Medical Council of Canada

A Gentle Introduction to Psychometrics for the Medical Educator: Key Concepts and How to Apply Them to Your Assessment

Workshop presented at the AMEE Meeting
Glasgow, UK, Monday, September 7th, 2015

André F. De Champlain, PhD | Director, Psychometrics and Assessment Services

Welcome – From Ottawa

General Objectives

- Provide you with an overview of common terminology used in psychometrics with practical illustrations
- Outline critical analyses that are routinely undertaken with all exams including item analysis, reliability, validity, etc.
- Allow you to apply these concepts via practical examples

Overview

- **Some basic terminology**
  - What is an examination?
  - What is a test score?
  - What is a measurement?

- **A first look at our data: Item analysis**
  - Item difficulty
  - Item discrimination
  - Distractor analysis (quintiles table)

Overview

- **Classical test theory reliability and related concepts**
  - What is reliability
  - The reliability coefficient
  - Types of reliability
  - Standard error of measurement (SEM)
  - Exercise

- **Beyond classical test theory reliability: Generalizability theory**
  - What is g-theory?
  - Examples

Overview

- **Validity and related concepts**
  - What is validity?
  - A history of validity
  - Kane’s validity framework
  - Exercise

Expectations

- **What this workshop is:**
  - An overview of common terminology and models that will demystify psychometrics and hopefully provide you with the knowledge necessary to participate in your assessment activities more fully

- **What this workshop is not:**
  - A condensed PhD in psychometrics!

Some Basic Terminology

- **What is an examination?**
  - A tool that allows us to obtain a sample of an individual's behaviour 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: Illustration

- **Examination**
  - 10 station comprehensive clinical sciences OSCE administered during the last year of the undergraduate degree as a graduation requirement

- **Domain**
  - The (theoretically infinite) pool of clinical sciences OSCE stations (OBGYN, medicine, psychiatry, surgery, etc.) from which you selected 10 stations to include in your graduation clinical sciences 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 OSCE performance presumably represents my overall clinical skill level (the property) in the domains that are targeted by the exam

- **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

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]
- Low values are indicative of “difficult” items
- High values are indicative of “easy” items

How “Difficult” Should Items Be?

Generally [**0.3-0.7**] since these values 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
- However, may still be needed for content validity reasons

How “Difficult” Should Items Be?

One exception: 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

A simple index: ***D***

$$D = p_{upper} - p_{lower}$$

- p_{upper}
  - Prop. of “high-scoring” candidates correctly answering item
- p_{lower}
  - Prop. of “low-scoring” candidates correctly answering item
- **Groups**
  - Median split
  - Top 25%; bottom 25%
  - Top 33%; bottom 33%, etc.

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 = e.g., absenteeism-grade
  - 0 = no relationship: shoe size – test score
  - Positive correlation: OSCE YR3 – OSCE YR4
  - Does **not** assume underlying performance on item is normally distributed

Item Discrimination

Biserial correlation coefficient

- r_{bis}
  - 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 = e.g., absenteeism-grade
  - 0 = no relationship: shoe size – test score
  - Positive correlation: OSCE YR3 – OSCE YR4
  - **Assumes** underlying performance on item is normally distributed

Point-Biserial Correlation

$$\rho_{pbis} = \frac{(\mu_{+} - \mu_{X})}{\sigma_{X}} \sqrt{p/q}$$

μ_{+}= Total score mean for those candidates who correctly answered the item

μ_{X}= Total score mean based on all candidates

σ_{X}= Total score standard deviation based on all candidates

p= Proportion of candidates who correctly answered the item

q= Proportion of candidates who incorrectly answered the item (1-p)

Biserial Correlation

$$\rho_{rbis} = \frac{(\mu_{+} - \mu_{X})}{\sigma_{X}} * \frac{p}{Y}$$

μ_{+}= Total score mean for those candidates who correctly answered the item

μ_{X}= Total score mean based on all candidates

σ_{X}= Total score standard deviation based on all candidates

p= Percentage of candidates who correctly answered the item (difficulty)

