**: Do the performances generalize to the domain of tasks?
3. **Extrapolation**: Do the performances generalize to other settings/performance formats?
4. **Explanation**: Do scores reflect performance on the constructs intended to be measured by the exam?
5. **Decision making**: Can the performances be used for placement decisions?

Validity of a test → **Validity of a score** → **Validity of an argument**

Approaches to Validity

- **Evaluation (scoring) argument**
  - Are exam performances (observations) scored accurately with respect to the construct(s) measured?
  - Does the (psychometric) model fit exam scores?
  - Is scoring properly documented?
- **Generalization argument**
  - Do the observed (actual exam) scores represent what would be obtained in the broader domain?
  - Relates to precision of measurement (reliability)

Approaches to Validity

- **Extrapolation argument**
  - Does test performance extrapolate to practice and other outcome indicators?
  - What are the sources that are likely to undermine this extrapolation?
    - Analysis of relationship between performance on exam (sample) and a broader criterion (e.g., workplace assessment)

Approaches to Validity

- **Explanation argument**
  - Do scores reflect performance on the constructs intended to be measured by the exam?
    - Confirmatory factor analysis of data set
    - Mapping of expert judgments of content on examination
- **Decision (consequential) argument**
  - Do scores reflect readiness for practice?
  - Do candidates with a low skill level fail the exam and those with a high skill level pass the exam?
    - Relationship of performance on exam to other criteria

What are critical steps in validating exams?

- Clearly lay out the claim/interpretive argument that you'd like to make based on the candidate test scores and challenge it
  - Is it clear and coherent?
  - Is it plausible given the empirical evidence at hand?
  - What claims/interpretations would your test NOT support?

- **Don't claim more than what is supported by evidence**
  - Avoid unsubstantiated "blanket statements"
    - "My test shows construct validity"
    - "My exam has face validity", etc.

Validity – Exercise

- **Scenario 1**
  - The admissions dean at your business school has asked you to develop an exam that will be used to admit students to your MBA program
- **Scenario 2**
  - Think of an examination program that you may have been involved in the past
- **For each scenario:**
  - What is the interpretation that you'd like to make based on the exam score?
  - What sources of evidence can you gather to support this/these argument(s)?
  - What inferences would NOT be supported by your sources of evidence?

Item Exposure

Considerations

- Item inventory
  - Number of times it has been used
  - Item format
  - Number of items on exam
- Candidate volume
- Purpose of the test
- Item production speed

Considerations (cont.)

- Test forms publishing cycle
  - Rotate forms
  - Retire/temporarily stopping using items that have certain amount of takes over so many years
- Test administration method and region
  - Paper and Pencil versus Computerized
  - e.g., China has a high percentage of suspected item compromise
- Type of exam
  - High stakes vs low stakes
- Exposure control parameter

Mitigating Over Exposure

- Item cloning
- Single-use items
- Web patrol monitoring
- Incorporate exposure control into testing methodology

Item Parameter Drift Analysis

- Drift occurs when item parameter estimates change between test administrations
- **Parameter change can be a result of:**
  - Changes in educational instructions
  - Changes in practice
  - Item compromise (test security breach)
- Conduct periodical statistical monitoring and consult subject matter experts

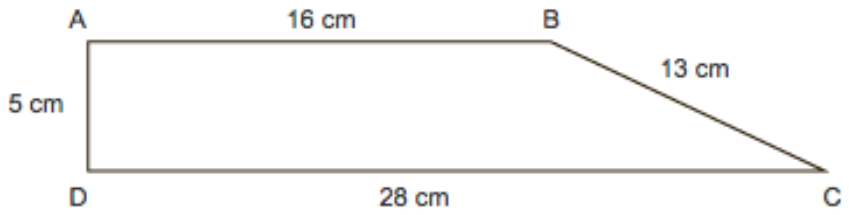
Test Fairness

Example #1 – Too Hard and Wrong

Question is too difficult for 5th graders

The Problem

Trapezoid ABCD is shown below.



