Norm and the Meaning of Test Scores

Lecture 3 Psychological Testing and Measurement

Sunthud Pornpresertmanit

Why interpreting raw score?

Student A
Math = 50
English = 40
Science = 70

Student B
Math = 60
English = 45
Science = 40

Why interpreting raw score?

- A raw score on any psychological test is meaningless.
  - Raw score
  - Percentage
- The difficulty level of the items making up each test will determine the meaning of the score.

How to interpret scores?

- Relative Score (Norm-referenced)
- Absolute Score (Criterion-referenced)

Norm-referenced Test

Norm-referenced

- The relative scores are designed to serve dual purpose.
  - Evaluate individual performance from normative sample
  - Provide comparable measures on different test
- The relative scores can be expressed in the same units.

Norm-referenced

- Relative scores are expressed in one of two major ways.
  - Developmental level attained
  - Relative position within a specified group

Developmental Norms

- How far along the normal developmental path the individual has progressed
- They are psychometrical crude and do not lend themselves well to precise statistical treatment.

Developmental Norms

- Mental Age
- Grade Equivalent
- Ordinal Scale

Mental Age: Age-equivalent items

- Items were grouped into year levels.
- Scores based on the highest year level that child attained.
- In actual practice, the individual performance showed certain amount of scatter.
- Therefore, it was customary to compute basal age and give partial credits for all test passed at higher year levels.

Mental Age: Age norms

- Standardized samples of each age constitute the age norm on the test.
- All raw scores on such a test can be transformed in a similar manner to the age norms.

Mental Age: Limitations

- The mental age unit tend to shrink with advancing years.
- Intellectual development progress and maturity

Grade-equivalent

- Like Mental Age
- Computing the mean raw score obtained by children in each grade and then equating raw score to grade mean

Grade-equivalent: Limitations

- Content of instruction varies from grade to grade
  - Unequal distance
- Grade norms tend to be incorrectly regarded as performance standards
- Grade norms are subject to misinterpretation

Ordinal scale

- Developmental norms derives from research in child psychology (Empirical observation).
- Gesell Developmental Schedules
  - Sequential development in motor, adaptive, language, personal-social from 4 weeks to 36 months

Ordinal scale

- Jean Piaget
  - Concerns in specific concepts in cognitive development than broad abilities (e.g. object permanence, conservation)
  - His theory is used in standardized scales.
  - Critical analyses and empirical evaluations have highlighted both its constructive features and limitations.

Ordinal scale

- In conclusion, ordinal scales are designed to identify the stage reached by the child in the development of specific behavior functions.
- Interpreting score in both quantitative (age) and qualitative (description)
- It share important features with the criterion-referenced tests.

Discussing Question

- Why grade norms are subject to misinterpretation?
- What is the criterion that discriminate whether behavior can be developed to ordinal scale or not?

Within-group Norms

- The individual's performance is evaluated in terms of the performance of the most nearly comparable standardized group.
- Clearly quantitative meaning defined

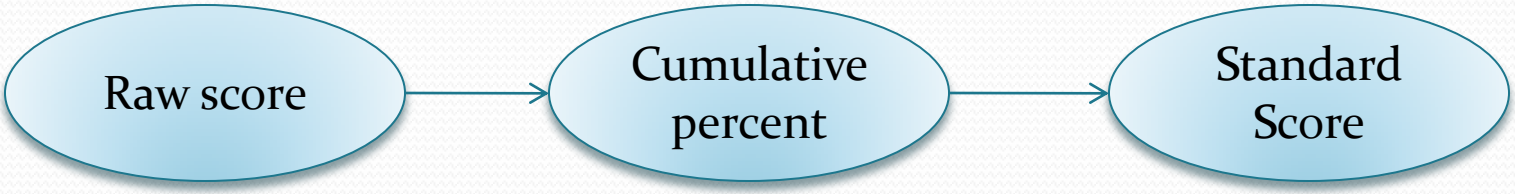
Within-group Norms

- Percentile: percentage of persons in the standardization sample who fall below a given raw score
- Standard Score (z-score): transformation from raw score to M = 0, SD = 1
- Derived Standard Score: transformation from raw score
  - Deviation IQ: M = 100, SD = 15 or 16 (depend on tests)
  - T-score: M = 50, SD = 10
  - CEEB score: M = 500, SD = 100
  - Stanine: M = 5, SD = 2

Within-group Norms

- Standard score and derived standard score
  - Linear Transformation

  - Normalized Transformation



Within-group Norms

- Why transform to normal distribution?
  - Most raw score distributions approximate the normal curve more closely than they do any other type of curve.
  - Normal curve has many useful mathematical properties.
- Normalized transformation should be carried out only when the sample is large and representative.
  - Deviation from normality results from defects in the test

Discussing Question

- What are the advantages and disadvantages of percentile, standard score and derived standard score?
- Why ratio IQ is not comparable at different age level?
- What is the characteristics of samples making up norms?
- Why population used for making norms affect interpreting scores and comparability of derived scores within a person?

Discussing Question

	Percentile	z score	Derived z
Easy to compute and understand			
Inequality of units			
Negative value and decimal			

Norms

- Test scores cannot be interpreted in abstract; they must be referred to particular test.
- There are three principal reasons to account for systematic variations among the scores obtained by the same individual on different test?
  - Test may differ in contest despite similar labels.
  - The scale unit may not be comparable.
  - The composition of the standardized samples used in establishing norms for different tests may vary.

