Lecture 3 Psychological Testing and Measurement

Sunthud Pornprasertmanit

Scaling and Process of Test Construction

Psychological Scaling

- The development of systematic rules and meaningful unit of measurement for quantifying empirical observations.
  - Such as centimeters, inches, feet

Psychological Scaling

- Three broad approaches for psychological scaling
  - Subject-centered methods
  - Stimulus-centered methods
  - Response-centered methods

Psychological Scaling

- Subject-Centered Methods
  - Locating individual in continuum
  - Such as Likert approach
  - Factor Analysis (When using summate scaling)

Psychological Scaling

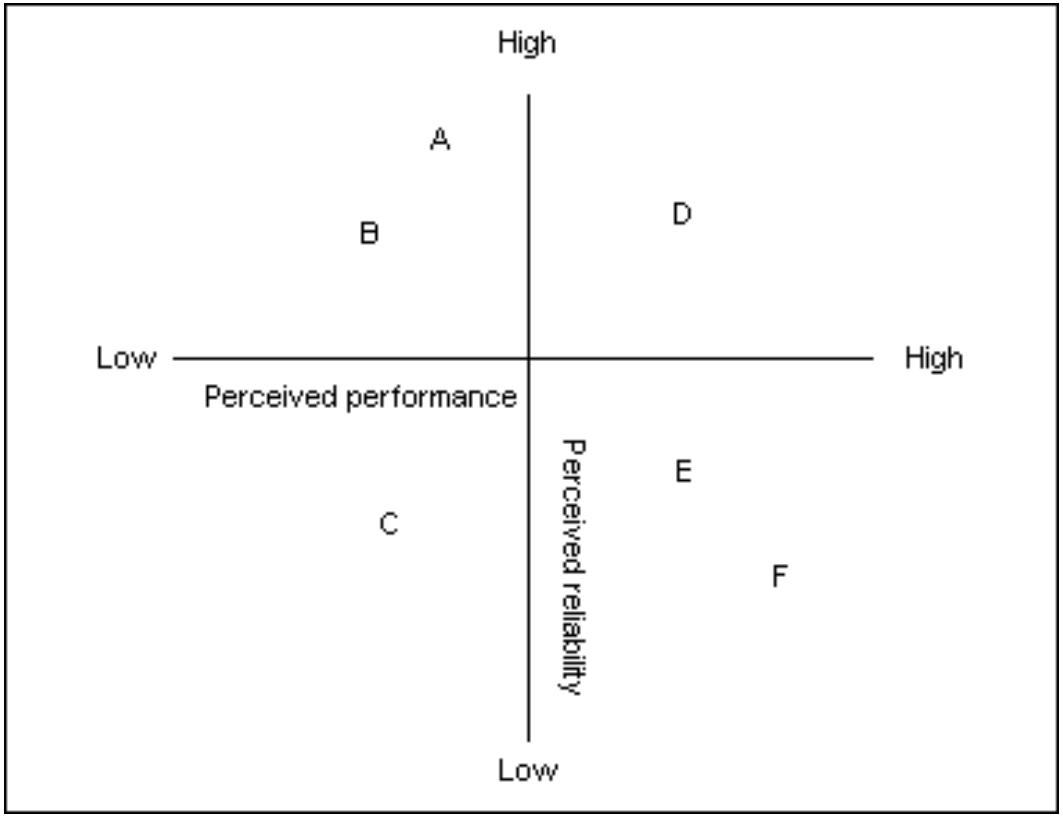
- Stimulus-Centered Methods
  - Locating stimulus in continuum
  - Such as measure perception of brightness
    - Just Noticeable Difference (JND)
  - Thurstone Measurement of Attitude
  - Multidimensional Scaling

Psychological Scaling

- Thurstone Measurement of Attitude
  - 1.00 ฉันจะแต่งงานกับคนนี้
  - 1.80 ฉันยอมรับคนนี้เป็นเพื่อนสนิท
  - 2.80 ฉันไว้ใจคนนี้
  - 3.50 ฉันจะร่วมมือกับคนนี้
  - 4.20 ฉันยอมรับคนนี้เป็นเพื่อนคุย
  - ...
  - 11.00 ฉันไม่อยากได้ยินชื่อของคนนี้