A new trapezoid is formed by doubling the lengths of sides AB and CD. Find the perimeter, in centimeters, of the new trapezoid.

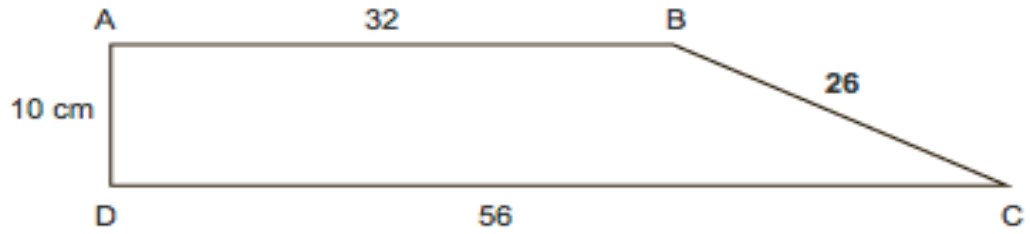
Show your work.

Source: http://www.wnyc.org/story/302903-state-officials-throw-out-another-pearson-test-question

Example #1 - Continued

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.



Perimeter = 10 + 32 + 26 + 56 = **124 cm**

Example #2 – Offensive Content

- Females in most societies don't have any authority in making healthcare decisions for themselves. Which of the following would be the reason this should be adapted in the United States?

What is Fairness?

- Many Definitions and Sometimes Contradictory
- Rules of Thumb Less Established

"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 Standards**
  - Equal Group Outcomes
    - Passing scores are relatively equal for subgroups (males and females)
      - Not very popular
  - Equal Treatment
    - Test conditions
  - Comparable opportunity to learn material

- **ETS General Guidelines**
  - 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

Zieky - Summary of Six ETS Guidelines from 2003

- Treat People with Respect
- Minimize the effects of construct-irrelevant knowledge/skills
- Avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting
- Use of appropriate terminology to refer to people
- Avoid stereotypes
- Represent diversity in depictions of people

ETS Listening. Learning. Leading.®

ETS Guidelines for Fairness Review of Assessments

2009

GUIDELINE 1. AVOID COGNITIVE SOURCES OF CONSTRUCT-IRRELEVANT VARIANCE

- Overall – Makes it cognitive difficult for all or specific groups
- Unnecessarily Difficulty in Language
- Topics to be avoided
  - Military
  - Regionalism
  - Religion
  - Specialized tools
  - Sports
  - US

WHAT IS THE ISSUE?

The item below is for an exam on basic statistics.

$$\overline{X} = \frac{\sum X}{N}$$

Q. Calculate the mean for the number of abortions among Catholic members in the United States?

Potential fairness issue:

1. Reference to religion and abortion for an exam on basic statistics

GUIDLEINE 2. AVOID COGNITIVE SOURCES OF CONSTRUCT-IRRELEVANT VARIANCE

- Primary purpose - Avoid emotional reactions from inappropriate content
- Topics recommended to be avoided
  - Accidents, incidents, or illness
  - Advocacy
  - Death and dying
  - Evolution
  - Group differences
  - Humor, irony, satire\
  - Images for international population
  - Inadvertent references
  - Luxuries
  - Personal questions
  - Religion
  - Sexual behavior
  - Slavery
  - Societal roles
  - Stereotypes
  - Substance abuse
  - Suicide
  - Violence

WHAT IS THE ISSUE?

