Test Item Considerations

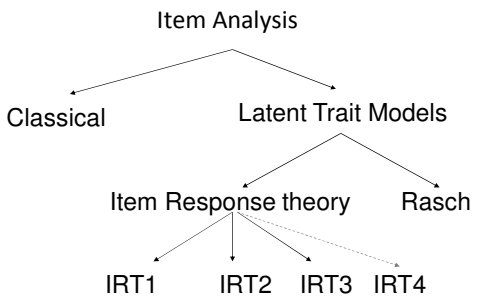
Item Analysis

The reliability of test scores and the validity of the interpretation of test scores are dependent on the quality of the items in the test

Item analysis is the general term for all the techniques used to assess test items during development and construction

- Contains both qualitative and quantitative procedures

What kinds of item analysis are there?



Test Development

Test users need to be familiar with how tests are developed to assist in evaluation

- Appropriateness for certain settings / persons
- Ease of administration and scoring
- Time involved
- Training needed

Test Development

Developing a test is a costly, time-consuming process

When planning a test, must be aware of

- Constructs that will be assessed
- Population of intended use
- Objectives of particular items within the testing framework
- Means which behavior samples will be gathered and scored

Steps in Test Development

1. Generating item pool, administration, and scoring procedures
2. Submitting item pool to qualitative analyses and then making revisions
3. Testing items on sample population
4. Evaluating results of trail administration through quantitative and qualitative analyses

Steps in Test Development

5. Adding / deleting / modifying items as needed based on step 4
6. Conduct additional trial administrations until obtaining satisfactory item set
7. Standardizing sequence and length of items, administration, and scoring procedures
8. Developing normative data set
9. Publishing test along with administration and scoring manual

Test Item Types

Immense variety of possible types of test items

- Can differ in terms of content, format, medium, scoring manner, and processing requirements

Can distinguish items based on types of responses

- Selected-response items
- Constructed-response items

Selected-response Items

Close-ended, present a limited number of alternatives which taker must choose from

Multiple-choice, true/false, ranking, matching

Dichotomous (two) and polytomous (three or more) formats

Forced-choice items require the user to choose which alternative is most/least characteristic of them

Selected-response Items

Advantages

- Easy to score – results in time savings and increased reliability
- Time-efficient
- Individual or group testing abilities
- Easy qualitative analysis of items

Selected-response Items

Disadvantages

- Can guess at answers – introduces possible error into test scores
- With personality testing, goals can be subverted
  - Random/careless responding or misleading responses
- Difficult and time-consuming to prepare test items
- Less flexible than CRI in terms of possible range of responses

Constructed-Response Items

Open-ended responses

Most common type is “fill in the blanks”

Must include thorough instructions on time limits, medium/manner/length of answers, and permitted materials

Includes interviews, biographical data questionnaires, behavioral observations, and projective techniques

Constructed-Response Items

Advantages

- Can provide rich samples of behavior
- Offer wider range of possible answers
- Elicit authentic samples of behavior, rather than choices among prepackaged alternatives

Constructed-Response Items

Disadvantages

- Reliability and validity are major concerns
- Complex and time-consuming to score due to degree of subjectivity involved
- Shorter test length results in higher content sampling errors
- Response length varies, which causes further reliability and validity difficulties

Item Analysis

Item validity is usually the most important quality for psych tests

Stats that measure this are indexes of *item discrimination*

- Includes item difficulty and item fairness

Role of Item Difficulty

Difficulty of a test item is a function of both the item and the taker

Thus, indexes of *relative difficulty* are needed for different groups of test takers

Allows determination of how appropriate test items are, as well as where to place the item in a test

Measuring Item Difficulty

First, use objective standards during item creation

- E.g., frequency of words, complexity of mathematical operations

Once items are created, use quantitative indexes of difficulty

Measuring Item Difficulty

Use proportion passing, or *p*

$$p = \frac{\text{\# of Subjects with Correct Answer}}{\text{Number of Subjects}}$$

