

Psychometric Tips that Every Credentialing Manager Should Know

AUSTIN FOSSEY

PSYCHOMETRICIAN AND ANALYTICS MANAGER
QUESTIONMARK CORP.

DIRECTOR OF PSYCHOMETRICS AND RESEARCH
ASSESSMENT, EDUCATION, AND RESEARCH EXPERTS
(AERE)

DR. MANNY STRAEHLE

PRESIDENT AND FOUNDER
ASSESSMENT, EDUCATION, AND RESEARCH EXPERTS
(AERE)

Thank you!

Liberty Munson

Andre DeChamplain

Ada Woo



Overview

Austin

Validity

Reliability

Item Analysis

Item Exposure

Manny

Intro

Basic Terminology

Sample Size

Test Fairness

Disclaimer

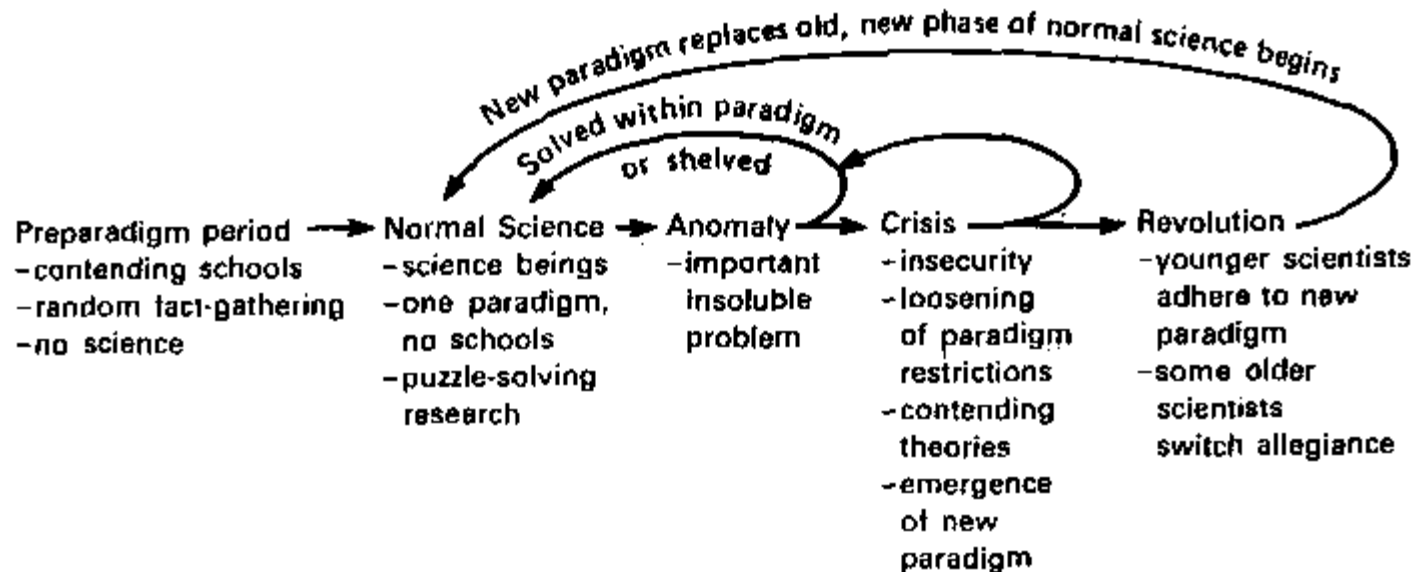
These rules of thumb can be considered as provisional guidelines.

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 yielding a measured result (e.g., scores).

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

Examples

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

Defining Validity

Evidence, Evidence, and More Evidence

“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)

“Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA, APA, & NCME, 2014).

Validity: An Ongoing Debate

Validity is a continually evolving concept.

Disagreements about what is important and what needs to be validated.

Validity is no longer restricted to interpretations of test scores (Sireci, 2013)

Important Steps to Follow

Clearly lay out the intended use of the test (Bachman, 2005):

- Is the interpretation of the score relevant to the decision being made?
- Is the interpretation of the score useful for the decision being made?
- Are the intended consequences of the assessment beneficial for the stakeholders?
- Does the assessment provide sufficient information for making the decision?

Collect evidence about:

1. Test content
2. Response process
3. Internal structure
4. Relations to other variables
5. Testing consequences

Integrate into a *validity argument* (AERA, APA, & NCME, 2014)

Argument-Based Validity

Criterion, content, and construct validity are crucial aspects of assessment result validity, but how do we demonstrate the link to the interpretations and uses of the assessment results?

