Item Analysis

Lecture 6 Psychological Testing and Measurement

Sunthud Pornpresertmanit

Item Analysis

- Item can be analyzed qualitatively or quantitatively.
- Reliability and validity of the test depends on characteristics of items.
- Item analysis makes it possible to shorten a test and at the same time to increase its validity and reliability, in contrast with Spearman-Brown formula.

Item Difficulty

- The difficulty of an item is defined in terms of the percentage (or proportion) of persons who answer it correctly.
- It is customary to arrange items in order of difficulty. It gives the test takers confidence and reduces the likelihood of their wasting much time on items beyond their ability.

Item Difficulty

- In process of test construction, a major reason for measuring item difficulty is to choose items of suitable difficulty level.
- Very easy or very difficult items do not provide information about individual difference.
- These items contribute nothing to reliability and validity of the test.

Item Difficulty

- However, very easy items improve test takers' morale.
- For maximum differentiation, one would seem that items at the .50 test difficulty level should be selected.
- Because of item intercorrelations, it is best to select items with a moderate spread of difficulty level, but whose average difficulty is .50

Item Difficulty

- Possibility of guessing in multiple-choice items
- Lord (1952) suggested that, for a five-option multiple-choice item, the average proportion correct should be approximately .69.

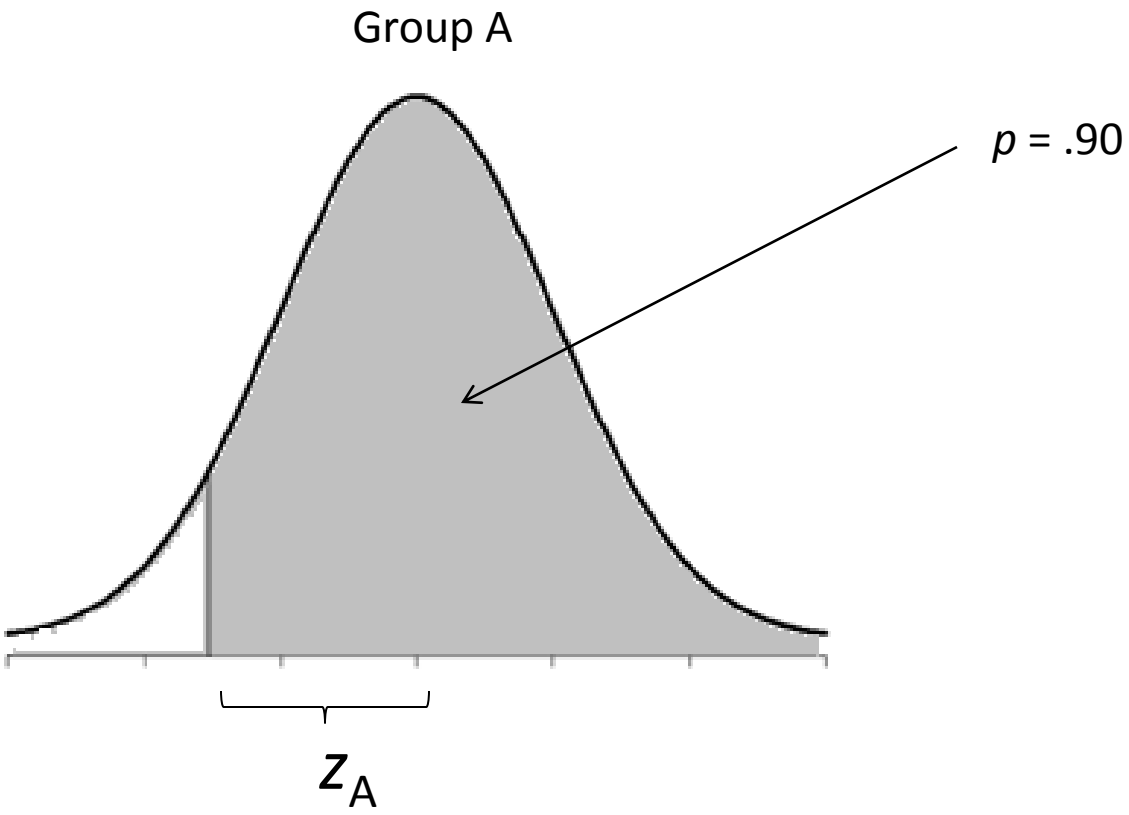
Item Difficulty

- We cannot infer that the difference in difficulty is equal distance in the range of scale.
- Equal percentage difference would correspond to equal differences in difficulty only in rectangular distribution.