Y= Y ordinate of the standard normal curve at the z-score associated with the p-value for this item (height of the curve)

Example

Assumptions:

- *p*-value = 0.60
- $p\text{-}_{upper}$ =0.70 (top 25%)
- $p\text{-}_{lower}$ = 0.40 (bottom 25%)
- μ_{+}= 20.33/30
- μ_{X} = 19.10/30
- σ_{X}= 5.17
- Y = 0.3867

Example

D = 0.70 – 0.4. = 0.30

$$\rho_{pbis} = \frac{(20.33 - 19.10)}{5.17}\sqrt{.60/.40} = 0.29$$

$$\rho_{rbis} = \frac{(20.33 - 19.10)}{5.17} * \frac{0.60}{0.3867} = 0.37$$

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

A Few Comments

- The **biserial** correlation coefficient is used to correlate an artificially dichotomized, normally distributed variable with a continuous variable
- The **point-biserial** correlation coefficient assumes a true dichotomous variable (e.g., gender)
- The point-biserial will always be lower in value than the biserial for a given item
- We generally prefer to compute the $r_{\text{pt-bis}}$

A Few Comments

Item difficulty and item discrimination indices are often confounded

- Items that are either very difficult (**low *p*-values**) or very easy (**high *p*-values**) allow for relatively few differentiations between more and less capable candidates

Content coverage also needs to be considered when determining whether an item should remain on an examination

A Few Comments

- **What about OSCEs and rating scales?**
  - For rating scales (polytomous data), we can calculate the mean scale score as an item difficulty index
  - Similarly, we can calculate a polyserial correlation coefficient as a measure of item (station) discrimination