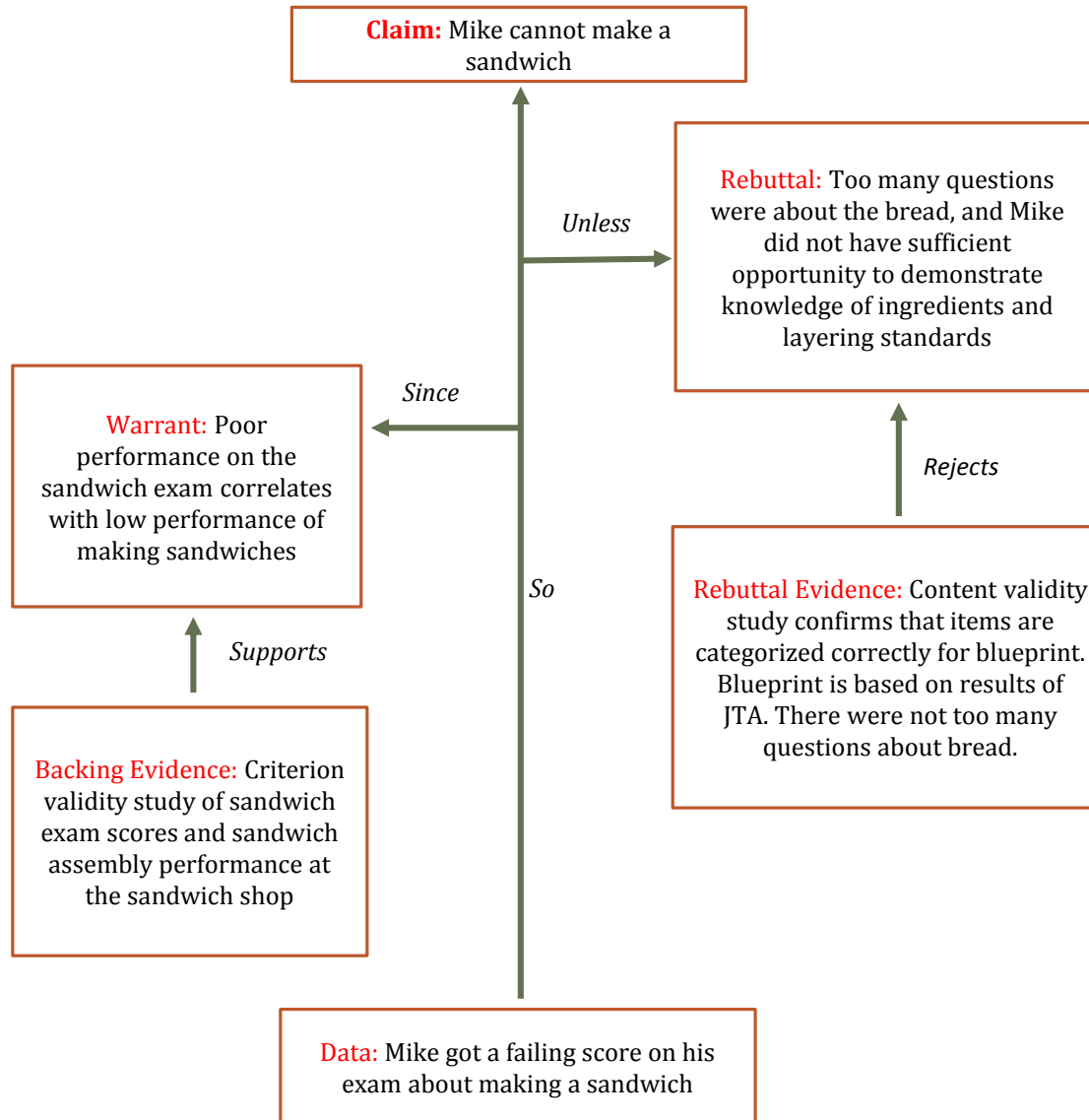
Argument-based validity (e.g., Kane, 1992) provides logic using Toulmin's structure of an argument to support claims about score-based interpretations.

Bachman (2005) expands this to include validity arguments for use cases.

Argument-Based Validity (Kane, 1992)

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

Example Toulmin Structure for Validity Interpretation



Five Types of Validity Arguments (Kane, 1992)

Evaluation: Can the results be used to evaluate the candidates' performance on the exam?

Generalization: Do the performances generalize to the domain of tasks?

Extrapolation: Do the performances generalize to other settings/performance formats?

Explanation: Can the performances be explained theoretically?

Decision making: Can the performances be used for placement decisions?

Examples of Warrants and Evidence

Argument	Example Warrant	Example Evidence
Evaluation	The scoring rule is appropriate and applied accurately and consistently.	Domain model Conceptual Assessment Framework Standard setting study
Generalization	The sample of items in the exam is representative of the domain of performance/behavior.	Domain analysis (e.g., JTA) Blueprint Content validity study
Extrapolation	Candidates' universal performance is related to their assessment performance.	Criterion validity study
Explanation	Performance on the exam can be explained as a function of the underlying skills/constructs.	Confirmatory factor analysis Content validity study Item development/item review documentation
Decision Making	Candidates who pass meet minimally acceptable performance levels for the domain.	Documentation of standard setting procedures Analysis of standard setting panel's ratings Criterion validity study

Validity Summary

Validation means gathering evidence to substantiate claims (arguments) about the assessment results.

- Clearly lay out the claim/interpretive argument based on candidates' results.
- “Challenge” the interpretive argument
 - Is it clear and coherent?
 - Is it plausible given the empirical evidence at hand?
 - Are there alternative explanations?
- State the proposed interpretation

Do not claim more than what is supported by evidence

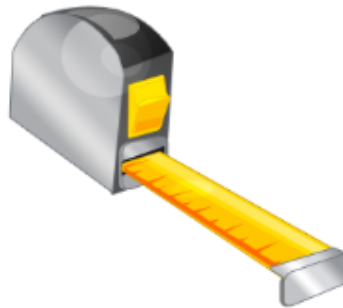
Reliability

Reliability

The reliability of a measure is its degree of **consistency** (Crocker & Algina, 2008).

Items on an exam are **mathematically related or clustering**.

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 Coefficients

Reliability is the ratio of true score variance to observed score variance, and it ranges from 0 to 1.

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}$$

- $\rho_{XX'}$ is the reliability coefficient. Proportion of observed score variance attributable to true score variance.
- σ_T^2 is the variance of the true scores
- σ_X^2 is the variance of the observed scores

Forms of Reliability

(Crocker & Algina, 2008)

Correlations Between Multiple Administrations

- Alternate Forms – coefficient of equivalence
- Test-Retest – coefficient of stability
- Alternate Forms and Test-Retest – coefficient of stability and equivalence

Single Administration Methods

- Split-Halves – Spearman Brown prophecy
- Item Covariances – Cronbach's Alpha (or KR 20, KR 21)

Participant Classification Methods

- Decision Consistency

Cronbach's Alpha (α)

Items are mathematically related

Measure of *internal consistency*

Theoretical lower bound of reliability, $\alpha \leq \rho_{XX'}$

Dependent on interrelatedness of the items and test length, assumes assessment is measuring a single construct (Crocker & Algina, 2008)

Interpretations:

- $\alpha \geq 0.9$ – Excellent (High-Stakes Assessment)
- $0.7 \leq \alpha < 0.9$ – Good (Low-Stake Assessment)
- $0.6 \leq \alpha < 0.7$ – Acceptable
- $0.5 \leq \alpha < 0.6$ – Poor
- $\alpha < 0.5$ – Unacceptable

Item Analysis

Classical Test Theory Item Statistics

P-Value

- Proportion of candidates answering correctly
- Ranges from 0 – 1 (when dichotomously scored)
- Low values = difficult, High values = easy

Item Discrimination

- Correlation coefficients between item and total scores
- To what extent does an item “discriminate” between candidates of low and high ability levels?