Item Difficulty

- If we assumed a normal distribution of the trait measured by any given item, the difficulty level of item can be expressed in terms of an equal-unit interval scale by reference to a table of standard normal distribution.
- The more difficult items have plus values, the easier items minus value.

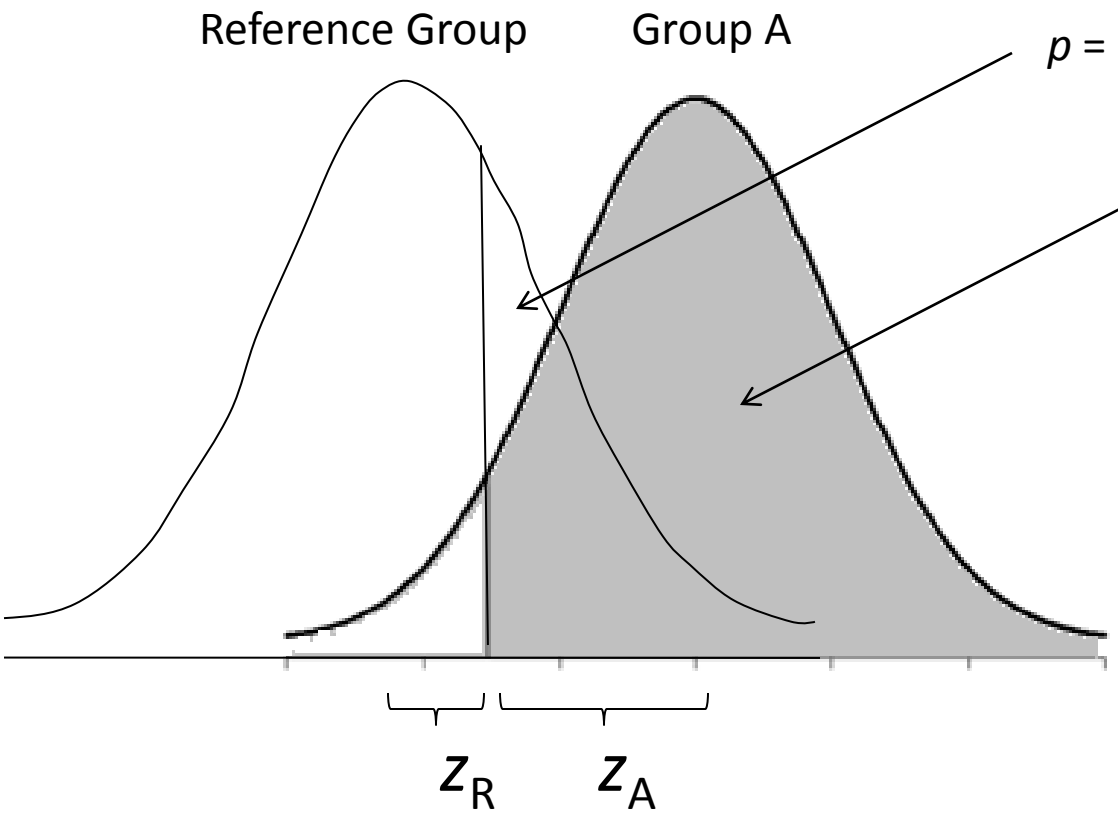
Item Difficulty



Item Difficulty: Thurstone Absolute Scaling

- Indices of item difficulty expressed as percentages or normal curve units are limited to the ability range covered by the sample from which they were obtained.
- There is need for a measure of item difficulty applicable across different samples varying in ability level.
- For example, age-difference, equivalent forms

Item Difficulty: Thurstone Absolute Scaling



Anchor Item

Item Difficulty: Thurstone Absolute Scaling

- The scale values of the same items in two (or more) group serve to define the relation between the groups and permit the transmutation of all item difficulty values from one group to another.