Ranges from 0.0 to 1.0

Large *p* indicates an easy item; a small *p* indicates a difficult item

Item Difficulty Index

Items that are correctly answered by every subject (p = 1.0) and items that are missed by every subject (p = 0.0) provide no information about individual differences and are of no value from a measurement perspective

For maximizing variability and reliability, the optimal item difficulty value is 0.5

- Too many difficult items and few subjects get them correct resulting in reduced variability; too many easy questions and most subjects get them correct, reducing variability

Optimal p Values for Items with Varying Numbers of Choices

Number of Choices	Optimal Mean p Value
2 (e.g., true/false)	.85
3	.77
4	.74
5	.69
Constructed Response (e.g., essay)	.50

Lord (1952)

Desirable p Values

In practice, a general recommendation is to use items with p values with a range of approximately 0.20 around the optimal values

For example, if your optimal p value is 0.50, you would select items ranging from 0.40 to 0.60, with a mean of 0.50

Distractors and Difficulty

Incorrect alternatives can have major influence on item difficulty

Higher numbers of distractors = lower probability of guessing correctly

Quality of distractors also plays a role

- Correct alternative should be obvious to takers who know the answer, while all alternatives should be plausible to someone who doesn't

For Example

What year did Einstein first publish his full general theory of relativity?

a) 1910	a) 1655
b) 1912	b) 1762
c) 1914	c) 1832
d) 1916	d) 1916
e) 1918	e) 2001

Item Discrimination

The extent to which items cause responses that differentiate test takers in terms of what the test evaluates

Criteria used to determine may include

- Internal criteria
- External criteria
- Combinations of both

Item Validation Criteria

Validation criteria chosen depend on the purpose of the test

- Ability tests require content/skill areas criteria
- Personality tests require traits/behaviors criteria

External validation increases validity of test scores as a whole

Internal validation increases homogeneity

Discrimination Stats

Requires information on item performance and criterion standing

Traditionally uses *D* (index of discrimination) and correlations

D = difference in percentage of test takers in the upper and lower groups

- Ranges from +100 to -100

D & *r* Example

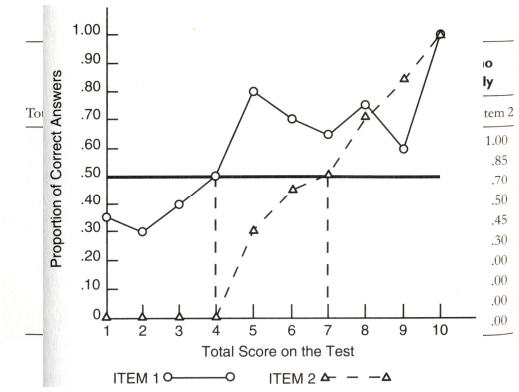
	Percentage Passing (*p* value)				
Item Number	Total Group	Upper[a] Group	Lower[a] Group	*D* index (Upper – Lower)	Point Biserial Correlation (r_{pb})[b]
1	100%	100%	100%	0	0.00
2	88%	100%	50%	50	0.67
3	38%	100%	0%	100	0.63
4	75%	50%	50%	0	0.13
5	75%	50%	100%	–50	–0.32
6	13%	50%	0%	50	0.43

Speed Tests

In closely timed tests, *p* and *D* values are a function of their position within the test

This happens because

- Fewer test takers get to the latter items
- Those who do tend to be high scorers



Item-Response Theory

General name for a group of models used to design, develop, and evaluate tests

Less widely used that classical test theory due to significant assumptions and more extensive data collection needed in IRT

CTT focuses on test as a whole, IRT focuses on responses to individual items

Item-Response Theory

Goals are to

- Generate items that provide maximum information about examinees
- Give examinees items tailored to their level
- Reduce the number of items needed while minimizing measurement error

CTT v. IRT

CTT is group-dependent, while IRT methods are invariant and provide a uniform measurement scale across groups

In CTT, the scores are test-dependent, while in IRT scores are independent of particular item set administered

In IRT, score reliability is more precisely generated via computerized adaptive testing

CTT v. IRT

Classical analysis has the test (not the item) as its basis

- Although the statistics generated are often generalised to similar persons taking a similar test; they only really apply to *those* persons taking *that* test

