2014 ICE EXCHANGE

Fresh Ideas for the CREDENTIALING Community

November 11-14, 2014 | San Antonio Marriott Rivercenter | San Antonio, TX

Psychometric Rules of Thumb that Every Credentialing Manager Should Know

Dr. Liberty Munson
Dr. Ada Woo
Dr. Andre De Champlain
Dr. Manny Straehle

Institute for Credentialing Excellence

Ask a Psychometrician?

1. What have been your experiences with psychometricians? Positive? Negative? Neutral? In what ways?
2. What questions have you asked psychometricians that have **not** been answered to your satisfaction?
3. Have you received conflicting information from psychometricians? What was your question? What was the conflicting information?
4. What don't you understand about psychometrics that you wish you did?

2014 ICE EXCHANGE

Fresh Ideas for the CREDENTIALING Community

Overview

Andre

Basic Terminology

Validity

Ada

Reliability

Number of Items per Form

Liberty

Item Analysis

Item Exposure

Manny

Intro

Sample Size

Test Fairness

Disclaimer

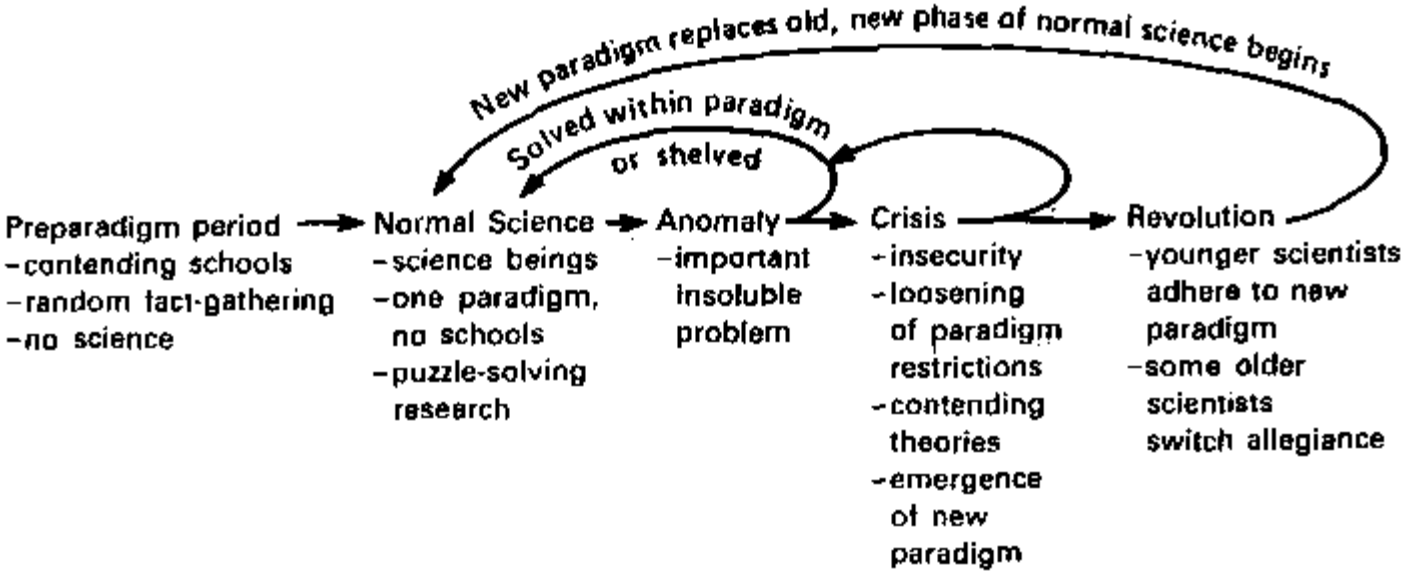
- These rules of thumb can be considered as provisional guidelines. For several of these, we provide supporting references (see handout).