Q. What is grammatically incorrect with the sentence below?

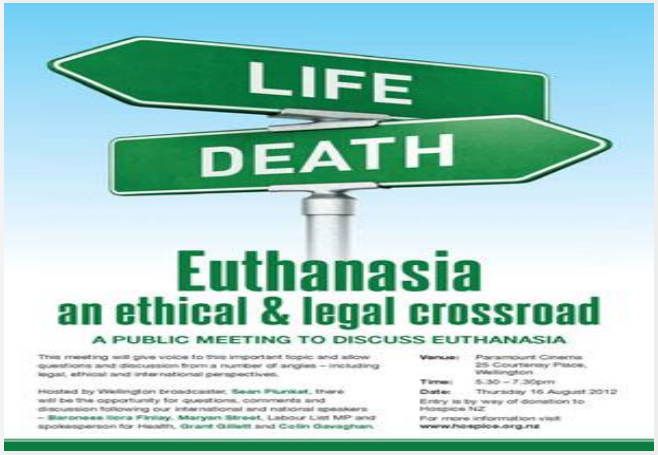
The US do not has the resource to protect Hawaii from the attacks on Pearl Harbor similar to the 9/11 attacks.

Potential fairness issues:
1. References to 9/11 and Pearl Harbor may elicit negative emotions.

2. International exam - US centric

Topics Best to Avoid

- Abortion
- Abuse of people (especially children) or animals
- Atrocities or genocide
- Contraception
- Euthanasia
- Experimentation on human beings or animals that is painful or harmful
- Hunting or trapping for sport
- Rape
- Satanism
- Torture
- Witchcraft



Use Appropriate Terminology for Groups

- African American people
- Asian American
- Hispanic American people
- Native American
- People who are bisexual, gay, lesbian, or transgendered
- People with disabilities
- Older people
- Below poverty line
- Gender

GUIDELINE 3. AVOID PHYSICAL SOURCES OF CONSTRUCT-IRRELEVANT VARIANCE

Avoid unnecessary physical barriers in items or stimulus materials

- Physical Barrier Examples
  - Irrelevant charts, maps, and other visual stimuli
  - Decorative charts
  - Visual stimuli in middle of paragraphs
  - Fonts that are hard to read
  - Non-english alphabet
  - Letters that look alike

What is the issue?

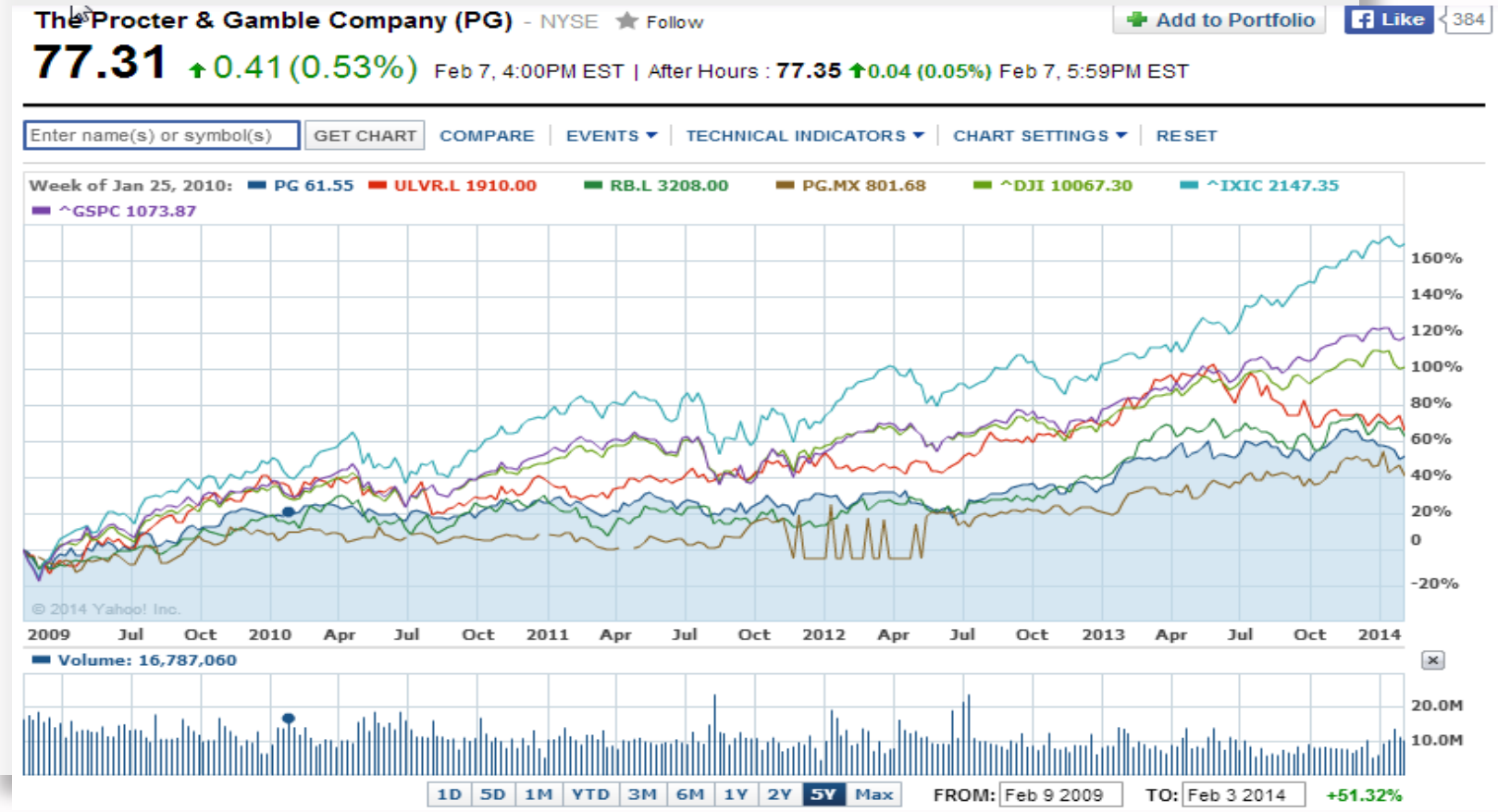
In Mr. T and Elmo's handbook of raising dogs,

