SES'S L.S.RAHEJA COLLEGE OF ARTS AND COMMERCE

Course: PSYCHOLOGICAL TESTING & STATISTICS Unit: 1 OF SEM VI Prepared by: Ms. Shivani Chande

TEST DEVELOPMENT

The process of developing a test occurs in 5 stages:

I. TEST CONCEPTUALIZATION

1. **Stimulus** for test conceptualization can be anything:

- A review of the available literature on existing tests designed to measure a particular construct might indicate that such tests leave much to be desired in psychometric soundness.
- An emerging social phenomenon or pattern of behaviour might serve as the stimulus for the development of a new test.
- The development of a new test may be in response to a need to assess mastery in an emerging occupation or profession.

2. **Preliminary questions**- Regardless of the stimulus for developing the new test, a number of questions immediately confront the prospective test developer.

What is the test designed to measure? What is the objective of the test? Is there a need for this test? Who will use this test? Who will take this test? What content will the test cover? How will the test be administered? What types of responses will be required of testtakers? Who benefits from an administration of this test? Is there any potential for harm as the result of an administration of this test? Etc.

3. Different approaches to test development and individual item analyses are necessary, depending upon whether the finished test is designed to be **norm-referenced or criterion-referenced**.

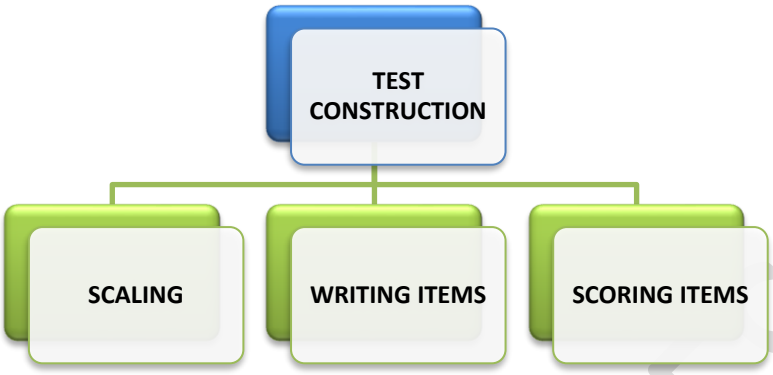
- A good item on a **norm-referenced achievement** test is an item for which high scorers on the test respond correctly. Low scorers on the test tend to respond to that same item incorrectly.
- Ideally, each item on a **criterion-oriented** test addresses the issue of whether the testtaker has met certain criteria. Criterion-referenced testing and assessment is commonly employed in licensing contexts, be it a license to practice medicine or to drive a car.

4. In the context of test development, terms such as **pilot work**, pilot study, and pilot research refer, in general, to the preliminary research surrounding the creation of a prototype of the test. In pilot work, the test developer typically attempts to determine how best to measure a targeted construct.

Pilot study may include:

- open-ended interviews with research subjects
- interviews with parents, teachers, friends, and others who know the subject
- physiological monitoring of the subjects (such as monitoring of heart rate) as a function of exposure to different types of stimuli.

II. TEST CONSTRUCTION



1. Scaling

Scaling may be defined as the process of setting rules for assigning numbers in measurement. In psychometrics, scales may also be conceived of as instruments used to measure a trait, a state, or an ability. Generally speaking, a testtaker is presumed to have more or less of the characteristic measured by a (valid) test as a function of the test score. The higher or lower the score, the more or less of the characteristic the testtaker presumably possesses. But how are numbers assigned to responses so that a test score can be calculated? This is done through scaling the test items.

A) **Types of Scales**- When we think of types of scales, we think of the different ways that scales can be categorized.