Distractor Analysis

- Evaluating effectiveness of distractors
- Relative performance groups
- P-Values of an item's options

Item Difficulty

Medium difficulty items ($p = 0.5$) usually discriminate best

Often target items with p between 0.3 and 0.7, though item discrimination is primary statistic for item selection

Only keep very easy or very hard items when required by blueprint / content validity

Item Discrimination

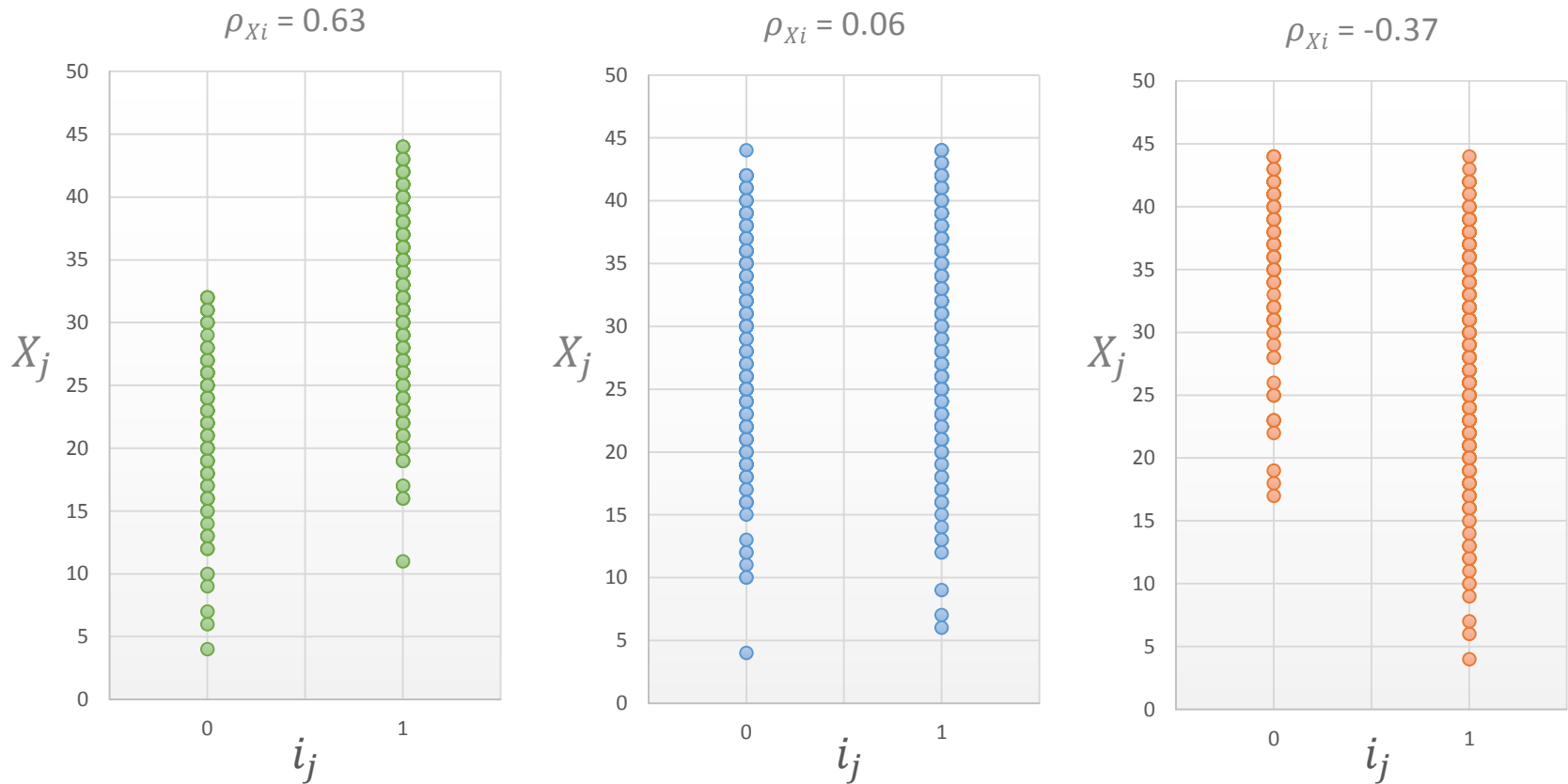
Candidates who are more proficient should correctly answer items in a higher proportion than those who are less proficient

Primary statistic for item selection in form building

Interpretation of values:

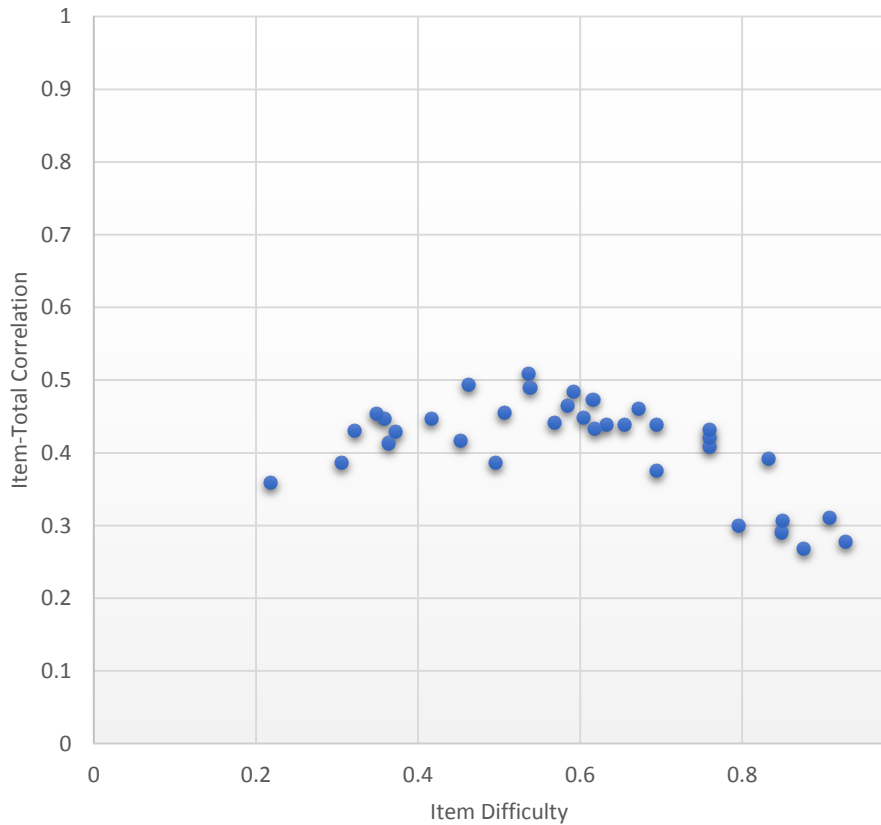
- Positive Values - High performers answering correctly, low performers are not.
- Values Near Zero - No relationship between item's score and test performance.
- Negative Values - Low performers answering correctly, high performers are not.