- 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

Some Basic Terminology

- **Measurement**
  - Process by which we assign a number (the test score) to candidates in a systematic fashion to represent properties of these individuals
  - E.g.: Assigning a score of “85%” to my math high-school graduation exam performance presumably represents my overall knowledge of the domains that are targeted by the examination
- **Psychometrics**
  - Branch of applied statistics that attempts to describe, categorize, and evaluate the quality of measurements, improve the usefulness, accuracy, and meaningfulness of measurements, and propose methods for developing new and better measurement instruments

Validity

Validity – What It Is

"Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment". (Messick, 1989)

- **E.G.**: The admissions dean at your medical school has asked you to develop a MCQ exam that will be used to admit students to your undergraduate program
- **Evaluative judgment**
  - Score on the admissions' exam is a good predictor of MD Year 1 GPA
- **Empirical evidence**
  - High correlation between exam scores and MD Year 1 GPA

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

Validity – Professional Standards

- **Standards for Educational & Psychological Testing (2014)**
  - "A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation" (Standard 1.2)

  - "If validity for some common or likely interpretation has not been investigated, of if the interpretation is inconsistent with available evidence, the fact should be made clear and potential users should be cautioned about making unsupported interpretations" (Standard 1.3)

Important Steps to Follow

- **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

Argument-based Approach to Validation (Kane, 1992)

- **Score-based interpretation is posited as an interpretive argument that describes the model which links (1) the test scores to the (score-based) inferences and (2) the score-based inferences to any decisions that are based on the latter conclusions**
  - **Intent is to make the argument as clear as possible**
  - **Focus on the weakest part of the argument (hypothetico-deductive model)**

1. State the interpretive argument as clearly as possible
2. Assemble evidence relevant to the interpretive argument
3. Evaluate the weakest part(s) of the interpretive argument
4. Restate the interpretive argument and repeat

Argument-based Approach to Validation

- **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**: Can the performances be explained theoretically?
5. **Decision making**: Can the performances be used for placement decisions?

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

Argument-based Approach to Validation

- **Evaluation argument**
  - The scoring rule is appropriate
  - The scoring rule is applied accurately and consistently
  - **Evidence**
    - Clearly documented scoring rules and processes
- **Generalization argument**
  - The sample of items/cases in the exam is representative of the domain (universe of items/cases)
  - **Evidence**
    - Practice analysis/blueprinting effort
    - Generalizability analyses

A Central Validity Argument with Most Exams

- Does the performance (score) on the sample of items included in any examination, as reflected by our test blueprint, correspond to my true competency level in those domains?

- How accurately does my score on a restricted sample of MCQs, direct observations, etc. correspond to what I would have obtained on a much larger collection of tasks?

A Central Validity Argument with Most Exams

- One way to assure that our judgments are as accurate as possible is to develop a blueprint via a practice analysis that dictates as clearly as possible what should appear as part of the MCC's Qualifying Examination decision

- **Practice analysis**: A study conducted to determine the frequency and criticality of the tasks performed in practice

- **Blueprint**: A plan which outlines the areas (domains) to be assessed in the exam (with weightings)

Argument-based Approach to Validation

- **Extrapolation argument**
  - The universe score is related to the target score
  - There are no systematic errors that are likely to undermine the extrapolation
  - **Evidence**
    - Analysis of relationship between performance on exam (sample) and on a broader criterion (e.g.: workplace assessment)
    - Convergence validity

Argument-based Approach to Validation

- **Explanation argument**
  - Scores on the exam can be explained as a function of the skills/constructs hypothesized to underlie performance
  - **Evidence**
    - Confirmatory factor analysis of data set
    - Mapping of expert judgments of content on examination

Argument-based Approach to Validation

- **Decision making argument**
  - Candidates with a low skill level are not likely to pass the exam
  - Candidates with a high skill level are likely to pass the exam

  - **Evidence**
    - Standard setting internal & external validity evidence
    - **Internal validity**
      - Documentation of process followed
      - Inter-judge reliability, generalizability analyses, etc.
    - **External validity**
      - Relationship of performance on exam to other criteria

Validity – In Summary

- **What is validation?**
  - Gathering evidence (empirical and other) to substantiate claims (arguments) that we would like to be able to make based on examination scores
    - Candidates who score higher on my high-school math graduation exam have better math skills
    - Candidates who do well on the SAT will have a higher 1st year undergraduate GPA

Validity – In Summary

- **What are critical steps in validating claims?**
  - Clearly lay out the claim/interpretive argument that you'd like to make based on the candidate test scores
  - "Challenge" the interpretive argument
    - Is it clear and coherent?
    - Is it plausible given the empirical evidence at hand?
  - State the proposed interpretation

- **<u>Don't claim more than what is supported by evidence</u>**
  - Avoid unsubstantiated "blanket statements"
    - "My test shows construct validity"
    - "My exam has face validity", etc.
  - **<u>These statements are devoid of meaning</u>**

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?