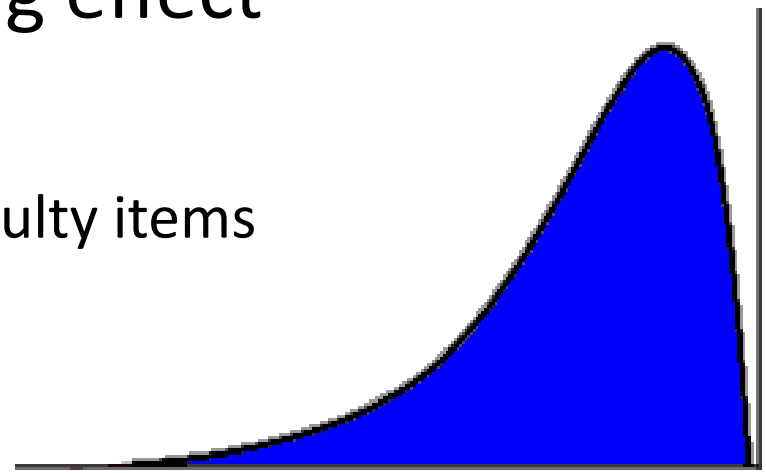
Item Difficulty

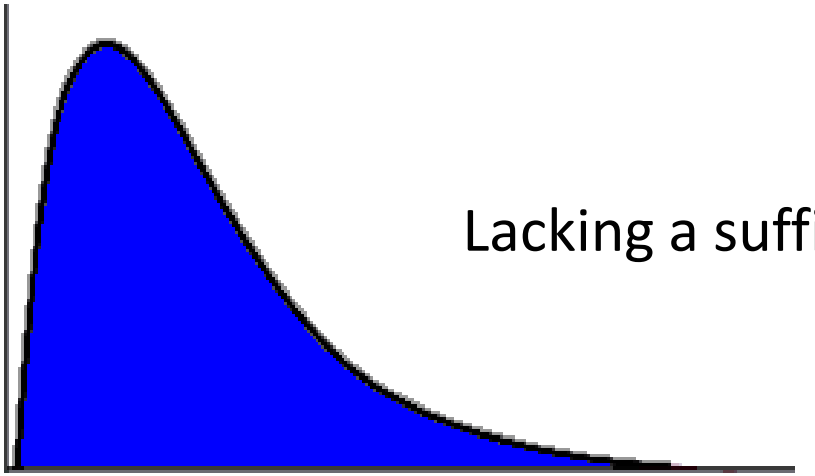
- The difficulty of the test as a whole is, directly dependent on the difficulty of the items that make up the test.
- Test difficulty for the population is checked by distribution of total scores.

Item Difficulty

- Floor effect and ceiling effect

Lacking a sufficient of difficulty items





Lacking a sufficient of easy items

Item Difficulty

- When the standardization sample yields a markedly nonnormal distribution on a test, the difficulty level of the test is ordinarily modified until a normal curve is approximated.
- Only in this way can the maximum differentiation between individuals at all ability levels be obtained with the test.

Item Difficulty

- In the construction of tests to serve special purposes, the choice of appropriate item difficulties, as well as the optimal form of the distribution of test scores, depends upon the type of discrimination sought.
- Accordingly, test designed for screening purposes should utilize items whose difficulty values come closet to the desired selection ratio.

Item Difficulty

- The choice of item difficulty in mastery testing should probably be at .80 or .90 level, in post training.
- The very easy items are the very items that would be included in a mastery test.
- For pre-training group, item difficulties should be low percentage.