Distractor Analysis Table

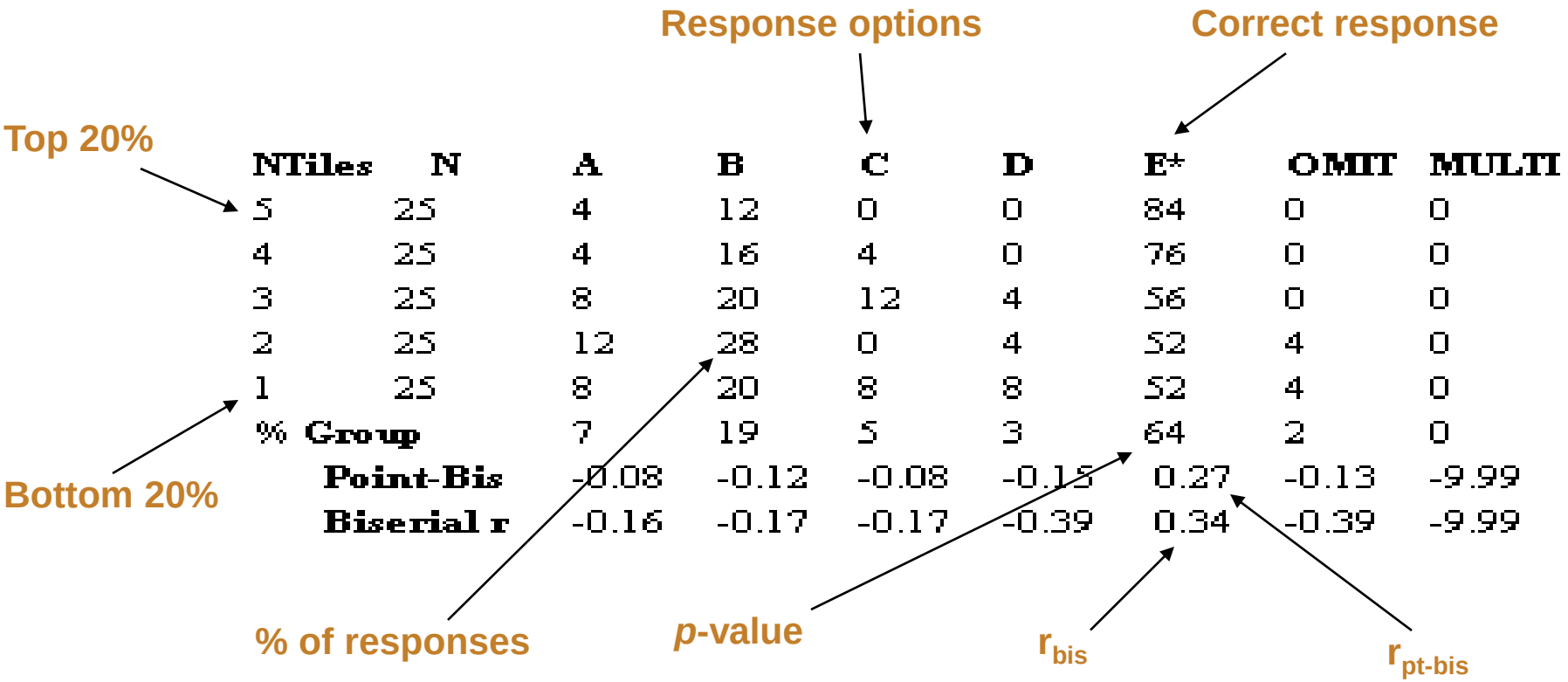
Response options

Correct response

	NTiles	N	A	B	C	D	E*	OMIT	MULTI
Top 20% →	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
Bottom 20% →	1	25	8	20	8	8	52	4	0
	% Group		7	19	5	3	64	2	0

% of responses

Distractor Analysis Table



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
Biserial r		-0.16	-0.17	-0.17	-0.39	0.34	-0.39	-9.99

Example 1

An 8-year old boy is stung by a bee. Within 5 minutes, he develops a 2-cm, raised, red, swollen lesion at the site of the injury. Which of the following findings will be predominant in tissue from the lesion?

A. Foreign body reaction
B. Hemorrhage
C. Lymphocytic infiltration
D. Neurophilic migration
→ E. Vasodilation

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
Biserial r		-0.29	-0.52	-0.17	-9.99	0.49	-9.99	-9.99

p-value

$r_{pt\text{-}bis}$

r_{bis}

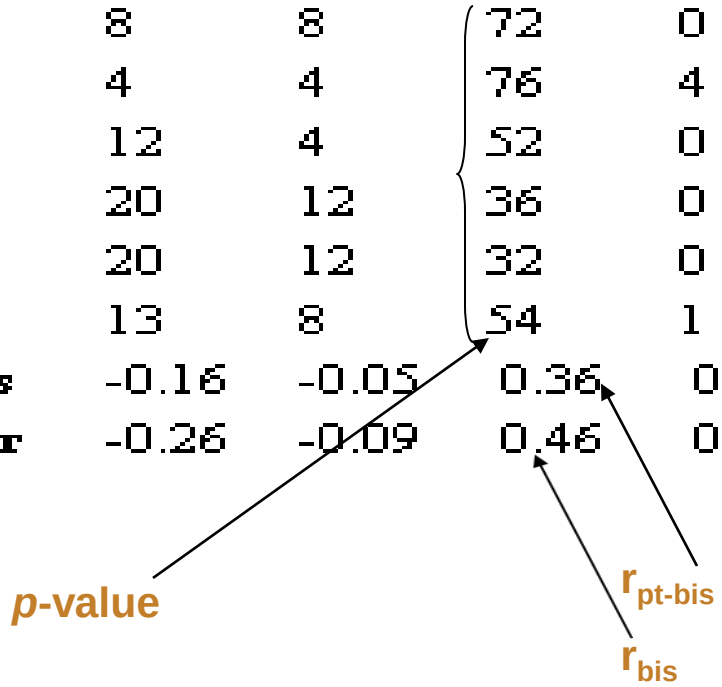
Example 2

A 6-month old boy bruises easily and has bleeding gums on several occasions for 2 months. A maternal uncle has a bleeding disorder. Examination shows several small bruises on the legs. Partial thromboplastin time is prolonged and prothrombin time is normal. Which of the following is the most likely coagulation factor deficiency?