Visualizing Correlation Discrimination

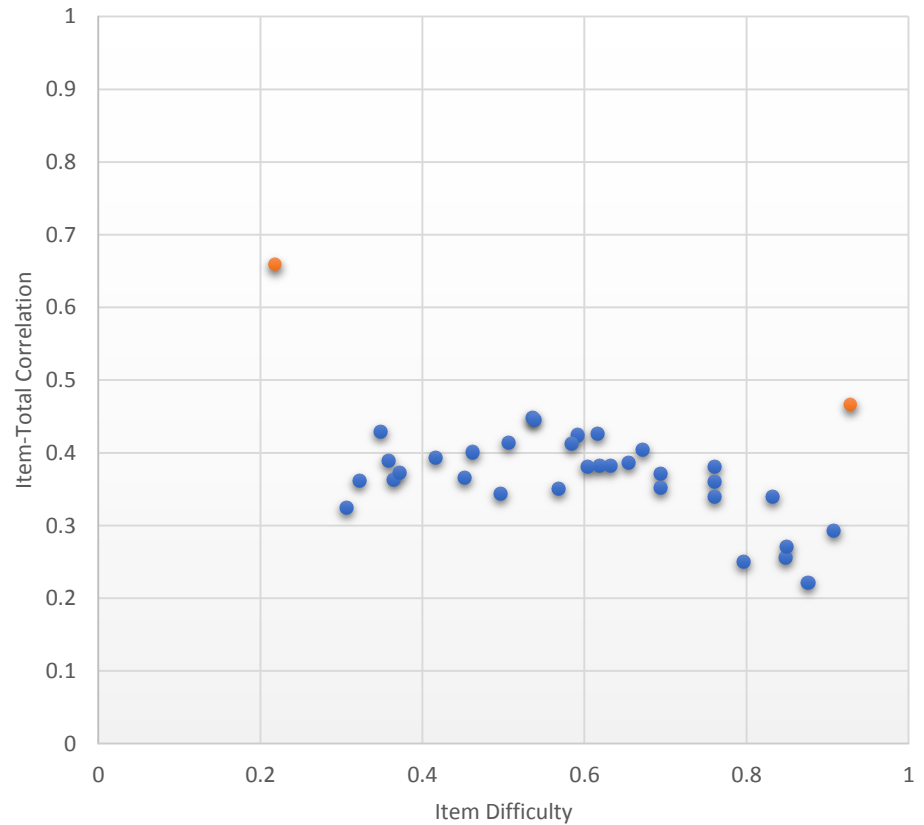


Discrimination and Difficulty

Item Difficulty by Discrimination:
Demo Assessment with Unweighted Scores



Item Difficulty by Discrimination:
Demo Assessment with Two Weighted Items



Two Discrimination Coefficients

Common to use *Item-Total Correlation* (Pearson correlation), but recall total score is partially made up of the item's score.

Use *Item-Rest Correlation* to correct for this dependency for these conditions (Crocker & Algina, 2008):

- Short forms (25 items or fewer)
- Weighted items
- Small sample sizes

Item difficulty p-value	◆ 0.694 (+/- 0.021)
Item-total correlation discrimination	≡ 0.358 (-0.04/+0.038)
Item-rest correlation discrimination	≡ 0.318 (-0.041/+0.04)

Distractor Analysis

Summarizes response distributions by option for participants based on their performance.

Identify miskeyed, items, overlapping options, response bias due to specialized knowledge.

Identify poor distractors and good distractors (common misconceptions).

Relative Performance Groups for Option Analysis

Helpful to see distributions of responses between high-performing and low-performing candidates.

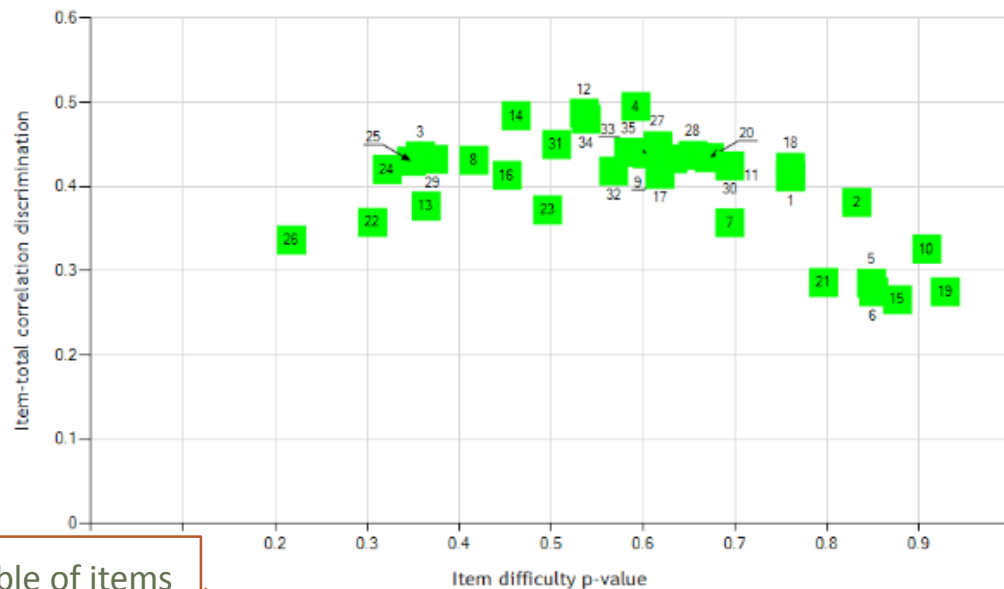
For this analysis, group candidates by scores (e.g., quartiles, quintiles).