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 References

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

Item Analysis

Basic Item Level Psychometric Analyses

Classical Test Theory Statistics: Fundamental Item Analysis

P-value

- Proportion of candidates answering correctly
- Ranges from 0 – 1 (when dichotomously scored) Polytomously scored: Average score and ranges from min to max score possible
- Low values = difficult
- High values = easy

Item Discrimination

- Discrimination index (D)
- Biserial/Point-biserial correlation coefficients
- To what extent does an item "discriminate" between candidates of low and high ability levels?

Distractor Analysis

- Evaluating effectiveness of distractors
- Quintiles, quartiles
- P-values and PBSE for each answer choice

How "Difficult" Should Items Be?

General rules of thumb:

Target: 0.3-0.7

- Maximizes information about differences between candidates

Item *p*-value of .5 provides max information

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

- No information provided about candidate
- Use only if needed for content validity reasons

Select items that maximize information near the cut-score

- Use IRT

Item Discrimination

Do Items Differentiate High and Low Performers?

Why?	Candidates who are more proficient should correctly answer items in a higher proportion than those who are less proficient
What?	An index indicating the degree of relationship between item score and a criterion score (e.g. total exam or section score)
Interpretation Overview	• 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

1. 1.aaaC1_	option	p-value	correlation	avg. time	0 to 42	43 to 48	49 to 51	52 to 53	54 to 60
	A	0.028	-0.199	137	17	3	3	2	3
	B	0.331	-0.065	67	90	90	80	45	20
>	C	0.472	0.160	96	96	103	101	74	90
	D	0.164	-0.004	59	30	48	46	31	6
	NULL	0.005	-0.210	56	5				

1. 1.EAaac_	option	p-value	correlation	avg. time	0 to 42	43 to 48	49 to 51	52 to 53	54 to 60
	A	0.010	-0.084	59	5	3	3		
>	B	0.901	0.491	39	128	190	219	194	248
	C	0.012	-0.130	89	7	4	1	1	
	D	0.013	-0.251	75	10	2	1		1
	E	0.009	-0.194	78	7	3			
	F	0.012	-0.236	87	11	1		1	
	G	0.012	-0.154	63	8	3	1		1
	H	0.001	-0.103	80	1				
	I	0.002	-0.050	1821	1			1	
	J	0.010	-0.088	83	5	3	1		2
	K	0.002	-0.123	84	2				
	L	0.016	-0.121	76	9	7		1	

1. 1.EAaad_	option	p-value	correlation	avg. time	0 to 42	43 to 48	49 to 51	52 to 53	54 to 60
	A	0.012	-0.120	81	8	3	1		
	B	0.044	-0.199	93	29	7	6		1
	C	0.018	-0.303	47	16	2			
	D	0.011	-0.175	53	9	1	1		
	E	0.002	-0.032	67	1	1			
	F	0.074	-0.251	73	44	16	7	4	2
>	G	0.762	0.552	43	95	190	207	145	112
	I	0.003	-0.066	89	2			1	
	J	0.060	-0.097	78	23	23	7	2	4
	L	0.008	-0.088	56	6	1	1		
	NULL	0.005	-0.287	0	5				

Psychometrically Sound Items

tem ID	n	P-value	Point biserial	Item Reliability
1.1.aaa	140	0.68	0.63	0.29
1.1.aab	122	0.84	0.29	0.11
1.1.aac	140	0.86	0.50	0.18
1.1.aad	122	0.88	0.42	0.14
1.1.aae	140	0.53	0.38	0.19
1.1.aaf	122	0.33	0.37	0.17
1.1.aag	140	0.62	0.41	0.20
1.2.aba	122	0.80	0.62	0.25

Item ID	Response	Count	p-value	Discrim					
1.1.aaa					15-36	37-49	50-57	58-64	65-82
	A	11	0.085	-0.337	8	2	1	0	0
	B	6	0.046	-0.312	5	1	0	0	0
	C	25	0.192	-0.340	9	8	6	2	0
>	D	88	0.677	0.627	6	12	17	21	32
1.1.aab					15-36	37-49	50-57	58-64	65-82
	A	10	0.086	-0.095	2	3	3	2	0
	B	1	0.009	-0.186	1	0	0	0	0
>	C	98	0.845	0.299	16	21	20	23	18
	D	7	0.060	-0.271	3	2	2	0	0
1.1.aac					15-36	37-49	50-57	58-64	65-82
	A	3	0.023	0.033	0	1	1	1	0
>	B	112	0.862	0.467	15	20	23	22	32
	C	8	0.062	-0.327	6	2	0	0	0
	D	7	0.054	-0.388	7	0	0	0	0
1.1.aad					15-36	37-49	50-57	58-64	65-82
	<<NULL>>	1	0.009	-0.174	1	0	0	0	0
>	A	102	0.879	0.408	14	21	25	25	17
	B	2	0.017	-0.118	1	1	0	0	0
	C	3	0.026	-0.293	3	0	0	0	0
	D	8	0.069	-0.218	3	4	0	0	1
1.1.aae					15-36	37-49	50-57	58-64	65-82
	<<NULL>>	1	0.008	-0.162	1	0	0	0	0
	A	15	0.115	-0.121	7	2	1	2	3
	B	36	0.277	-0.196	11	8	4	8	5
>	C	70	0.538	0.380	5	12	17	12	24
	D	8	0.062	-0.204	4	1	2	1	0