Psychological Scaling

- Multidimensional Scaling



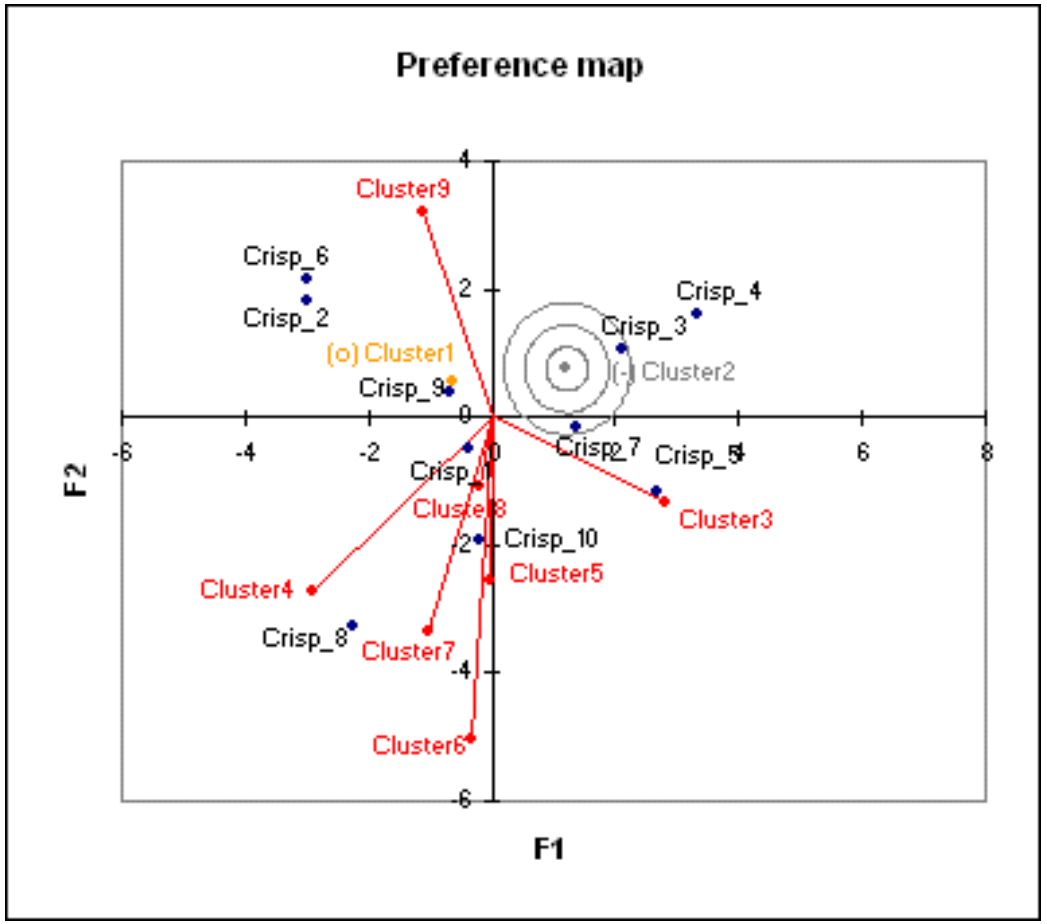
Psychological Scaling

- Response-Centered Methods
  - Locating individual and stimulus in continuum
  - Such as unfolding techniques

    Multidimensional Analysis of Preference

Psychological Scaling

- Multidimensional Analysis of Preference



Psychological Scaling

- At the moment, the subject-centered methods can quantify the relative weight of stimuli. (e.g. by factor analysis)
- The stimuli-centered methods can quantify the relative position of individual (e.g. by item response theory)
- Therefore, the classification is vague.

Meaning of Test Score

- Norm Referenced Test (NRT)
- Criterion Referenced Test (CRT)

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.

Process of Test Construction

1. Identifying the primary purpose(s) for which the test score will be used
2. Identify behaviors that represent the construct or define the domain
3. Prepare test specifications
4. Construct initial pool of items
5. Have items reviewed
6. Hold preliminary item tryouts
7. Field-test items with large sample

Process of Test Construction

8. Determined statistical properties of test scores
9. Design and conduct reliability and validity
10. Develop guidelines for administration
11. (For NRT) Develop norms of the test
    (For CRT) Making cutoff

Process of Test Construction

0. Identifying construct
- Definition
- Relationship to related construct
- Specifying the type of construct
  - Maximal Performance
  - Typical Performance

Process of Test Construction

1. Identifying purposes of test score use
   - Admission
   - Placement
   - Diagnostic Decisions
   - Research

Process of Test Construction

2. Identifying behavior to represent the construct

- Not only think up
- There are many useful ways to help identifying behaviors
- Content Analysis
- Review of Research
- Critical Incidents
- Direct Observations

Process of Test Construction

2. Identifying behavior to represent the construct

- Not only think up
- There are many useful ways to help identifying behaviors
- Expert Judgments
- Instruction Objectives

Process of Test Construction

2.1 Interpreting Scores

- Norm-Referenced Test
  - Use high discriminative items
  - Define the domain or dimensionality
- Criterion-Referenced Test
  - Items discriminate between cutoff
  - Define the domain

Process of Test Construction

3. Test Specification

- Define the relative emphasis of each domain
- In addition, categorizing cognitive operations

Knowledge
Comprehension
Application

Analysis
Synthesis
Evaluation

Process of Test Construction

3. Test Specification

- Table of specification

	Knowledge	Comprehension	Application	Totals
Content 1	1	10	9	20
Content 2	1	11	11	23
Content 3	1	6	2	9
Content 4	1	12	5	18
Totals	4	39	27	70