- nominal, ordinal, interval, or ratio
- age-based scale/ grade-based scale
- stanine scale
- unidimensional/multidimensional
- comparative/categorical

B) **Scaling Methods**-

- **Rating scale**- One type of summative rating scale is the **Likert scale**- Each item presents the testtaker with five alternative responses (sometimes seven), usually on an agree–disagree or approve–disapprove continuum. Rating scales can be unidimensional or multidimensional.
- **Method of paired comparisons**- Testtakers are presented with pairs of stimuli (two photographs, two objects, two statements), which they are asked to compare. They must select one of the stimuli

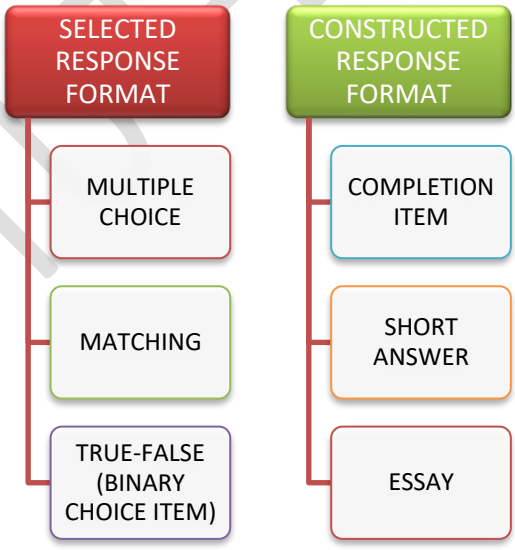
according to some rule. For each pair of options, testtakers receive a higher score for selecting the option deemed more justifiable by the majority of a group of judges.

- **Sorting tasks**:
  a) **comparative scaling-** entails judgments of a stimulus in comparison with every other stimulus on the scale.
  b) **Categorical scaling**- Stimuli are placed into one of two or more alternative categories that differ quantitatively with respect to some continuum.
- **Guttman scale**- Items on it range sequentially from weaker to stronger expressions of the attitude, belief, or feeling being measured. A feature of Guttman scales is that all respondents who agree with the stronger statements of the attitude will also agree with milder statements.
- **Method of equal-appearing intervals**- first described by Thurstone (1929), is one scaling method used to obtain data that are presumed to be interval in nature.

2. Writing Items

a) **Item Pool**- An item pool is the reservoir or well from which items will or will not be drawn for the final version of the test. When devising a standardized test using a multiple-choice format, it is usually advisable that the first draft contain approximately twice the number of items that the final version of the test will contain.

b) **Item format** – Variables such as the form, plan, structure, arrangement, and layout of individual test items are collectively referred to as item format. Two types of item format are the selected-response format and the constructed-response format. Items presented in a selected-response format require testtakers to select a response from a set of alternative responses. Items presented in a constructed-response format require testtakers to supply or to create the correct answer, not merely to select it.

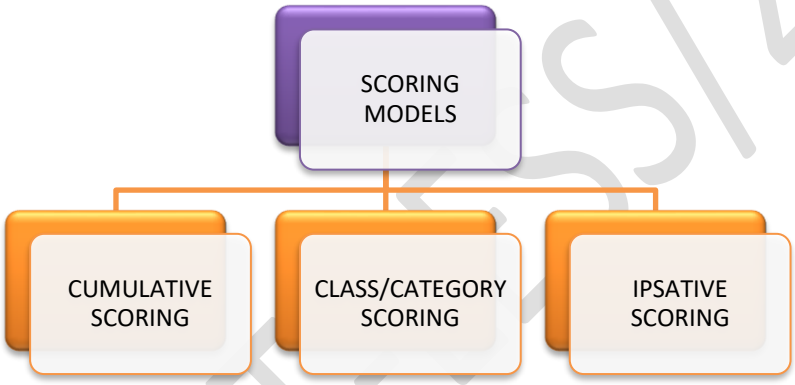


c) **Writing items for computer administration**.
A number of widely available computer programs are designed to facilitate the construction of tests as well as their administration, scoring, and interpretation. These programs typically make use of two advantages of digital media: the ability to store items in an item bank and the ability to individualize testing through a technique called item branching.

- **Item bank** - a relatively large and easily accessible collection of test questions.