An Example of a Psychometrically Poor Item

You are creating an ASP.NET a… for an online store that sells movies on videocassette. Each user is assigned a profile based on … 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 Applicati
    ex));
\}
```

B.
```
catch (FileNotFoundExc
{
  throw (new ApplicationExceptio
    ex));
\}
```

C.
```
catch (ApplicationException ex)
{
  throw (new ApplicationExceptio
    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

What is the correct answer?

You have a computer that runs Windows XP Professional. You create a new partition and install Windows 2000 Professional on the new partition. You discover that you can no longer start Windows XP Professional. You verify that Windows 2000 starts successfully.

You need to ensure that you can start both operating systems.

What should you do?

A. Start the computer from the Recovery Console and run **fixmbr**.

B. Start the computer from a MS-DOS-based startup disk and run **sys c:**.

C. From the Windows XP Professional installation CD, copy the ntldr file to the active partition on the computer.

D. From the Windows XP Professional installation CD, copy the ntdetect.com file to the active partition on the computer.

Does this question perform psychometrically?

How Does This Item Perform Psychometrically?

You have a computer that runs Windows XP Professional. You create a new partition and install Windows 2000 Professional on the new partition. You discover that you can no longer start Windows XP Professional. You verify that Windows 2000 starts successfully.

You need to ensure that you can start both operating systems.

What should you do?

A. Start the computer from the Recovery Console and run **fixmbr**.

B. Start the computer from a MS-DOS-based startup disk and run **sys c:**.

C. From the Windows XP Professional installation CD, copy the ntldr file to the active partition on the computer.

D. From the Windows XP Professional installation CD, copy the ntdetect.com file to the active partition on the computer.

	p-value	correlation	avg. time	0 to 40	41 to 48	49 to 51	52 to 55	56 to 65
A	0.360	-0.147	86	19	23	10	19	18
B	0.081	0.374	128	6	4	5	5	
>C	0.364	-0.092	71	15	20	16	18	21
D	0.190	-0.236	115	11	15	5	11	5
NULL	0.004	-0.311	0	1				

Item Exposure

How Concerned Should You Be About Item Exposure?

Considerations:

Item Inventory
- Number of times it has been used
- Item format
- Number of items on exam

Resources required?
- How long does it take to create items?
- How much does it cost?

Publishing cycle
- Rotate forms
- Retire/temporarily stopping using items that have been used X times

Exam Purpose & Type
- High or low stakes?

Delivery Regions
- China, Pakistan, India, Turkey

Test administration method
- Paper & Pencil
- Computerized
- Online

Candidate Volume

Mitigating Over Exposure

Item cloning

Single-use items

Web patrol monitoring

Set item exposure parameters

Publishing Strategies

Item Drift

What?

Item parameter estimates change between test administrations

Why?

Parameter change can be a result of:

- Changes in educational instructions
- Changes in practice
- Item compromise (test security breach)

Action:

Conduct periodical statistical monitoring and consult subject matter experts

Sample Size

CHOCOLATE DEMONSTRATION

How Many SMEs?

Does the sample represent the population?