Kelley (1939) recommended top and bottom 27% of scores:

Answer option information		Number and percentage of participants achieving scores			
Outcome #	Answer option	All	Upper 27%	Middle 46%	Lower 27%
1	A	53 (10.6%)	6 (4.4%)	26 (11.3%)	21 (15.6%)
2	B	50 (10%)	3 (2.2%)	18 (7.8%)	29 (21.5%)
✓ 3	C	347 (69.4%)	120 (88.9%)	165 (71.7%)	62 (45.9%)
4	D	50 (10%)	6 (4.4%)	21 (9.1%)	23 (17%)
5	No response	0 (N/A %)	0 (N/A %)	0 (N/A %)	0 (N/A %)
Total assessment mean score		63.1 %	82 %	64.2 %	42.5 %

☒ View item total scatter plot
 ☐ View item rest scatter plot

Item difficulty by item-total correlation discrimination



Sortable table of items

not plotted on this chart.

Item-level statistics

Select an item from the table below in order to view more item analysis details.

Presentation order	Question wording	Question description	Question type	Perception question id (Revision)	Topic	Item difficulty p-value	Item-total correlation discrimination	Item-rest correlation discrimination	Average Item score	Max possible score
26	Q26	Q26	Multiple Choice	0913660629624699 (1)	Demo Assessment - Form C	0.218	0.337	0.301	0.218	1
22	Q22	Q22	Multiple Choice	3660787948457636 (1)	Demo Assessment - Form C	0.306	0.358	0.318	0.306	1
24	Q24	Q24	Multiple Choice	2911667287558086 (1)	Demo Assessment - Form C	0.322	0.421	0.383	0.322	1
25	Q25	Q25	Multiple Choice	7553197457195108 (1)	Demo Assessment - Form C	0.348	0.431	0.392	0.348	1
3	Q3	Q3	Multiple Choice	8527303538053956 (1)	Demo Assessment - Form C	0.358	0.437	0.398	0.358	1

Item details

Question type	Multiple Choice
Question status	Normal
Question minimum possible score	0
Question maximum possible score	1
Number of participants presented the question	1000
Number of participants who responded to the question	1000
Item difficulty p-value	♦ 0.159 (+/- 0.012)
Item-total correlation discrimination	■ 0.1 (-0.031/+0.031)
Item-rest correlation discrimination	■ -0.046 (-0.032/+0.032)
High-Low discrimination	0.10
Item reliability	0.036
Perception question id	9743753565705035
Topic	Geology

Item Difficulty (p-value)

Item Discrimination

Distractor
Analysis

Answer option information		Number and percentage of participants achieving scores			
Outcome #	Answer option	All	Upper 27%	Middle 46%	Lower 27%
1	The Paleocene Epoch	304 (30.4%)	34 (12.6%)	145 (31.5%)	125 (46.3%)
✓ 2	The Pliocene Epoch	159 (15.9%)	63 (23.3%)	61 (13.3%)	35 (13%)
3	The Pleistocene Epoch	284 (28.4%)	130 (48.1%)	141 (30.7%)	13 (4.8%)
4	The Holocene Epoch	253 (25.3%)	43 (15.9%)	113 (24.6%)	97 (35.9%)
5	No response	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Total assessment mean score		46.8 %	66.5 %	48 %	25.2 %

Other Common Item-Level Stats

Classical Test Theory (CTT)

- Item Reliability
- Distractor discrimination

Crocker & Algina, 2008

Item Response Theory (IRT)

- Item parameters
- INFIT and OUTFIT
- Differential Item Functioning (DIF)

de Ayala, 2009

Item Exposure

Considerations

Item inventory

Test purpose

Candidate volume

Candidate population

Investment

Test delivery design

Publishing cycle



Mitigating Item Exposure

LIMITING OPPORTUNITIES

Multiple forms

Frequent republishing

Large or rotating item banks

Exposure limits for items or forms
(e.g., one time use)

Short test windows

IDENTIFYING BREACH

Channel for reporting breach

Web patrol monitoring

Item drift

Statistical models for detecting
item breach at candidate or class
level

Item Drift

IRT parameter estimates change over time/between administrations (de Ayala, 2009)

Drift is a result of changes in instruction, changes in practice, or item exposure

Schedule monitoring of item drift, consult with SMEs

Sample Size

Goal of Sampling

THE SAMPLE SHOULD REPRESENT THE POPULATION

Definitions

Sample = a small group of people from the larger group you are trying to investigate (e.g., US State such as **Ohio**)

Population = Larger group you are trying to investigate (e.g., United States)

Example

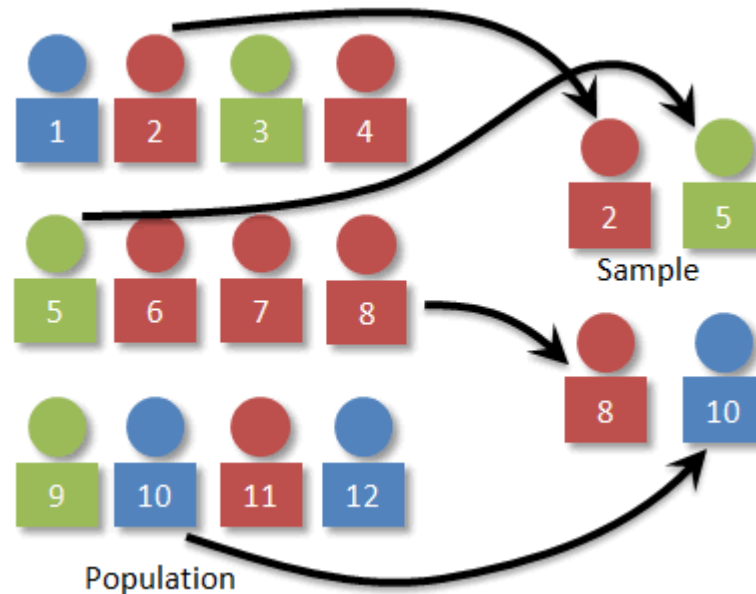
In presidential elections, Ohio tends to be a good predictor of presidential election outcomes because it represents the entire US population as a state representing key demographics and other factors.