which breed is considered to be one of the shyest of all dog breeds?

Picture in middle of question.

What is the Issue?

- What is the issue?
  - *Below is ta chart of P&G stock what is the cost of the stock July 2014?*



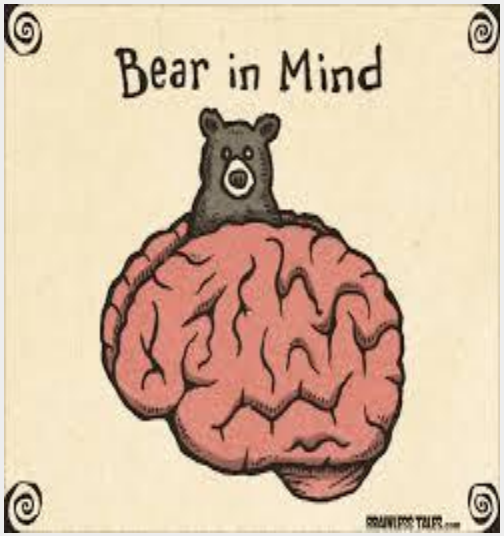
Other Fairness Considerations

- Selection of subject matter experts (sampling)
- Selection and execution of test development and psychometric methods and activities throughout the test development lifecycle
- Minimizing external influences on testing process (e.g., increase passing rates to minimize customer service complaints)
- Statistical methods - differential item functioning (DIF)
  - Uncovers bias towards one group
    - Classical example
      - Gender and sports item questions