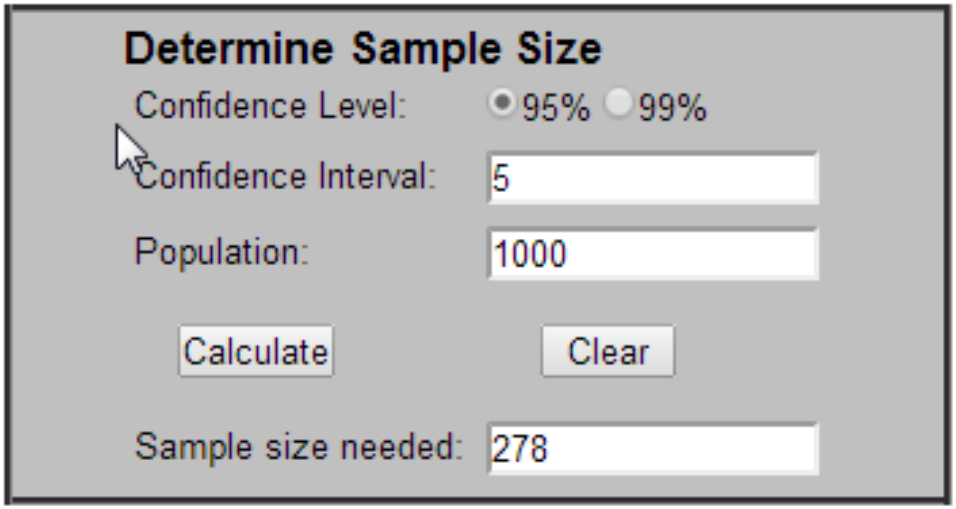
- Sample of Sommeliers (n=100)
  - Experience
  - Education/Training
  - Industry
  - Geography
  - Gender
  - Ethnicity
- Population of Sommeliers (N=1000)
  - Experience
  - Education/Training
  - Industry
  - Geography
  - Gender
  - Ethnicity

