PRINCIPLES OF LANGUAGE TESTING

Acep Haryudin, M.Pd

Overview

- What are the principles of language testing?
- How can we define them?
- What factors can influence them?
- How can we measure them?
- How do they interrelate?

Reliability

Related to accuracy, dependability and consistency e.g. 20°C here today, 20°C in North Italy – are they the same?

According to Henning [1987], reliability is

- a measure of accuracy, consistency, dependability, or fairness of scores resulting from the administration of a particular examination e.g. 75% on a test today, 83% tomorrow – problem with reliability.

Validity: internal & external

Construct validity [internal]

- the extent to which evidence can be found to support the underlying theoretical construct on which the test is based

Content validity [internal]

- the extent to which the content of a test can be said to be sufficiently representative and comprehensive of the purpose for which it has been designed

Validity [2]

Response validity [internal]

- the extent to which test takers respond in the way expected by the test developers

Concurrent validity [external]

- the extent to which test takers' scores on one test relate to those on another externally recognized test or measure

Validity [3]

Predictive validity [external]

- the extent to which scores on test Y predict test takers' ability to do X e.g. IELTS + success in academic studies at university

Face validity [internal/external]

- the extent to which the test is perceived to reflect the stated purpose e.g. writing in a listening test – is this appropriate? depends on the target language situation i.e. academic environment

KKM (Kriteria Ketuntasan Minimal/Minimum)

Inteks Siswa (3, 2, 1)
Kompleksitas (1, 2, 3)
Daya Dukung/ Fasiltas (3, 2, 1)

Inteks 3
Kompelksitas 3
Daya Dukung 3
KKM?? 3+3+3= 9/3 = 3 (85-100)

Validity [4]

- 'Validity is not a characteristic of a test, but a feature of the inferences made on the basis of test scores and the uses to which a test is put.'

Alderson [2002: 5]

Practicality

The ease with which the test:

- items can be replicated in terms of resources needed e.g. time, materials, people
- can be administered
- can be graded
- results can be interpreted

Factors which can influence reliability, validity and practicality.

Test [1]

- quality of items
- number of items
- difficulty level of items
- level of item discrimination
- type of test methods
- number of test methods

1. Susan ……………… to Bandung everyday.

a. Going

b. Goes

c. Gone

d. Went

which animal lives in the water?

a. Elephant, zebra, and Crocodile
b. Shark, fish and crocodile
c. Shark and Fish
d. Lion, Spider and Butterflies

From the image above, the …. Animal is the elephant.

a. Most

b. Large

c. Bigges

d. high

These animals are wild animals.
These animals are :
A. elephants and butterfly
B. lion and crocodile .
C. Zebra and bird
D. spider and butterfly

Based on the type of food, from the picture above elephants and zebras are included in the type of animal...

a. carnivore
b. herbivore
c. insektivore
d. omnivore

both these professions are engaged in entertainment...

a. teacher and clown,
b. clown and musician,
c. businesman and musician,
d. doctor and clown

Which job is most often outdoors?

a. basketball athlete, doctor, teacher
b. architect, pilot, soldier
c. musician, architect, clown
d. basketball athlete, pilot, architect

From the picture above, the profession of a teacher is included in the profession related to...

a. medical
b. art
c. education
d. technique

Dr. R. Green, Aug 2006

Getty/indianexpress.com

Test [2]

- time allowed
- clarity of instructions
- use of the test
- selection of content
- sampling of content
- invalid constructs

Test taker

- familiarity with test method
- attitude towards the test i.e. interest, motivation, emotional/mental state
- degree of guessing employed
- level of ability

Test administration

- consistency of administration procedure
- degree of interaction between invigilators and test takers
- time of day the test is administered
- clarity of instructions
- test environment – light / heat / noise / space / layout of room
- quality of equipment used e.g. for listening tests

Scoring

- accuracy of the key e.g. does it include all possible alternatives?
- inter-rater reliability e.g. in writing, speaking
- ***intra-rater*** reliability e.g. in writing, speaking
- machine vs. human

intra-rater reliability e.g. in writing, speaking

Fina (Test maker)	Fahlevi (Test maker)
Speaking (Skill)	Speaking (Skill)
Nadia (Test Taker)	Nadia (Test Taker)
score 4, 3, 2, 1	score 4, 3, 2, 1
4	4+3=7 /2 = **3,5**

How can we measure reliability?