Process of Test Construction

4. Item Format

- Selecting an appropriate item format
- Verifying that the proposed format is feasible for intended examinees
- Selecting and training the item writers
- Writing items
- Monitoring

Process of Test Construction

1. Maximal Performance
   - Dichotomous format
   - Multiple choice
   - Matching
   - Blank
   - Essay
2. Typical Performance
   - Likert Format
   - Semantic Differential Format
   - Forced Choice
   - Q-sort

Process of Test Construction

4. Item Format

- Guidelines for rating scale
  - มีลักษณะสอดคล้องกับลักษณะเป้าหมาย
  - ถามถึงลักษณะในปัจจุบัน
  - หลีกเลี่ยงข้อคำถามที่แสดงถึงลักษณะความเป็นจริง
  - หลีกเลี่ยงข้อคำถามที่กำกวม
  - หลีกเลี่ยงข้อคำถามที่แปลความหมาย ได้มากกว่าหนึ่งความหมาย
  - พยายามให้ข้อคำถามสั้น กระชับ
  - ไม่ใช้คำเชิงคุณภาพ เช่น มาก ปานกลาง น้อย ในข้อคำถาม
  - หลีกเลี่ยงประโยคปฏิเสธซ้อนปฏิเสธ

Process of Test Construction

4. Item Format

- Individual/Group Test
- Test Length
- Item Difficulty

Process of Test Construction

5. Item Review

- Should consider
  - Accuracy
  - Appropriateness or relevance to test specifications
  - Technical item-construction flaws
  - Grammar
  - Offensiveness or appearance of 'bias'
  - Level of readability

Process of Test Construction

5. Item Review
- Using expert or panel of experts
- Various types of experts
  - Test Construction
  - Domain of the Test
  - (Optional) Readability Level (such as test takers are children)
- Reviewing Problematic Items

Process of Test Construction

6. Preliminary Item Tryout

- Small samples of test takers
- It is possible to administer subsets of items
- Informal
- Observe reactions
- After testing, Debriefing and providing the opportunities to making comments
- Examination of descriptive statistics

Process of Test Construction

Next Steps (different between NRT and CRT)

- Item Analysis
- Reliability and Validity
- Test Manual
- Making Norm or Cutoff (Standards)

Exercise 2.1 – 2.4

- Mechanical Comprehension.sav
- Outline Lecture 2 from educational year 50

Case

3.2 Consulting project on performance assessment

HW Developing performance assessment sheet

Exercise 4.1 – 4.2

- Defining a personality trait
- Item Writing

Next Lecture Reliabilty

Lecture 3 Psychological Testing and Measurement
Sunthud Pornprasertmanit

What is Reliability?

- Reliability refers to the consistency of scores obtained by the same persons when they are reexamined with the same test on
  - Different occasions
  - Different sets of equivalent items
  - Under other variable examining conditions

What is Reliability?

- The concept of reliability underlies the computation of the error of measurement
- We can predict the range of fluctuation likely to occur in a single individual's score as a result of irrelevant or unknown chance factors.

What is Reliability?

$$O = T + E$$

- O = Observed score (Individual differences by test
- T = True score (Real individual differences)
- E = Error of measurement

What is Reliability?

- Because *E* is chance factor, it does correlate with *T*.

$$\sigma_O^2 = \sigma_T^2 + \sigma_E^2$$

- Therefore, variance of observed score is the sum of variance of true score and error of measurement variance

What is Reliability?

- Test reliability indicates the extent to which individual differences in test scores are attributable to "true" difference.

$$r_{xx} = \frac{\sigma_T^2}{\sigma_O^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2}$$

What is Reliability?

- Any condition that is irrelevant to the purpose of the test represents error variance: test taking time, rapport, instructions etc.
- Factors that might be considered error variance for one purpose would be classified under true variance for another.

What is Reliability?

- Such a measure of reliability characterizes the test when it is administered under standard conditions and given to persons similar to those constitute the normative sample.

Type of Reliability

Type of Reliability Coefficient	Error variance
Test-retest Reliability	Time sampling
Alternate-Form (Immediate)	Content sampling
Alternate-Form (Delayed)	Time and content sampling
Split-Half	Content sampling
KR and Coefficient Alpha	Content Heterogeneity
Scorer Reliability	Interscorer differences

Standard Error of Measurement

- Standard error of measurement is standard deviation of error scores.
- The more reliability coefficient, the less standard error of measurement.
- Computed by:

$$SEM = SD_t \sqrt{1 - r_{tt}}$$

Standard Error of Measurement

- Standard error of measurement can be used for true score estimate (by confidence interval)

$$CI_{\%} = X \pm z_{\%/2} SEM$$

Standard Error of Measurement

- Unlike the reliability coefficient, the error of measurement is independent of the variability of the group on which it is computed.
- However, SEM cannot be directly comparable from test to test.