A. Factor IIII
B. Factor VII
→ C. Factor VIII
D. Factor X
E. Factor XIII

Example 2

NTiles	N	A	B	C*	D	E	OMIT	MULTI
5	25	8	8	72	0	12	0	0
4	25	4	4	76	4	12	0	0
3	25	12	4	52	0	32	0	0
2	25	20	12	36	0	32	0	0
1	25	20	12	32	0	36	0	0
% Group		13	8	54	1	25	0	0
Point-Bis		-0.16	-0.05	0.36	0.05	-0.27	-9.99	-9.99
Biserial r		-0.26	-0.09	0.46	0.19	-0.37	-9.99	-9.99



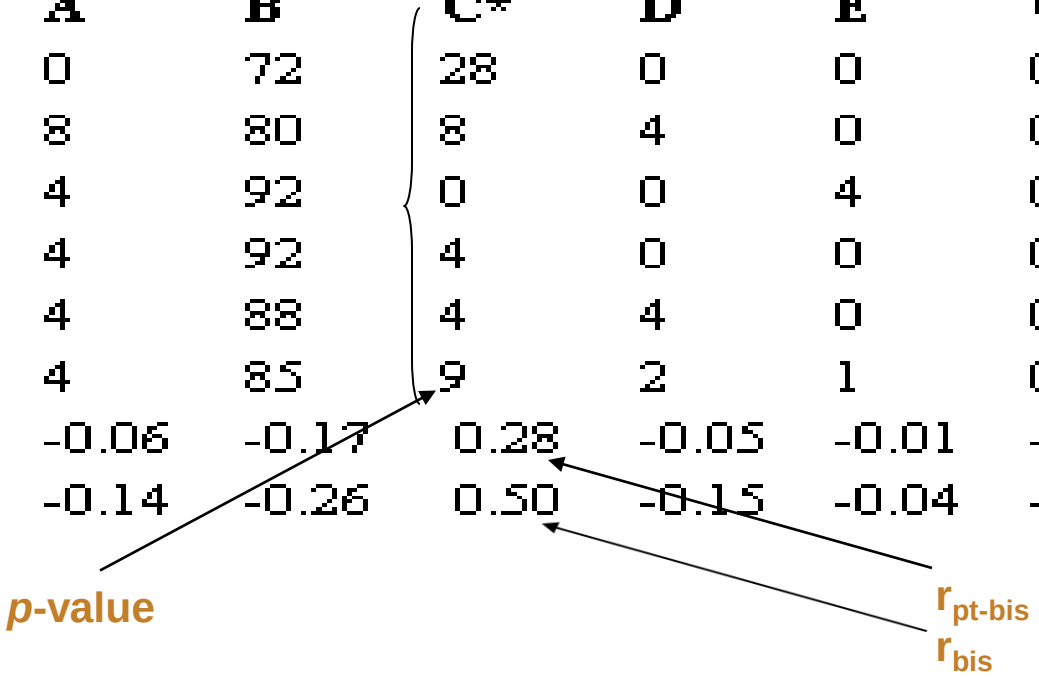
Example 3

Large amounts of the artificial sweetener aspartame should be avoided in children who have which of the following metabolic disorders?