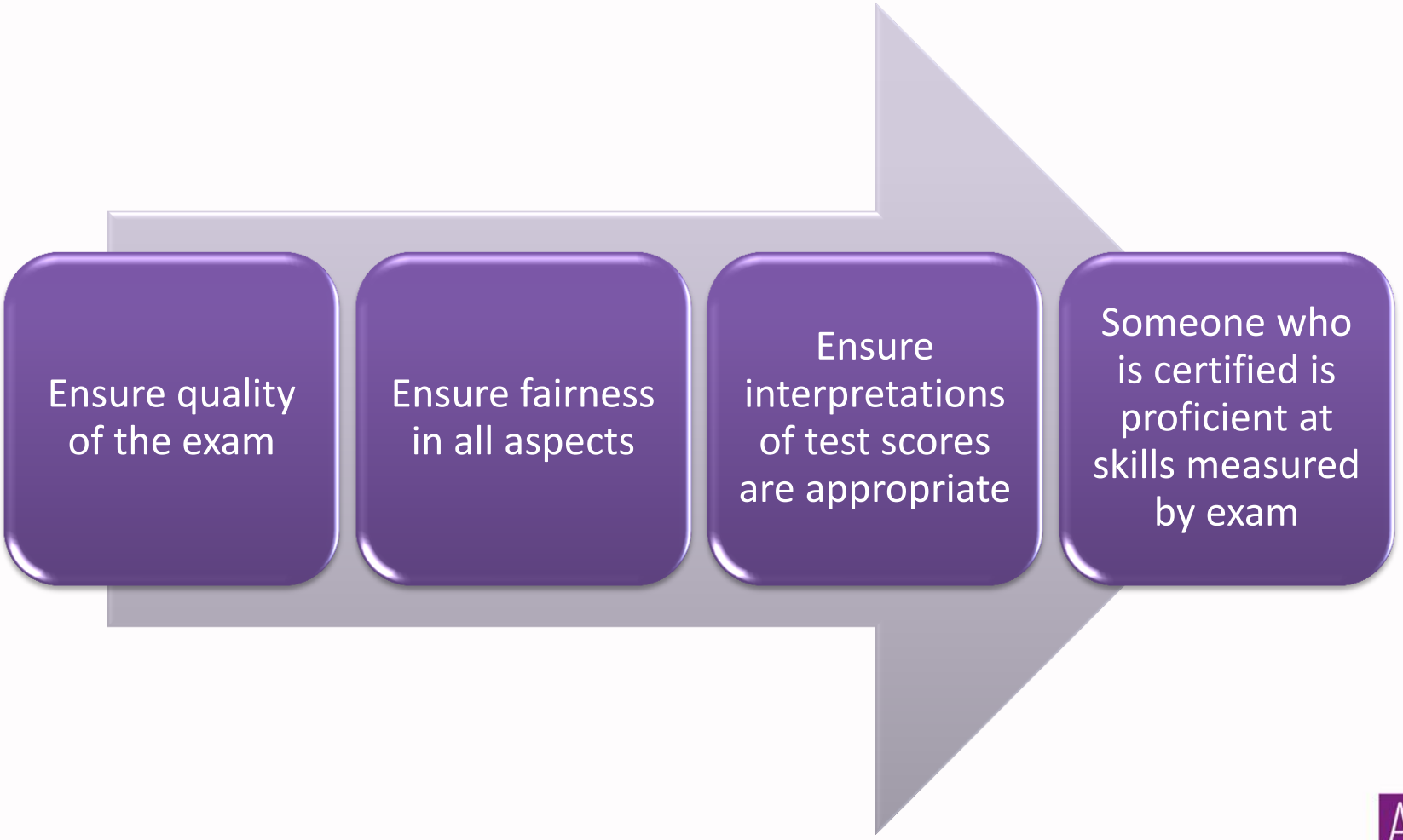
Summary

"…any characteristics of items that affect test scores and are unrelated to what is being measured is unfair"

p. 25 Haladyna and Rodriguez (2013)



Why is Psychometrics Important?



Questions

- Liberty
  - Liberty.Munson@microsoft.com
- Ada
  - awoo@ncsbn.org
- Manny
  - manny@intlcred.com

Microsoft

NCSBN
National Council of State Boards of Nursing

International Credentialing Associates

Innovations In Testing.org

ATP Innovations in Testing 2014 Scottsdale, AZ

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.