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

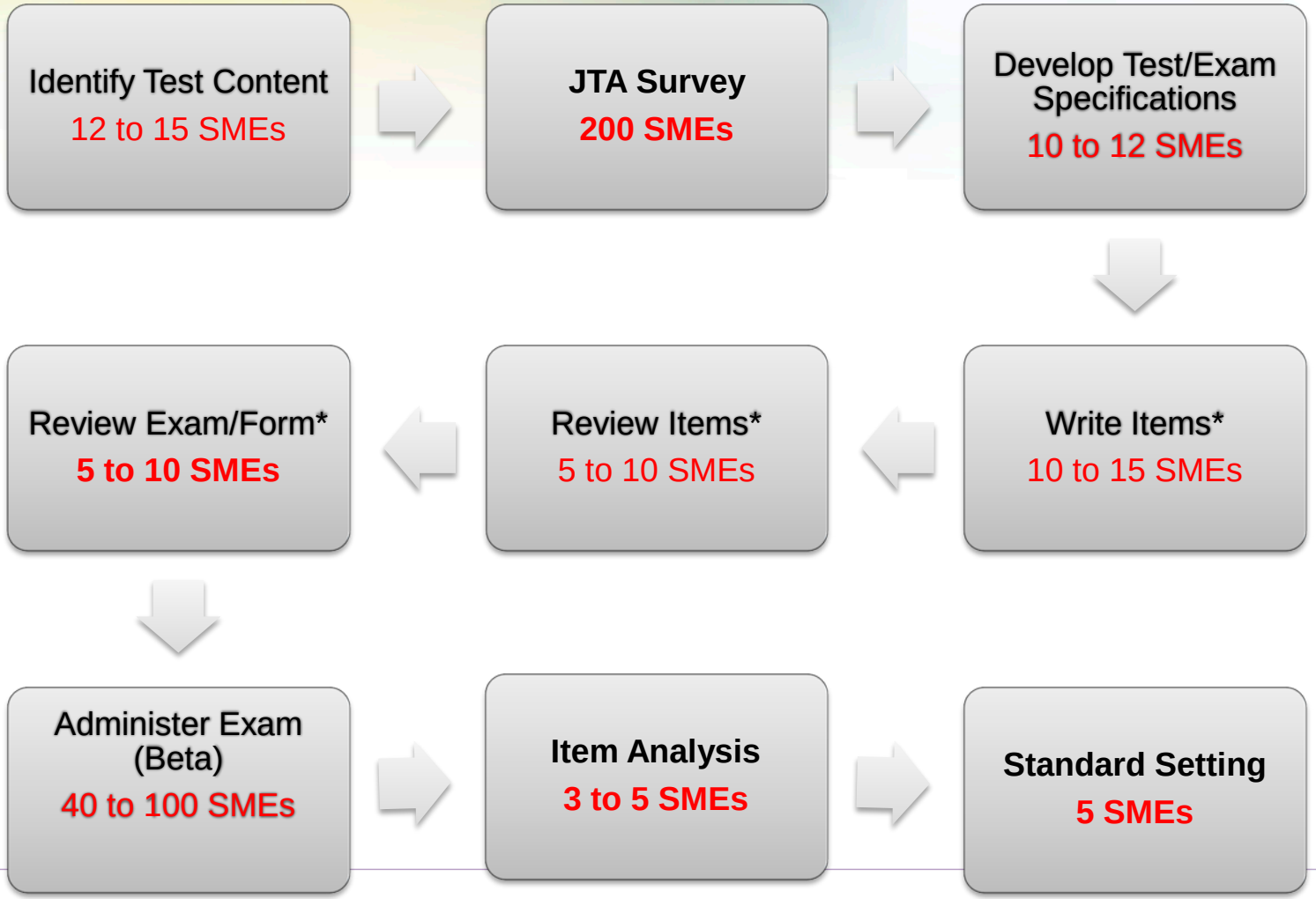
Sampling Exercise

- Total of 1000 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



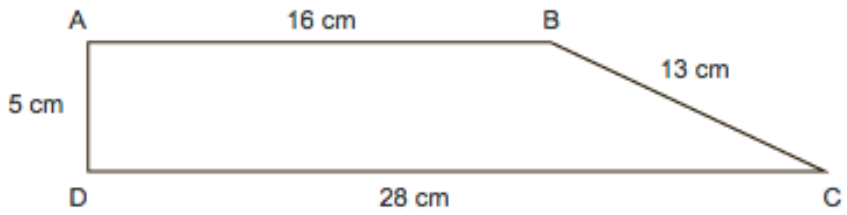
Test Fairness

Example #1 – Too Hard and Wrong

Question is too difficult for 5th graders

The Problem

Trapezoid ABCD is shown below.



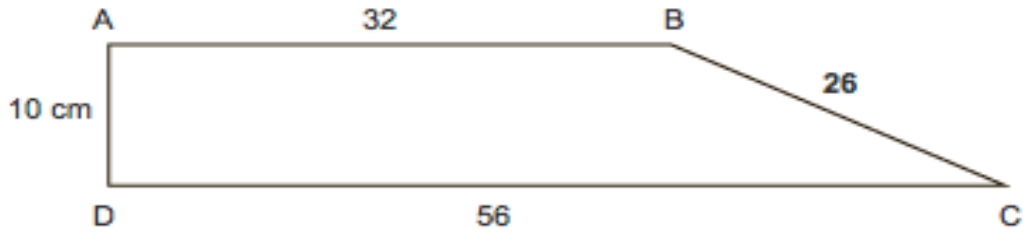
A new trapezoid is formed by doubling the lengths of sides AB and CD. Find the perimeter, in centimeters, of the new trapezoid.

Show your work.

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

Q1. 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 constru irrelevant knowledge/skills
- Avoid material that is unneces controversial, inflammatory, offensive, or upsetting
- Use of appropriate terminolog refer to people
- Avoid stereotypes
- Represent diversity in depictic people

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?

T…or 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?

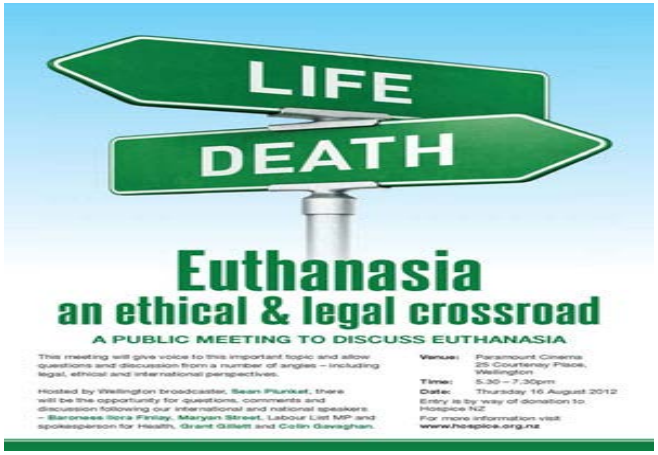
Q. What is grammatically incorrect with the sentence below?