A. Diabetes mellitus
B. Phenylketonuria
→ C. Hereditary fructose intolerance
D. Lactose intolerance
E. Maple syrup urine disease

Example 3

NTiles	N	A	B	C*	D	E	OMIT	MULTI
5	25	0	72	28	0	0	0	0
4	25	8	80	8	4	0	0	0
3	25	4	92	0	0	4	0	0
2	25	4	92	4	0	0	0	0
1	25	4	88	4	4	0	0	0
% Group		4	85	9	2	1	0	0
Point-Bis		-0.06	-0.17	0.28	-0.05	-0.01	-9.99	-9.99
Biserial r		-0.14	-0.26	0.50	-0.15	-0.04	-9.99	-9.99



Software

- **SAS**
  - **PROC FREQ**
    - *p*-values & fifths tables
  - **PROC CORR Alpha**
    - Item-total correlations (point-biserial)
    - BISERIAL macro (Behavior Research Methods, 2007, 39 (3), 527-530)
- **SPSS**
  - **FREQUENCIES**
    - *p*-values & fifths tables
  - **RELIABILITY**
    - Item-scale correlations (point-biserial)
    - Biserial macro (Behavior Research Methods, 2007, 39 (3), 527-530)

Reliability

- **Standards for Educational & Psychological Testing (2014)**
  - "For each total score, subscore or combination of scores that is to be interpreted, estimates of relevant reliabilities should be reported" (Standard 2.3)
  - "Each method of quantifying the precision or consistency of scores and/or decisions should be described clearly and expressed in terms of statistics appropriate to the method" (Standard 2.19)

Reliability

- **What is reliability?**
  - Reliability provides us with an indication of the *degree of consistency* (or precision) with which test scores and/or decisions are being measured by a given examination (*sample* of OSCE stations, sample of MCQs, sample workplace-based assessments, etc.)
  - A person's observed (actual) test score is composed of a "true score" and measurement error, i.e., X= T + E – or -

$$\sigma^2_X = \sigma^2_T + \sigma^2_E$$

  - A student's true score is the score that would be obtained if the test was measuring the ability of interest in a perfectly consistent fashion

Reliability

Test score – a reminder:

- Any examination, by virtue of practical constraints (e.g., available testing time) is comprised of a very restricted number of items, stations, tasks that compose the domain of interest
  - My undergraduate clinical sciences OSCE contains 10 stations that can be administered in a 4-hour exam

- But as a test score user, are you really interested in the performance of candidates on those very specific 10 stations?

- **No** - you're interested in generalizing from the performance on those 10 very specific OSCE stations to the broader domains of interest

Reliability

- **Test score – a reminder:**
  - If my undergraduate OSCE targets the *Hx*, *CM*, *PE* and *IP* skills of candidates, you are inferring or generalizing my ability level to those domains based on my performance on those very specific 10 OSCE stations
  - Reliability refers to the level of accuracy (or consistency, precision) with which we can make that generalization (from my performance on the exam to the broader domains)
  - How well does my score of 85% on the 10 OSCE stations reflect my *true* clinical skills?

Reliability

- **Reliability coefficient**
  - A reliability coefficient allows us to estimate the degree of consistency (or precision) with which test scores are being measured by our examination for a given group of candidates
  - The higher the reliability coefficient value, the greater the degree of consistency (or precision) (ranges 0 ---> +1.00)
  - According to classical test theory, a reliability coefficient provides an estimate of the ratio of true score variance to observed score variance

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

Reliability

- **Estimating a reliability coefficient**
  - How do we estimate a reliability coefficient?
  - Degree of consistency in generalizing across which factor (facet)?
    - Time (test-retest or coefficient of stability)?
    - Equivalent forms (coefficient of equivalence)?
    - Two halves of an exam (split-half coefficient)?
    - Etc.
  - Generally interested in estimating the degree of consistency in performance from item to item throughout a test form that is due to the candidates' true ability level (their true scores)
  - Cronbach's coefficient alpha requires one single test administration

Cronbach's Alpha

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma^2{}_i}{\sigma^2{}_X}\right)$$

$$KR-20 = \frac{n}{n-1}\left(1 - \frac{\sum p_i q_i}{\sigma^2{}_X}\right)$$

n = Number of test items (whether MCQs or OSCE stations)

p_i = Proportion of candidates who correctly answer an item

q_i= Proportion of candidates who incorrectly answer an item ($1-p_i$)

