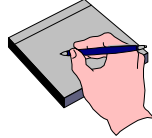


Overview of Ch14: Classroom Assessment Grading, & Standardized Testing

- What is a standardized test?
- Types of standardized tests
- Norms & Standardization
- Reliability
- Interpreting Test Scores
- Validity
- Issues in testing



1

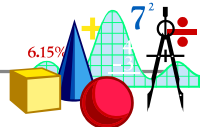
What is a Standardized Test?

- Objective
- Standardized
- Measure (i.e. yields scores or categories)
- of a sample of behavior
- from which predictions can be made
 - Achievement: Predicts previous learning
 - Aptitude: Used to predict future performance
 - Diagnostic: Predicts present characteristics/ disabilities.

2

Types of Tests

- Norm-referenced
 - Compare people to people
 - Useful when selecting top candidates
- Criterion-referenced
 - Assesses mastery of basic skills
 - Useful when grouping for instruction



Which is more appropriate for school settings?

3

Why Do We Need Norms?

- An examinee's raw score is meaningless for comparisons, even when expressed as % correct.
- Using norms based on the performance of the standardization (or norm) sample, raw score of any examinee can be converted into a standard score (a score derived from the raw score just like percentile scores) which is meaningful because it allows
 - evaluation of individual's performance relative to norm group
 - comparison of individual's performance on different tests

4

How Norms are Developed

1. Test appropriate norm group
2. Calculate the Mean, a measure of central tendency
 - Mean: Add up scores & divide by # of scores
3. Check against the other 2 measures of central tendency (median and mode) to make sure all 3 more or less equal
 - Median: the middle score. Scores must be in ascending order
 - Mode: score that occurs most often
 - If Mean=Median=Mode, then distribution normal

5

How Norms are Developed Cont'd



4. Also look at the shape of the frequency distribution polygon (curve) or histogram (bar graph) to make sure scores are more or less normally distributed (i.e. the curve is Bell-shaped).
 - Characteristics of Normal Distribution
 - 50% of scores above the mean, 50% below
 - 68% of scores within 1 SD from the mean
 - 95% of scores within 2 SD from the mean
 - 99% of scores within 3 SD from the mean

6

How Norms are Developed Cont'd

5. Calculate the Standard Deviation (S.D.), a measure of score variability around the mean.
- » Why range not sufficient indicator of variability?
 - » Formula for S.D.: $SD = \text{Square root of (the sum of the squared deviations from the mean divided by the number of students)}$

7

Calculating S.D.

Calculate the mean: $\bar{\chi}$

Subtract the mean from each score: $(\chi - \bar{\chi})$

Square each difference: $(\chi - \bar{\chi})^2$

Add all the squared differences: $\Sigma(\chi - \bar{\chi})^2$

Divide by the number of scores: $\frac{\Sigma(\chi - \bar{\chi})^2}{N}$

Find the square root: $\sqrt{\frac{\Sigma(\chi - \bar{\chi})^2}{N}}$

Non-Normal Distributions

Note: If scores not normally distributed then formulas we have for calculating SD & z-scores are not applicable.

Examples of non-normal distributions:

- » Positively skewed distributions
- » Negatively skewed distributions
- » Bi-modal distributions.

Think about possible problems that the distribution of scores indicate.

Rewrite the test if necessary in order to obtain a normal distribution.

9

How Norms Are Used to Compute Standard Scores

- Now that you have the norms (Mean and SD), you are ready to compute standard scores, which are derived from raw scores and allow meaningful comparisons.
- You can convert a raw score to a standard z-score.
Formula: $z = (\text{raw score} - \text{Mean}) / \text{SD}$
- Standard z-scores have Mean=0 & SD=1. This can be inconvenient. To avoid this problem you may:
 - Use standardized t-scores (common in education) which have a mean on 50 & a SD of 10.
 - » Formula for t-scores: $t = 10z + 50$
 - Or, use any other standardized score by plugging in the desired mean & SD into the formula. Examples:
 - » Formula for IQ scores: $\text{IQ} = 15z + 100$
 - » Formula for SAT scores: $\text{SAT} = 100z + 500$

10

Problems with Other Derived Scores

Warning #1: When making comparisons, you should not use percentile rank scores.

- Why? Because have a major drawback--they distort the measurement scale at the extremes.