Item Discrimination

- Item discrimination refers to the degree to which an item differentiates correctly among test takers in the behavior that the test is designed to measure.
- Which items should discriminate?
  - External Criteria (such as performance, mastery)
  - Internal Criteria (Total score)

Item Discrimination

- Statistical analysis that related to Item discrimination
  - Correlation (Pearson, Point-biserial, Biserial, Phi)
    - Corrected Item Total Correlation (CITC)
  - Regression
  - Mean difference (T-test)
    - Categorical Criterion
    - Exterme groups

Item Discrimination

- Statistical analysis that related to Item discrimination
  - Proportion difference (Chi-square, Fisher exact test)
  - Alpha if item deleted
- Although the numerical values of the indices may differ, the items that are retained and those that are rejected on the basis of different discrimination indices are largely the same.

Item Discrimination

- When using extreme groups,
  - The more extreme the groups, the sharper will be the differentiation.
  - The use of very extreme groups would reduce the reliability of the results because of the small number of cases utilized.
  - Kelly (1939) found that the optimum point is upper and lower 27%, in a normal distribution.
  - … found that this optimum point the rejected items are similar to those analyzed by CITC.

Item Discrimination

- Under certain conditions, the external and internal discrimination may lead to opposite results.
- When rejecting item that have low correlations with total score, this method will increase test validity only when original pool of items measures a single trait.

Item Discrimination

- When using regression techniques to relate external criterion, the most satisfactory items are those with the highest external validities and the lowest coefficients of internal consistency.
- Such a procedure, however, is neither feasible nor theoretically defensible.
- Interitem correlations are subject to wide sampling fluctuation.

Item Discrimination

- The resulting regression weights would be too unstable to provide a satisfactory basis for item selection.
- Its content will be heterogeneous that preclude meaningful interpretation of test score.

Item Discrimination

- For many testing purposes, a satisfactory compromise is to sort the relatively homogeneous items into separate tests or subtests, each of which covers a different aspect of the external criterion.

Item Discrimination

- The item discrimination indices are not independent of item difficulty but are biased in favor of intermediate difficulty levels.

Analysis of Distracters

- The analysis of distracters in multiple-choice may reveal some knowledge from test takers.
- The distracter that cannot deceive any test takers to choose may be changed. However, when changing distracters, the item difficulty and discrimination will change.

Item Response Theory

Item Analysis of Speeded Tests

- Whether or not speed is relevant to the function being measured, item indices computed from a speeded test may be misleading.
- The item indices found from a speed test will reflect the position of the item in the test rather than its intrinsic difficulty or discriminative power.

Item Analysis of Speeded Tests

- Item discrimination indices tend to be overestimated for those items that have not been reached by all test takers.
- To avoid some of these difficulties, we could limit the analysis of each item to those persons who have reached the them.
- This is not a completely satisfactory solution, because of shrinking number of cases, and selected sample.

Item Analysis of Speeded Tests

- The sample on which late-appearing item is analyzed is likely to consist of some very poor respondents, who guess, and a larger number of very proficient and fast respondents.
- One empirical solution is to administer the test with a long time limit to the group on which item analysis is to be carried out.
- This solution is satisfactory provided that speed itself is not an important aspect of the ability to be measured by the test.

Cross-Validation

- It is essential that test validity be computed on a different sample of persons from that on which items were selected.
- Any validity coefficient computed on the same sample that used for item-selection purposes (by external criterion) will capitalize on random sampling errors within that particular sample and will consequently be spuriously high.

Cross-Validation

- Classic example from Kurtz (1948): A research test on the Rorschach test for selection effective managers for insurance companies.
  - Item analysis group: correctly classified 70 of the 80 managers
  - Cross validate group: correctly classified 21 of the 41 managers

Cross-Validation

- The amount of shrinkage of a validity coefficient in cross-validation depends in part on the size of the original item pool and the proportion of items retained.
- When the number of original items is large and the proportion retained is small, there is more opportunity to capitalize on chance.
- Another condition affecting amount of shrinkage is size of sample

Cross-Validation

- If items are chosen on the basis of previously formulated hypotheses, validity shrinkage in cross-validation will be minimized.