σ^2_X = Total score variance

Reliability: Example

Item (*p*-value)	Item (*q*-value)	Item Variances (*p***q*)
1 (.30)	1 (.70)	1 (.21)
2 (.40)	2 (.60)	2 (.24)
3 (.10)	3 (.90)	3 (.09)
4 (.20)	4 (.80)	4 (.16)
5 (.80)	5 (.20)	5 (.16)
6 (.30)	6 (.70)	6 (.21)
7 (.40)	7 (.60)	7 (.24)
8 (.50)	8 (.50)	8 (.25)

- Sum ($\Sigma\sigma^2_i$) =1.56
- Assume total score variance (σ^2_x) is 5

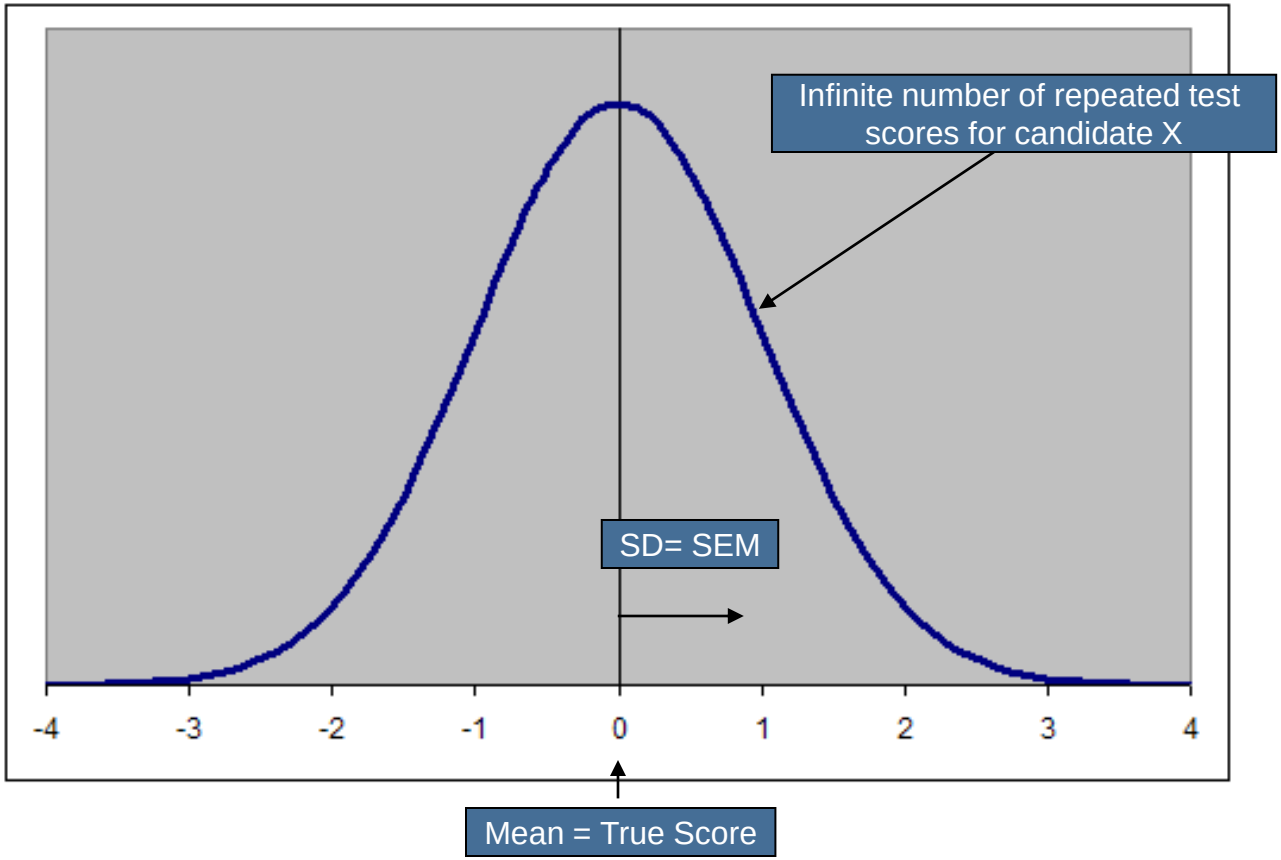
Reliability: Example

- $\infty = 8/7 * \{1 – 1.56/5\}$
- $\infty = 1.1429 * 0.688 = 0.7863$ or 0.79
- **79% of observed score variance in performance from item to item throughout the test is accounted for by true score variance**

Reliability

- **Precision of score estimates?**
  - **Accuracy of a given score estimate**
    - Reliability coefficient applies to a group of scores
    - What about the consistency of one score point?
    - We need an estimate of measurement error
    - **Standard error of measurement (SEM)**
  - **Standard error of measurement**
    - Expected amount of error in individual observed scores (X)
    - SEM = SD * (1 - Reliability estimate)$^{1/2}$
    - Using NCT, we can estimate the interval within which a student's "T" falls given a certain probability

Standard Error of Measurement (SEM)



SEM: Example

- **How accurately does my score of ‘4/8’ reflect my true ability?**
  - SEM = $\{5\}^{1/2} * \{1-0.7863\}^{1/2}$
  - SEM = 2.2361 * 0.4623 = 1.03
  - **95 C.I. = 4 ± (1.96 * 1.03) = [1.98 – 6.02]**
- If we re-tested our candidate 100 times, we would expect his/her observed score to fall between 1.98 and 6.02 95% of the time
- Not a very accurate reflection of their true ability level!

Beyond CTT Reliability: G-Theory

- **Measurement error arises from multiple sources (multifaceted)**
  - For an OSCE, measurement error could be attributable to:
    - The selection of a particular set of stations
    - SP portrayal effects
    - Occasion effects
    - Rater effects