Latent trait models aim to look beyond that at the underlying traits which are producing the test performance

- They are measured at item level and provide sample-free measurement

Essentials of IRT

Unidimensional models of IRT assume

- Items comprising a test / test segment measure only a single trait
- Item responses of examinees depends only on their standing on the trait

Aim to measure the underlying ability (or trait) which is producing the test performance rather than measuring performance *per se*

As the statistics are not dependant on the test situation which generated them, they can be used more flexibly and are thus **sample-free**

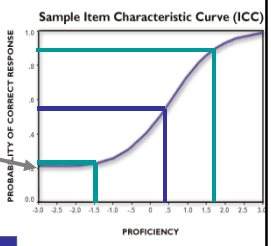
Essentials of IRT

IRT attempts to predict traits based on responses (unobservable vs. observable), and are evaluated based on those predictions

Data must be gathered from samples that differ on ability/trait to be assessed

Item Characteristic Curves

An ICC is a plot of the candidates ability over the probability of them correctly answering the questio[illegible] higher the [illegible] higher the chance that they will respond c[illegible]



c - intercept

b - ability at max (a)

a - gradient

Difficulty

Although there is no "correct" difficulty for any one item, it is clearly desirable that the difficulty of the test is centred around the average ability of the takers

The higher the "b" parameter the more difficult the question - this is inversely proportionate to the probability of the question being answered correctly

Discrimination

In IRT, maximal discrimination is sought

- the higher the "a" parameter, the more desirable the question

Differences in the discrimination of questions can lead to differences in the difficulties of questions across the ability range

Guessing

A high "c" parameter suggests that candidates with very little ability may choose the correct answer

This is rarely a valid parameter outwith multiple choice testing…and the value should not vary excessively from the reciprocal of the number of choices

Item Fairness

Qualitative methods for ensuring fairness include removing stereotypes, offensive content, and representation of diverse subgroups in materials

Qualitative methods include examination of *differential item functioning*

- Compares item difficulty and discrimination across different types of groups

Test Usage

Should I Use?

What information are you seeking?

How will it be used?

How much of it is available from other sources?

What other ways might I get the information?

What are the advantages or disadvantages of testing rather than other methods?

No!

- Purpose of testing is unclear to user
- User is not familiar with the test procedures and documentation
- User is unaware of where the results will go or how they will be used
- Information is already available or more easily gained through other methods
- Taker is unwilling or likely to be harmed

No!

- Environmental conditions are inadequate
- Test is inappropriate based on taker's demographic or other factors
- Test norms are outdated, inadequate, or inapplicable
- Reliability and validity of test is unknown or inadequate

Yes!

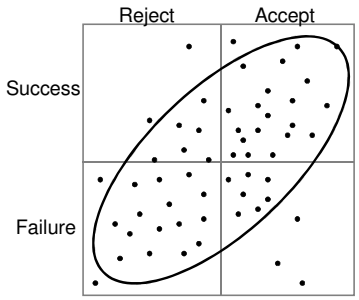
- Test is efficient and psychometrically sound for required information
- Objective means of data gathering is needed
- User is qualified to give and interpret the test
- Test is applicable to the taker

Utility

Stems from the value and priorities of decision makers

Which is most important quadrant changes depending on what is being tested for

Hidden Data



This shows all outcomes, both for accepted and rejected employees

How would our criteria of utility change if we were looking at suicide prediction?

Other Sources

Biodata / life-history

Interviews

Observation

Choosing a Test

Does it have adequate psychometrics?

Is it appropriate for the test taker?

Does it answer my question?

Is the benefit greater than the cost?

Administration

Testing environments should be

- Quiet
- Well lit
- Comfortable temperature
- Appropriate seating / tables
- Free of other stimuli
- As similar as possible to conditions where test was standardized
- Private – only the tester and the testee

Administration

Obtaining informed consent is the first step

Rapport needs to be established prior to testing being conducted

- Lack of rapport will negatively impact test scores

The taker's test anxiety and test sophistication are also highly influential