Norms

- Which population establishing norms?
- How to draw a sample from population?
- How much sample to be used for establishing norms?
- How much response rates?

Discussing Question

- Why institutional samples are not representative of population?
  - School
  - Hospital
  - Prison
- What are the advantages and disadvantages of restricted population?

Equating Norms Procedures

- How to compare scores of individuals or groups across time or test
  - Alternate form
  - Anchor test
  - Fixed referenced group
  - Simultaneous norming

Alternate form

- Two or more versions of test
- Alternate form = tests that are exactly the same in content sampling
- Parallel form = content, M, SD, Reliability and validity

Simultaneous norming

- By norming tests at the same time and on the same group of people, one can readily compare the performance of individuals or subgroups on more than one test, using same standard.
- Equipercentile method
- Example: Metropolitan Achievement Test
  - 300,000 four-, fifth-, sixth grade schoolchildren
  - 7 batteries that consist of reading comprehension and vocabulary subtests

Anchor tests

- Anchor test consist of common sets of items administered to different groups in the context of two or more tests.
- These items allow comparability of test scores from one test to another.
- The accuracy of comparability depends on correlations between tests equated, reliability and validity of each test.

Specific Norms

- For many testing purposes, highly specific norms are desirable.
- Subgroups may be formed with respect to age, grade, type of curriculum, sex, geographical region, urban or rural environment, socioeconomic level etc.
- Local norm developed by the test users within particular settings.
- Local norms are more appropriate than broad national norms for many testing purposes, such as prediction of subsequent performance or college achievement.

Fixed reference groups

- Fixed norms from one test (Reference group like specific norm)
- When develop new test, use some items from the test in new test for equating the new test to the norm (like anchor test)
- Score scale maintain over time

Fixed reference groups

- Example: Scholastic Aptitude Test
  - 1926-1941 → Norms in each year
  - After 1941 → 11,000 test takers in 1941 used as reference group
  - Any change in the candidate over time can be detected only with a fixed-score scale.
  - In 1995, SAT changed fixed reference group from 1941 to 1990 test takers.

Item response theory

- Item response theory (IRT) or latent trait theory
- Based on probability that a person of specified ability (the so-called latent trait) succeeds on an item of specified difficulty.
- IRT models are used to establish a uniform “sample-free” scale of measurement.

Item response theory

- IRT models place both persons and test items on a common scale
- If item parameter estimates (such as item difficulty) are invariant across population, it is not necessarily to fix performance to referenced group.

Item response theory

- IRT is suited for use in computerized adaptive testing (CAT)
- Their ability levels can be estimated based on their responses to test items during the testing process
- These estimates are used to select subsequent sets of test items that are appropriate to the test takers' ability levels
- Problems of CAT are test security, test costs, and the inability of examinees to review and amend their responses.

Criterion-referenced Test

Criterion-referenced test

- Criterion-, content-, domain-, or objective-referenced test
- Criterion-referenced testing uses as its interpretive frame of reference in specified criterion or standard.
- Most found in education and occupational settings.
- The CRT focus is on what test takers can do and what they know, not on how they compare with others.

Criterion-referenced test

- The CRT can be used in 2 perspectives
  - Knowledge testing
  - Mastery testing

Knowledge testing

- The important procedure to construct this test is a clearly defined domain of knowledge or skills to be assessed by the test and application of that knowledge.
- Table of specifications
- Then, items are prepared to sample each objective.
- Test items may be written in protocol.
- These tests are usually described as measures of “achievement.”

Knowledge testing

- This CRT score has qualitative meaning.
- This CRT is equivalent to interpreting test scores in the light of the demonstrated validity of the particular test. (can combined with NRT)

Performance Assessment

- Sometimes, the CRT can be used to ascertain or certify competence in tasks that are more realistic.
- The question is that “Does this test taker display mastery of the skill in question?”
- This question can be rated by subjective judgment or develop method for applying the criteria.

Mastery testing

- Procedures that evaluate test performance on the basis of whether test taker does or does not demonstrate a preestablished level of mastery are known as mastery test.
- After suitable training, nearly everyone can achieved complete mastery.
- Therefore, individual differences in performance are of little or no interest.

Mastery testing

- Two important questions?
  - How many items must be used for reliable assessment of each of the specific instructional objectives covered by the test? (Number of items)
  - What proportions of item must be correct for the reliable establishment of mastery? (Cutoff)
- These questions can be solved by statistical techniques.
- The good example is qualifying for a driver's license test.

Mastery testing

- The utility of mastery and nonmastery can be used to specify cutoff score.
- Mastery tests are suitable for basic skills.

Relation between NRT and CRT

- Beyond basic skills, mastery testing is inapplicable or insufficient.
- Therefore, NRT is usually used in complex skills.
- Some published test are so constructed as to permit both norm-referenced and domain-referenced applications.
- An example is provided by Stanford Diagnostic Test in reading and in mathematics.

Relation between NRT and CRT

- A normative framework is implicit in all testing.
- Applying a cutoff point to dichotomize performance simply ignores the remaining individual differences within the two categories and discards potentially useful information.

Discussing Question

- Are letter grades suitable for assessing academic performance? Why?
- Suppose that you are clinical psychologists license tests developers. Making this license test by these steps.
  - Define content domain and objective
  - Create how to measure
  - Develop cutoff standard (by your opinion)