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,

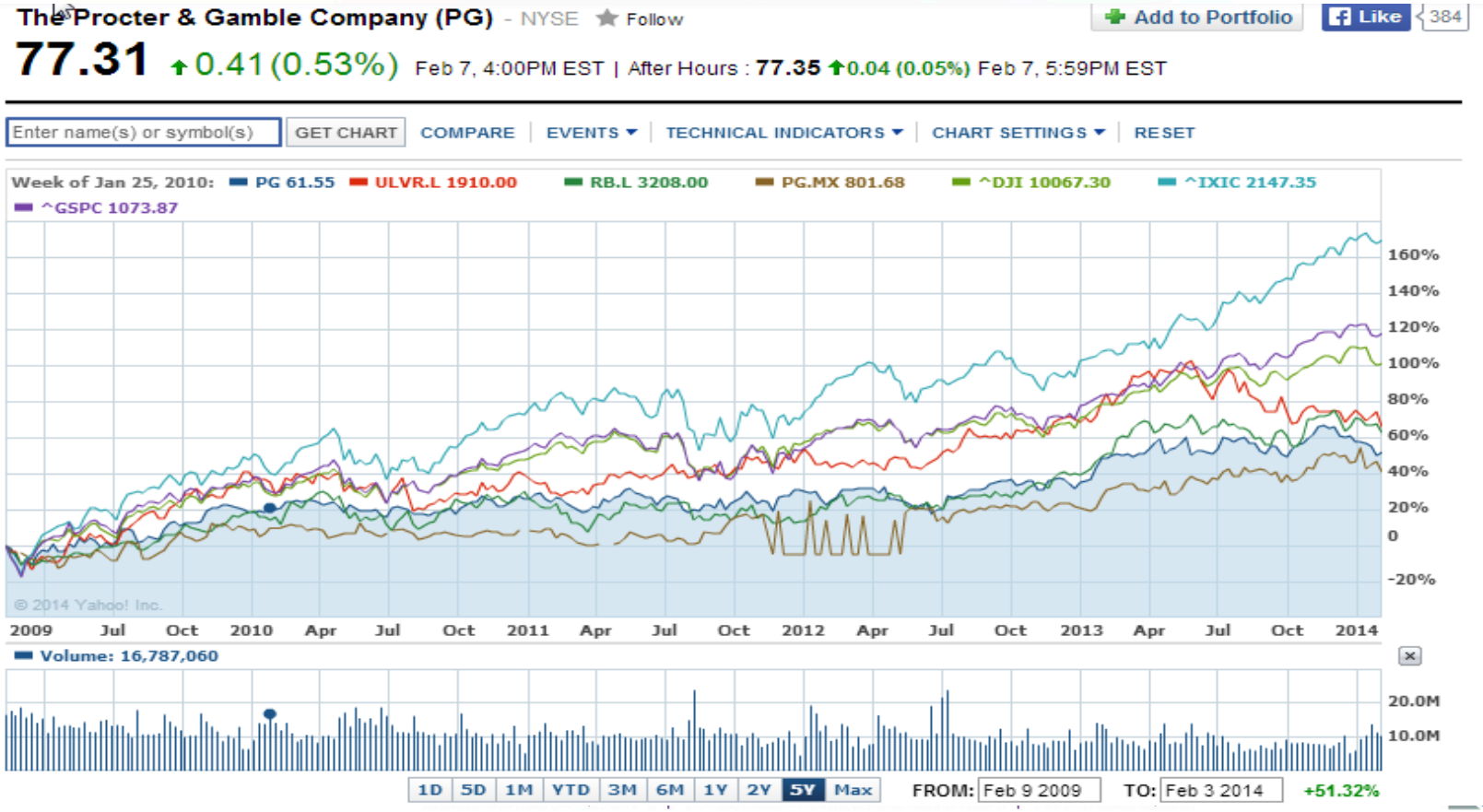
which breed is considered to be one of the shyest of all dog breeds?