Warning #2: Grade Equivalent Scores should not be used in making comparisons.

- Why? Can be misleading.

So, always use standard scores for comparisons.

11

Reliability

- Definition: Rel. is the consistency of scores obtained by same person when tested on different occasions
- Classical theory of measurement: $X = T + e$ (obtained score=true score + random error)
 - e is always there because it is random. Best we can hope for is to minimize e
- Rel=degree to which e is minimized so that obtained score (X) reflects true score (T)
- Rel=relationship between 1st score (X) & 2nd score (Y) on a test = the correlation coefficient between X & Y = "r"

12

Types and Size of Reliability

- Different estimates of reliability are based on what the test writer/user suspects to be possible sources of error.
 - Test-Retest (stability): $e = \text{time}$
 - Alternate-Form: $e = \text{content}$
 - Inter-Rater/observer: $e = \text{Raters/observers}$
 - Internal consistency: $e = \text{content (test heterogeneity)}$
 - » Split-Half
 - » Coefficient Alpha (α): Conceptually, the average of all split-half reliabilities
- What is the size of an acceptable Reliability?

13

Using Rel. in Interpretation of Individual Scores



- Assume you can give the test to a person an infinite # of times. The person would obtain various scores on the test.
- The mean of these scores would be T .
- These scores would vary around T as a function of the unreliability of the test. If test highly unreliable then wide range of scores.
- Conceptually: Standard Error of Measurement (SEM) = SD of this hypothetical distribution of scores = the degree to which we are confident that the score the student obtained reflects his/her true ability

14

SEM and Confidence Intervals

- Computationally: $SEM = SD \text{ of test} * \sqrt{(1-Rel)}$
Example: If $SD=10$, $Rel=.84$, then
$$SEM = 10 * \sqrt{(1-.84)} = 10 * .4 = 4$$
- SEM used to construct confidence intervals.
Example: if a person has an obtained score (X) of 30 on a test that has a SEM of 4, we interpret this to mean:
 - we are 68% sure that his true score lies anywhere between 26 & 34 ($X \pm (1 * SEM)$ i.e. 30 ± 4)
 - we are 95% sure that his true score lies anywhere between 22 & 38 ($X \pm (2 * SEM)$ i.e. 30 ± 8)
 - we are 99% sure that his true score lies anywhere between 18 & 42 ($X \pm (3 * SEM)$ i.e. 30 ± 12)

15

Validity

- Definition: Degree to which the inferences we make from test scores are accurate
 - Can a test be reliable but not valid?
 - Can a test be valid but not reliable?
- Validation Strategies
 - Content Validation: Degree to which test content constitutes a representative sample of the behavior domain intended, as judged by experts.
 - Criterion-Related validation: correlating test scores with present criterion (concurrent validity) or future criterion (predictive validity) about which you are hoping to make inferences. Examples?

16

Validation Strategies Cont'd

- Construct Validation: Determining extent to which a test measures a theoretical construct through looking at
 - » Developmental changes
 - » Experimental studies & interventions
 - » Internal consistency (Alpha)
 - » Discriminant validity (low correlations with tests measuring un-related constructs)
 - » Convergent validity (high correlations with tests measuring related constructs)
- What is an acceptable size for a validity coefficient?

17

Issues in Testing

- Test use in controlled. Why?
 - So that only qualified examiners give it. This ensures:
 - » Selection of quality, appropriate tests for the individuals to be tested
 - » Correct administration & scoring : clerical errors biggest source of error in test scoring
 - » Proper interpretation of test scores
 - » Protection of privacy & confidentiality
 - To prevent familiarity with test content, which invalidates the test. Note distinction between:
 - » Familiarity with test content
 - » Test orientation procedures
 - » Training in underlying skills

18

Issues in Testing Cont'd

- Discussing test scores with families:
 - Explain meaning of percentiles and standard scores
 - Caution them about misinterpreting grade equivalents
 - Talk in terms of confidence intervals to emphasize imperfection of test & never use it as only basis for decision
- ADA of 1989 protects the physically/mentally impaired. Requires providing reasonable accommodations for them in testing, including:
 - use of signers/readers, extended time limits, alternate forms of test responses, accessible testing locations and private sessions
 - not taking the test altogether