Exercise: Is this Sample Representative by Color?

Population
N=12

Blue = 3 (25%)
Red = 6 (50%)
Green = 3 (25%)



Sample
N=4

Blue = 1/4 (25%)
Red = 2/4 (50%)
Green = 1/4 (25%)

How Many SMEs?

Sample of Sommeliers (n=100)

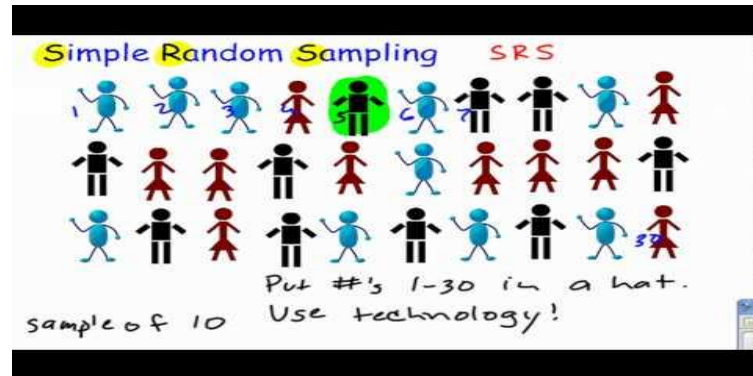
- Experience
- Education/Training
- Industry
- Geography
- Gender
- Ethnicity



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 sampled **100** 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%**

Determine Sample Size
Confidence Level: ☒ 95% ☐ 99%
Confidence Interval:
Population:

Sample size needed:

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
 - $\pm 3\%$ 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?

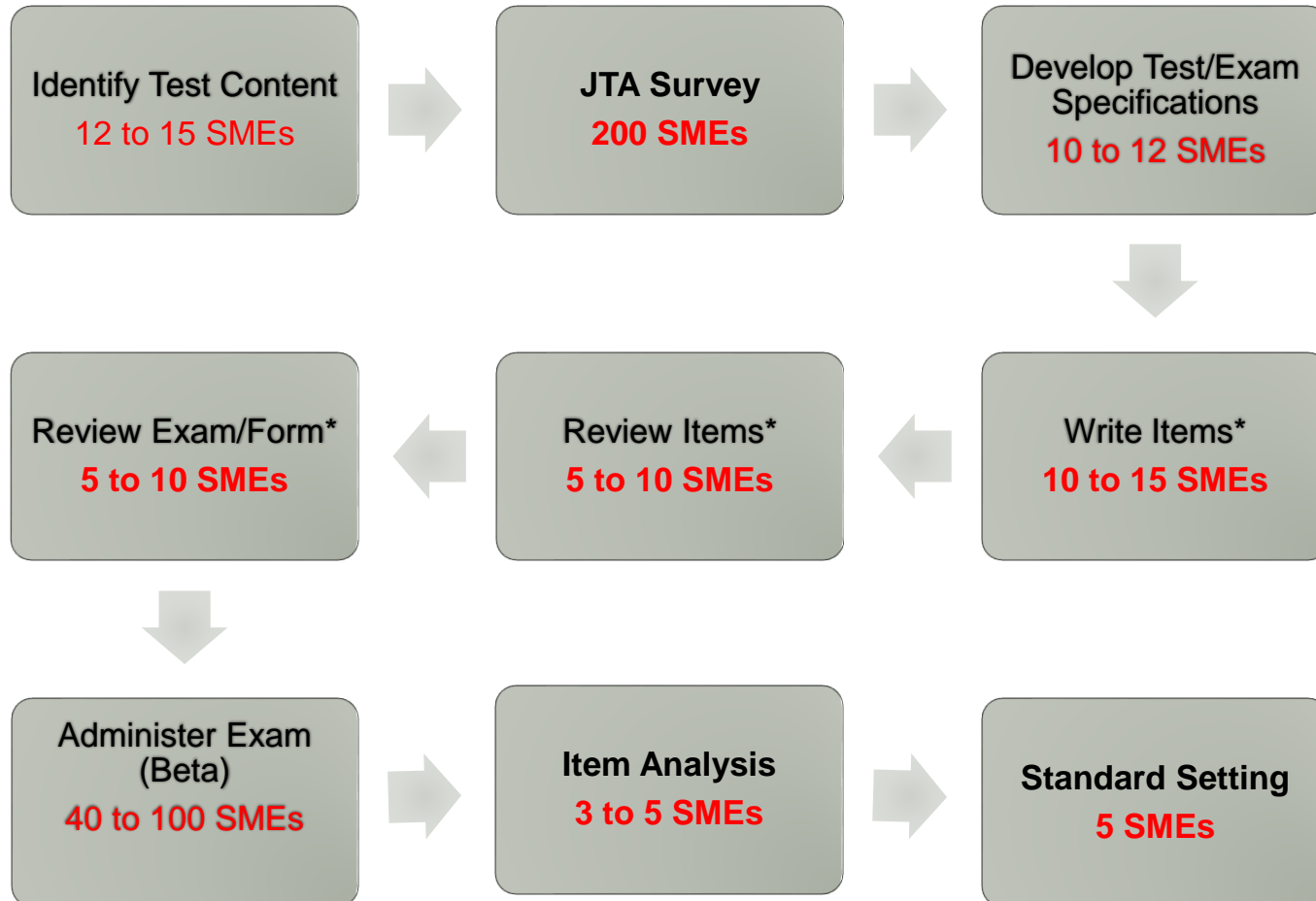
Confidence Interval is set at **95%**

Margin of Error **5%**

ANSWER: 278

Required Sample Size [†]								
Population Size	Confidence = 95%				Confidence = 99%			
	Margin of Error				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



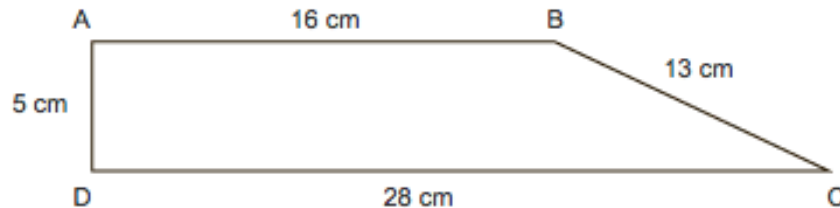
Test Fairness

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.