    - Setting (if your OSCE is given at multiple locations)
  - We need to clearly identify these sources and address them *a priori*
  - **We can quantify these sources of error using *G*-theory**

Use of Generalizability Theory

- **What is Generalizability Theory (g-theory)?**
  - G-theory is a framework that allows us to generalize the consistency of performance to other conditions
    - What would be the reliability of my OSCE if I had 7 stations instead of 10?
    - Will 2 raters and 10 stations permit me to attain a reliability of .7?
    - How is the standard error of measurement affected by decreasing the number of stations by 25%?
  - The general idea is to identify all the sources of error (facets). We can then estimate reliability when manipulating some or all of these components of error

Use of Generalizability Theory

- **Classical Test Theory vs. G-theory**
  - **Similarities between CTT and G-theory**
    - Decomposes observed variance into true-score and error components
  - **Differences between CTT and G-theory**
    - G-theory uses other information in the design to decompose the variance further
    - G-theory allows for much more substantive “what if” questions to be answered

Reliability Exercise

- **Scenario**
  - A Medicine course director asks you to estimate the reliability of end-of-clerkship OSCE scores for the assessment of clinical & communication skills of students

- “**Parameters**”
  - The assessment is to be given over a 1-week period
  - All students are assessed on 6 stations using a 1-5 global rating scale (Total score: 6-30)
  - Passing mark = 15

Reliability Exercise

- **Questions**
  - What is the reliability of scores?
  - How much measurement error is associated with the cut-score (95% CI)?
  - How do I interpret these values?
  - What are some of the facets that might impact reliability?
  - How can I minimize their impact before I administer the exam again?