- **Computerized Adaptive Testing (CAT)** - an interactive, computer-administered testtaking process wherein items presented to the testtaker are based in part on the testtaker's performance on previous items.
  CAT tends to reduce floor effects and ceiling effects.
  - A **floor effect** refers to the diminished utility of an assessment tool for distinguishing testtakers at the low end of the ability, trait, or other attribute being measured.
  - A **ceiling effect** refers to the diminished utility of an assessment tool for distinguishing testtakers at the high end of the ability, trait, or other attribute being measured.
- **Item Branching** - The ability of the computer to tailor the content and order of presentation of test items on the basis of responses to previous items. A computer that has stored a bank of test items of different difficulty levels can be programmed to present items according to an algorithm or rule.

3. Scoring Items



a) **Cumulative model** - the higher the score on the test, the higher the testtaker is on the ability, trait, or other characteristic that the test purports to measure.
b) **Class or category scoring** - testtaker responses earn credit toward placement in a particular class or category with other testtakers whose pattern of responses is presumably similar in some way.
c) **Ipsative scoring** - comparing a testtaker's score on one scale within a test to another scale within that same test. On the basis of such an ipsatively scored personality test, it would be possible to draw only intra-individual conclusions about the testtaker. (example: "John's need for achievement is higher than his need for affiliation.)

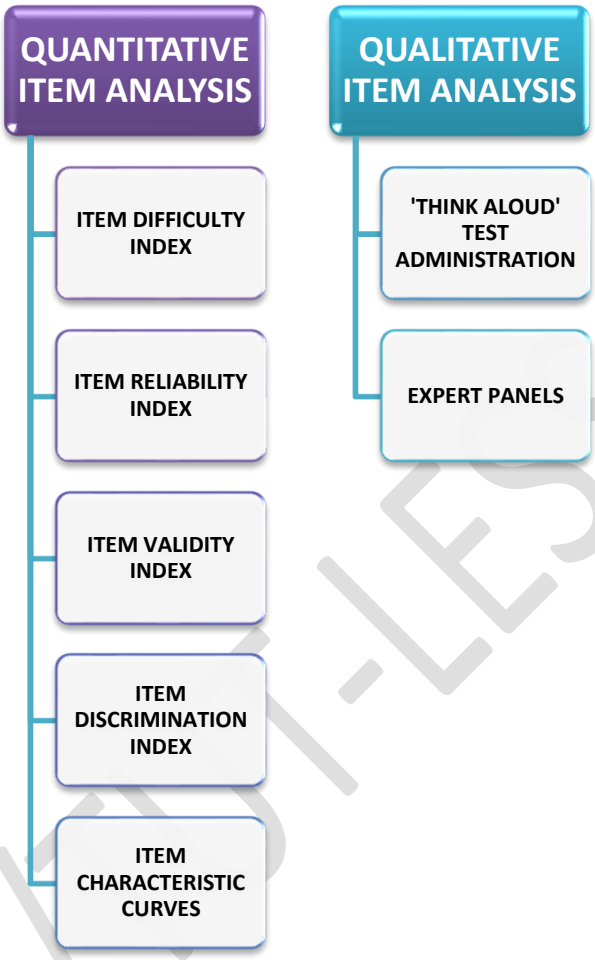
III. TEST TRYOUT

- Having created a pool of items from which the final version of the test will be developed, the test developer will try out the test.
- The test should be tried out on people who are similar in critical respects to the people for whom the test was designed.
- number of people on whom the test should be tried out - there should be no fewer than five subjects and preferably as many as ten for each item on the test. In general, the more subjects in the tryout the better.
- The more subjects employed, the weaker the role of chance in subsequent data analysis.
- The test tryout should be executed under conditions as identical as possible to the conditions under which the standardized test will be administered.

IV. ITEM ANALYSIS

After the first draft of the test has been administered to a representative group of examinees, the test developer analyses test scores and responses to individual items. The different types of statistical scrutiny that the test data can potentially undergo at this point are referred to collectively as item analysis.