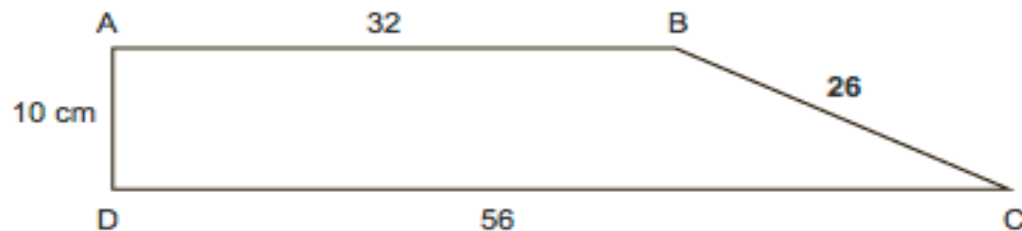
Show your work.

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.



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

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 Fairness?

Summary of Standard	Examples to Avoid
3.0 Minimize construct-irrelevant variance and promote valid score interpretations Construct-Irrelevant	Who was the best general? (the question is very vague hence creating construct irrelevancy)
3.1 Develop, revise, and administer exams to promote valid score interpretations Construct-Validity	An exam on mathematics that is suppose to equally cover algebra and geometry items but there are only geometry items
3.2 Measure intended construct and minimize construct irrelevant characteristics (e.g., linguistic, communicative, cognitive, cultural) Construct-Validity	A test on raising poultry for 3 rd graders in the US ask questions only in Taushiro .
3.3 Include relevant subgroups in studies (reliability, validity, ...) Subgroups	A science exam was developed only using data from white males only which did not reflect the population (Recall the sampling section)

What is Fairness?

Summary of Standard	Examples to Avoid
3.4 Comparable treatment during test admin and scoring process Equal Treatment	Poor resolution of screen shots in Test Center A but in Test Center B the resolution is very good
3.5 Provisions made to test administration and scoring procedures to remove construct irrelevant barriers for all subgroups	Test developers find differences between subgroups and don't specify the provisions made during the scoring or test administration
3.6 Evidence indicates test scores differ in meaning for relevant subgroups Interpretation	Army candidates score significantly higher than non-Army candidates and the testing organization does not examine and analyze the reason for this difference
3.7 Differential predictions should be used for criterion validity for score predictions Criterion Validity	Using all individuals in the study for predictions rather than doing a subanalysis across various demographic response categories.

Summary of Standard	Examples to Avoid
<p>3.8 Constructed responses should collect and report validity of score interpretations for relevant subgroups</p> <p>Rubric Development</p>	<p>A candidate is observing internet traffic of a potential hacker. Judges were not trained and calibrated on the rubric properly when evaluating this candidates ability to perform intrusion analysis.</p>
<p>3.9 Develop and provide accommodations.</p> <p>Accommodations</p>	<p>Not adding more time for an individual who has a documented learning disability</p>
<p>3.10 Document and monitor accommodations</p> <p>Accommodations</p>	<p>Test administrator fails to provide proper documentation to the candidate requiring those accommodations</p>
<p>3.11 Provide validity for any alteration of the test to make accommodations for the candidate</p> <p>Accommodations</p>	<p>An exam is shortened to accommodate for individuals who require a substantial amount of time to review test questions. The test developers did not provide any evidence to support the validity of this exam.</p>

Summary of Standard	Examples to Avoid
<p>3.12 Provide and document methods that support the validity of the translation and adaptation models used</p> <p>Translation</p>	<p>Translating an exam without examining any differences or using common methods such as DIF or professional judgements among SMEs</p>
<p>3.13 Administered in the language that is most appropriate and most relevant</p> <p>Translation</p>	<p>A test only asking it in Taushiro.</p>
<p>3.14 When using an interpreter, it should follow a standardized procedure for the interpreter. Interpreter should be qualified (fluent in the language and culture)</p> <p>Translation</p>	<p>Used an interpreter you located at the local Walmart and is unqualified.</p>
<p>3.15 Using a test for various subgroups requires additional information in various documents (e.g, test manuals) about the validity of these subgroups</p>	

Summary of Standard	Examples to Avoid
<p>3.16 Scores are affected by construct irrelevant characteristics, when legally permissible, these scores can be used if evidence of validity exists for the subgroups</p>	<p>Not offering an exam that was not designed for a specific disability, when the standard form was unavailable.</p>
<p>3.17 Reporting aggregate scores publicly requires providing evidence of comparability cautionary statements of interpretation</p> <p>Interpretation Guidance</p>	<p>Providing only mean scores across subgroups without any statement stating (or similar statement) that these findings may not be 100% conclusive</p>
<p>3.18 Diagnostic and placement testing should use other sources of information than a single test</p> <p>Multiple Assessments</p>	

Summary of Standard	Examples to Avoid
<p>3.19 Opportunity to learn the test content and that the educational materials are related to the test content</p> <p>Relate Education to Test Content</p>	
<p>3.20 When using multiple assessments, use evidence of subgroup differences in the mean scores or percentages of examinees whose scores exceed the cut scores, in deciding which test or score to use</p> <p>Subgroups</p>	

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

ETS Fairness Guidelines

(Zieky, 2003)

Guideline	Poor Examples
1. Treat People with Respect	
2. Minimize the effects of construct-irrelevant knowledge/skills	Calculate the mean of abortions occurring among Catholics on a national statistics tests
3. Avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting	Asking a question about 9/11 on a national writing exam
4. Avoid stereotypes	
5. Represent diversity in depictions of people	Using only one race, gender, age group, and etc.