Test-retest

- same test administered to the same test takers following an interval of no more than 2 weeks

Inter-rater reliability

- two or more independent estimates on a test e.g. written scripts marked by two raters independently and results compared

Measuring reliability [2]

Internal consistency reliability estimates e.g.

- Split half reliability
- Cronbach’s alpha / Kuder Richardson 20 [KR20]

Split half reliability

- test to be administered to a group of test takers is divided into halves, scores on each half correlated with the other half
- the resulting coefficient is then adjusted by Spearman-Brown Prophecy Formula to allow for the fact that the total score is based on an instrument that is twice as long as its halves

Cronbach's Alpha [KR 20]

- this approach looks at how test takers perform on each individual item and then compares that performance against their performance on the test as a whole
- measured on a -1 to +1 scale like discrimination

Reliability is influenced by …..

- the longer the test, the more reliable it is likely to be [though there is a point of no extra return]
- items which discriminate will add to reliability, therefore, if the items are too easy / too difficult, reliability is likely to be lower
- if there is a wide range of abilities amongst the test takers, test is likely to have higher reliability
- the more homogeneous the items are, the higher the reliability is likely to be

How can we measure validity?

According to Henning [1987]

- non-empirically, involving inspection, intuition and common sense
- empirically, involving the collection and analysis of qualitative and quantitative data

Construct validity

- evidence is usually obtained through such statistical analyses as factor analysis [looks for items which group together], discrimination; also through retrospection procedures

Content validity

- this type of validity cannot be measured statistically; need to involve experts in an analysis of the test; detailed specifications should be drawn up to ensure the content is both representative and comprehensive

Response validity

- can be ascertained by means of interviewing test takers [Henning]; asking them to take part in introspection / retrospection procedures [Alderson]

Concurrent validity

- determined by correlating the results on the test with another externally recognised measure. Care needs to be taken that the two measures are measuring similar skills and using similar test methods

Predictive validity

- can be determined by investigating the relationship between a test taker's score e.g. on IELTS/TOEFL and his/her success in the academic program chosen
- problem - other factors may influence success e.g. life abroad, ability in chosen field, peers, tutors, personal issues, etc.; also time factor element

Reliability vs. validity?

- 'an observation can be reliable without being valid, but cannot be valid without first being reliable. In other words, reliability is a necessary, but not sufficient, condition for validity.'

  [Hubley & Zumbo 1996]

- ‘Of all the concepts in testing and measurement, it may be argued, validity is the most basic and far-reaching, for without validity, a test, measure or observation and any inferences made from it are meaningless’

  [Hubley & Zumbo 1996, 207]

Reliability vs. validity [2]

- even an ideal test which is perfectly reliable and possessing perfect criterion-related validity will be invalid for some purposes

[Henning 1987]

Practicality

Designing and developing good test items requires

- working with other colleagues
- materials i.e. paper, computer, printer etc.
- time

Some items look very attractive but this attraction has to be weighed against these factors.

References

- Alderson, J. C 2002 *Conceptions of validity and validation*. Paper presented at a conference in Bucharest, June 2002.
- Angoff, 1988 Validity: An evolving concept. In H. Wainer & H. Braun [Eds.] *Test validity* [pp. 19-32], Hillsdale, NJ: Erlbaum.
- Bachman, L. F. 1990 *Fundamental considerations in language testing*. Oxford: O.U.P.
- Cumming A. & Berwick R. [Eds.] *Validation in Language Testing* Multilingual Matters 1996
- Hatch, E. & Lazaraton, A. 1991 *The Research Manual - Design & Statistics for Applied Linguistics* Newbury House

References [2]

- Henning, G. 1987 *A guide to language testing: Development, evaluation and research* Cambridge, Mass: Newbury House
- Hubley, A. M. & Zumbo, B. D. A dialectic on validity: where we have been and where we are going. *The Journal of General Psychology 1996. 123[3] 207-215*
- Messick, S. 1988 The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun [Eds.] *Test validity* [pp. 33-45], Hillsdale, NJ: Erlbaum.
- Messick, S. 1989 Validity. In R. L. Linn [Ed.] *Educational measurement*. [3rd ed., pp 13-103]. New York: Macmillan.