Reliability Exercise: Data

Station	Station Variance (σ^2_i)	
1	1.7	
2	1.5	
3	2.0	
4	0.8	
5	0.7	
6	1.5	
Total	$\Sigma\sigma^2_i = 8.2$	$\sigma^2_x = 13$

Reliability Exercise: Data

Station	Station Variance (σ^2_i)	
1	1.7	
2	1.5	
3	2.0	
4	0.8	
5	0.7	
6	1.5	
Total	$\Sigma\sigma^2_i = 8.2$	$\sigma^2_x = 13$

Reliability Exercise: ∞, SEM & 95% CI

$$\infty = \frac{6}{6-1}\left(1 - \frac{8.2}{13}\right) = 0.44$$

$$SEM = 13 * \sqrt{1 - 0.44 =} \qquad 9.728$$

95% CI = $15 \pm 1.96(9.728) =$ (-4.06688 , 34.06688)

MEDICAL COUNCIL OF CANADA | LE CONSEIL MÉDICAL DU CANADA

Reliability Example: Discussion & Recommendations?

What Validity 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)

What Validity Is 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 & Psychological Testing (2014)

“The test developer should set forth clearly how test scores are intended to be interpreted and used. The populations for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described” (Standard 1.1)

“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)

Standards for Educational & Psychological Testing (2014)

“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)

“If the test is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, providing a rationale and collecting new evidence if necessary” (Standard 1.4)

Important Steps to Follow Prior to Collecting a Single Piece of Data

- 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

Gathering Validity Evidence

Validity Theory – A Brief History

Criterion-based model of validity

How well does the test score predict the criterion?

Cureton (1950)

Content-based model of validity

How well do my items represent the domain?

APA (1954)

Construct-based model of validity

Can I specify a theoretical framework for what I am trying to measure and an empirical framework for how I will test out the model & linkages?

(Cronbach & Meehl 1955)

Current models of validity

Unified framework of construct validity (Messick, 1989)

Argument-based approach to validation (Kane, 1992)

Previous Validity Frameworks: Concerns

- **“Trinitarian” model of validity (1954)**
  - **“Content, Criterion-related and Construct Validity”**
    - **Criterion-based model of validity**
    - Very difficult to identify well-defined, reliable criteria
    - Also need to validate criterion measures!
    - Can become a circular argument
  - **Content-based model of validity**
    - **Subjective**
      - Prone to confirmation biases
      - Item relevance judgments often made by test developers!

Previous Validity Frameworks: Concerns

- **“Trinitarian” model of validity (1954)**
  - **Construct-based model of validity**
    - Very few clearly definable nomological networks in education

- **Encourages a “toolbox” approach to validation efforts**
  - Leads to an “opportunistic” choice of validity evidence
    - Researcher uses whatever data are available
    - Leads to proliferation of “weak” validation research

Modern Validity Frameworks

- **Unified framework of construct validity (Messick, 1989)**
  - Formalization of earlier work – “all validity is construct validity” (Loevinger, 1957)
  - “Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment”

Modern Validity Frameworks

- **Unified framework of construct validity (Messick, 1989)**
  - Unified framework doesn't aid the practitioner in identifying questions that need to be answered as part of validation efforts – "it's all construct validity"
    - **Where do I start?**

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 (Kane, 1992)

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 (Kane, 1992)

- **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

Argument-Based Approach to Validation (Kane, 1992)

- **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 (Kane, 1992)

- **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 (Kane, 1992)

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

  - **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 this OSCE have better clinical and communication skills
    - Candidates who score higher on family medicine clerkship exam will do well in a family medicine residency, etc.

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 medical school has asked you to develop a MCQ exam that will be used to admit students to your undergraduate program

- **Scenario 2**
  - The minister of health has asked you to assemble a practice-based assessment as part of revalidation efforts for physicians in your country; assessments include mini-CEXs and multisource feedback

Validity: Exercise

For each scenario:

- Lay out the main interpretive argument(s)
- Indicate which sources of evidence will be used to support these arguments
- Indicate which arguments (inferences) are NOT supported by your sources of evidence

MEDICAL COUNCIL OF CANADA
LE CONSEIL MÉDICAL DU CANADA

THANK YOU!