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?*



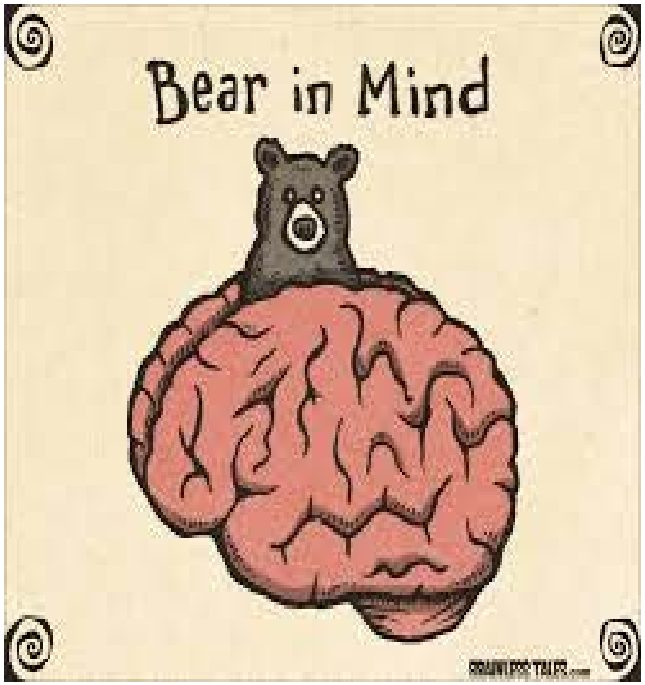
Other Fairness Considerations

- 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

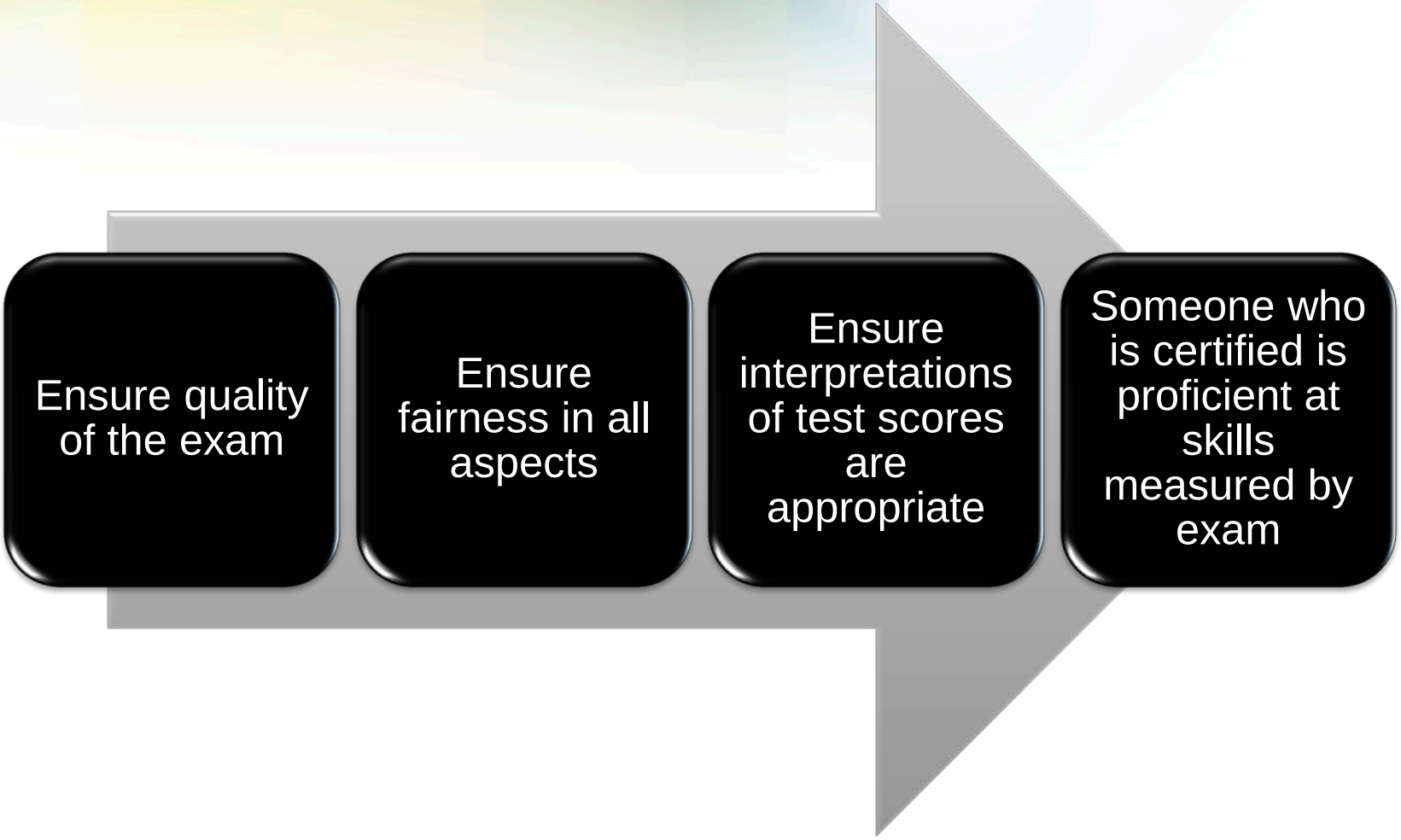
Summary

"…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

Andre

adechamplain@mcc.ca

Medical Council of Canada

Ada

awoo@ncsbn.org

NCSBN
National Council of State Boards of Nursing

Liberty

Liberty.Munson@microsoft.com

Microsoft

Manny

manny@intlcred.com

International Credentialing Associates

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.