Item-total Statistics

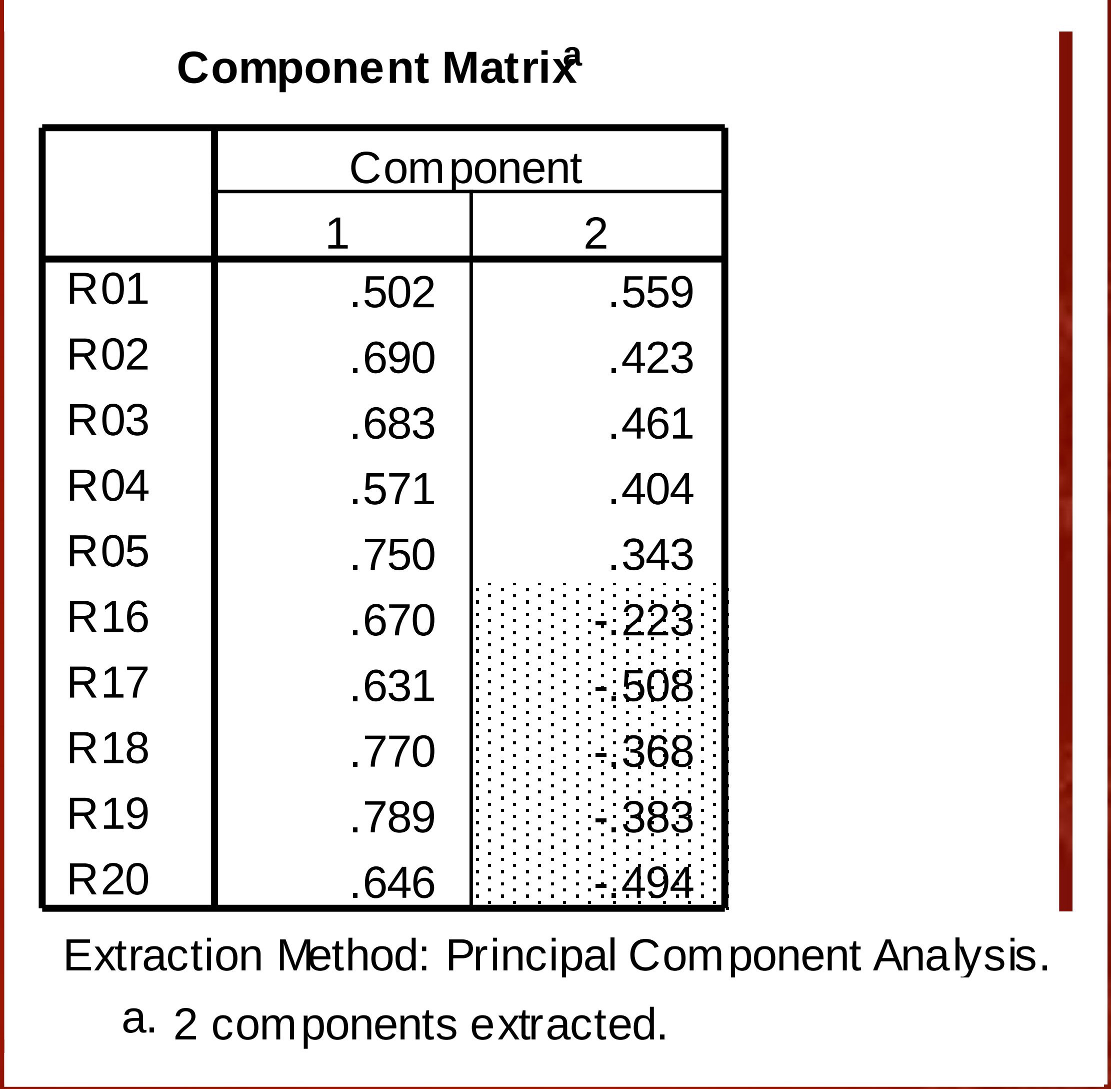
	Corrected Item-Total Correlation	Alpha if Item Deleted
R01	.5259	.7964
R02	.6804	.7594
R03	.6683	.7623
R04	.5516	.7940
R05	.7173	.7489
R16	.3946	.8288

N of Cases = 194.0 N of Items = 6 Alpha = .8121

Item-total Statistics

	Corrected Item Total Correlation	Alpha if Item Deleted
R16	.5773	.7909
R17	.5995	.7863
R18	.7351	.7553
R19	.7920	.7419
R20	.6490	.7753
R01	.1939	.8663

N of Cases = 194.0 N of Items = 6 Alpha = .8185



Component Matrix[a]

	Component	
	1	2
R01	.502	.559
R02	.690	.423
R03	.683	.461
R04	.571	.404
R05	.750	.343
R16	.670	-.223
R17	.631	-.508
R18	.770	-.368
R19	.789	-.383
R20	.646	-.494

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Thank you

for your attention!