Item analysis can be **quantitative or qualitative.**



1. Quantitative item analysis

a) **Item-Difficulty Index**- An index of an item's difficulty is obtained by calculating the proportion of the total number of testtakers who answered the item correctly. The statistic referred to as an item-difficulty index in the context of achievement testing may be an item-endorsement index in other contexts, such as personality testing.

b) **The item-reliability index** - provides an indication of the internal consistency of a test. A statistical tool useful in determining whether items on a test appear to be measuring the same thing(s) is factor analysis.

c) **The item-validity index** is a statistic designed to provide an indication of the degree to which a test is measuring what it purports to measure. The higher the item-validity index, the greater the test's criterion-related validity.

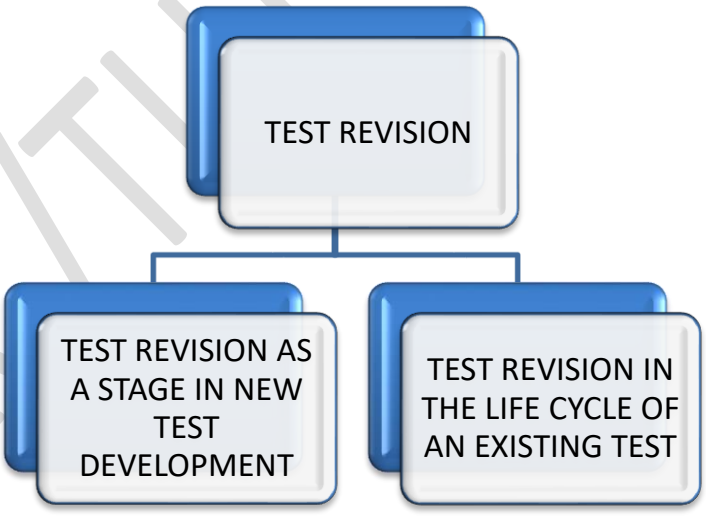
d) **The Item-Discrimination Index** - Measures of item discrimination indicate how adequately an item separates or discriminates between high scorers and low scorers on an entire test. The item-discrimination index is a measure of the difference between the proportion of high scorers answering an item correctly and the proportion of low scorers answering the item correctly.
e) **Item-Characteristic Curve** - is a graphic representation of item difficulty and discrimination.
f) Other consideration in item analysis:
    i. Guessing
    ii. Item fairness
    iii. Speed tests

2. Qualitative item analysis

Qualitative item analysis is a general term for various nonstatistical procedures designed to explore how individual test items work.

a) **"Think aloud" test administration** - An innovative approach to cognitive assessment entails having respondents verbalize thoughts as they occur. Cohen et al. (1988) proposed the use of "think aloud" test administration as a qualitative research tool designed to shed light on the testtaker's thought processes during the administration of a test.
b) **Expert panels** - expert panels may also provide qualitative analyses of test items. A sensitivity review is a study of test items, typically conducted during the test development process, in which items are examined for fairness to all prospective testtakers and for the presence of offensive language, stereotypes, or situations.

V. TEST REVISION



- **Cross-validation and co-validation**
  - The term **cross-validation** refers to the revalidation of a test on a sample of testtakers other than those on whom test performance was originally found to be a valid predictor of some criterion.
  - **co-validation** may be defined as a test validation process conducted on two or more tests using the same sample of testtakers. When used in conjunction with the creation of norms or the revision of existing norms, this process may also be referred to as **co-norming**.
- **Quality assurance during test revision**.
  - Mechanisms of quality assurance are put into place by test publishers in the course of standardizing a new test or restandardizing an existing test.
  - Quality control mechanisms are generally involved with examiners, protocol scoring, and data entry.
  - The examiner is the front-line person in test development, and it is critically important that examiners adhere to standardized procedures.
  - An **anchor protocol** is a test protocol scored by a highly authoritative scorer that is designed as a model for scoring and a mechanism for resolving scoring discrepancies. A discrepancy between scoring in an anchor protocol and the scoring of another protocol is referred to as scoring drift
  - Once protocols are scored, the data from them must be entered into a database. For quality assurance during the data entry phase of test development, test developers may employ computer programs to seek out and identify any irregularities in score reporting.

THE USE OF IRT IN BUILDING AND REVISING TESTS:

a) Evaluating the properties of existing tests and guiding test revision
b) Determining measurement equivalence across testtaker populations
c) Developing item banks