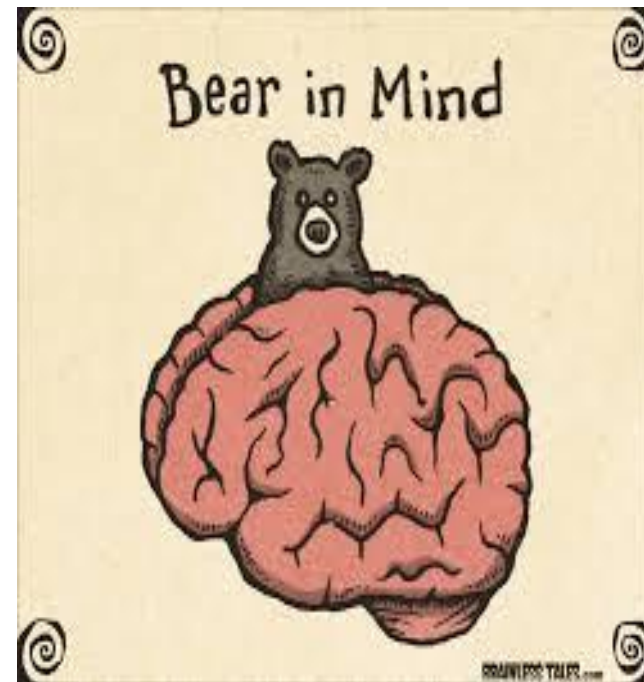
Other Fairness Considerations

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

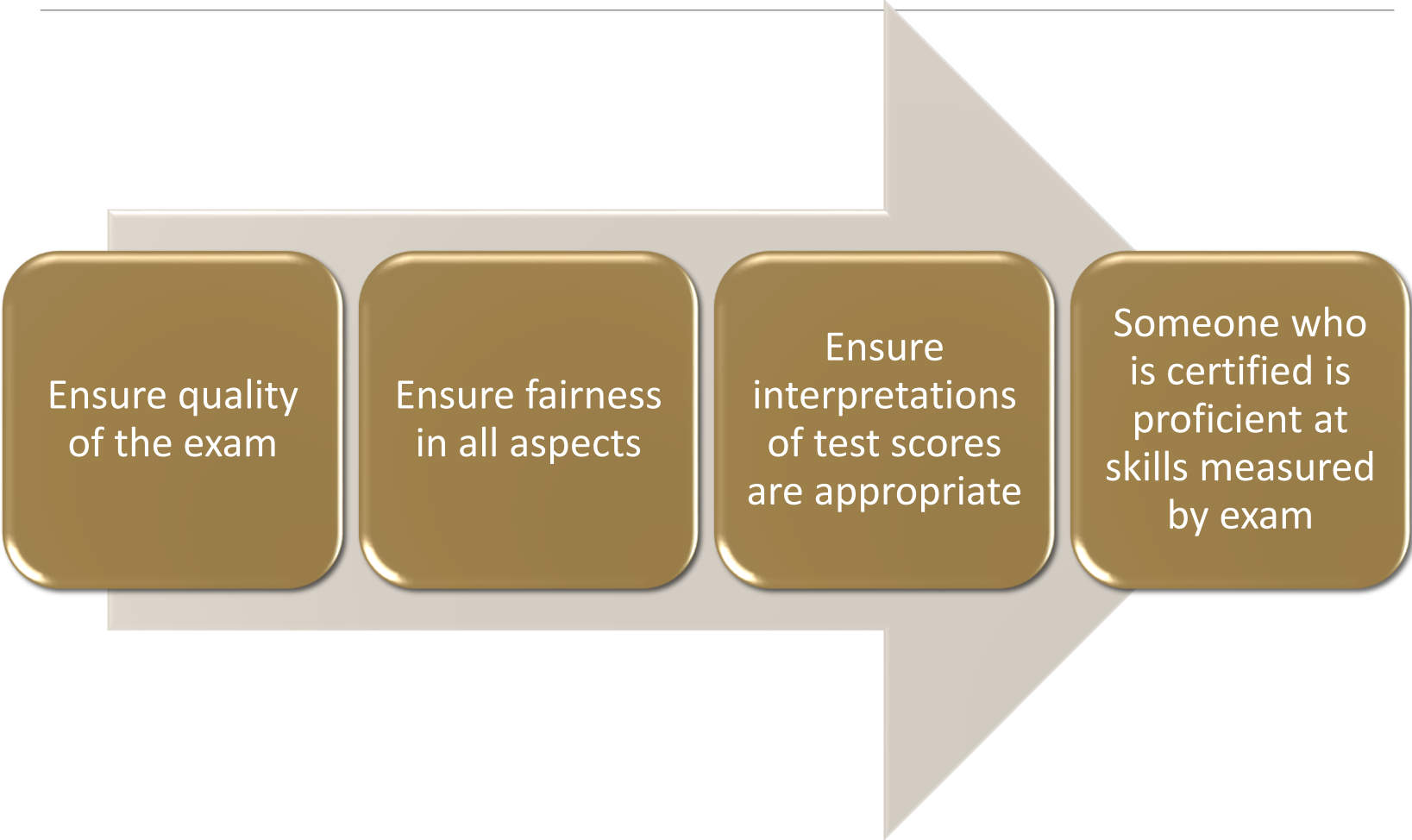
Summary

“...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?



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

Questions

Assessment, Education, and Research Experts

We are passionate about assessing and researching people, places, and things. And once in a while, we help organizations develop and deliver their educational content about people, places, and things.



Austin – Director of Psychometrics and Research

austin@aerexperts.com



Manny – President and Founder

manny@aerexperts.com

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council of Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1-34.

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

Crocker, L, & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guildford Press.

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

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

References

- Kelley, T. L. (1939). Selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30, 17-24.
- Linn, R. L. (Ed.). (1989). *Educational Measurement*. New York: Macmillan.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement: Third Edition* (pp. 13-103). New York, NY: Macmillan.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.
- Sireci, S. G. (2013). *A theory of action of test validation*. Proceedings from the Maryland Assessment Research Center Conference. Available from [http://marces.org/conference/commoncore/MARCES SteveSireci.pdf](http://marces.org/conference/commoncore/MARCES_SteveSireci.pdf)
- Whitley, B. E. (1996). *Principles of Research in Behavioral Science*. Mountain View, CA, Mayfield.