



Scales Development – Reliability and Validity.

DR CHARLOTTE GREENWAY WITH SUPPLEMENTARY INFORMATION BY DR
STUART JONES.



Components of Attitude

What is an attitude?

“a psychological tendency that is expressed by evaluating a particular entity with some degree of favour or disfavour”

(Eagly & Chaiken, 1993)



Attitude Scales





Types of Scales



Types of Scales

Question A: Are you taller than 1.70 m?	yes/no
Question B: Are you taller than 1.80 m?	yes/no
<i>Arithmetic Ability</i>	
Question A: Did the student give the right answer to $2 + 2$?	yes/no
Question B: Did the student give the right answer to $23 * 17$?	yes/no
<i>Political Ideology</i>	
Question A: Should income differences be decreased?	yes/no
Question B: Should income differences be abolished?	yes/no
<i>Religious Belief</i>	
Question A: Do you believe in heaven?	yes/no
Question B: Do you believe in hell?	yes/no
<i>Affluence</i>	
Question A: Do you own a car?	yes/no

Guttman Scale

___	5. Abortion weakens the moral fiber of our society. (1.6)
___	6. We must allow abortions for certain situations. (8.7)
___	7. Abortion is never justified. (1.1)
___	8. It doesn't make any difference to me whether we allow abortions or not. (5.6)
___	9. Most people are too lax about allowing abortions. (3.2)
___	10. People who get abortions deserve our understanding, but not special treatment. (7.6)
___	11. Abortions should be easier to obtain than they are now. (9.1)
___	12. Abortion is justified only in instances of rape. (3.6)
___	13. We can't call ourselves civilized as long as we allow abortions. (2.5)
___	14. Abortion may be unacceptable, but people deserve our sympathy. (4.6)

Thurstone Scale

___	5. Compared to our competitors, is our product quality better, worse, or about the same?
___	6. Our product is more reliable than our competitors'.
___	7. Our product is more durable than our competitors'.
___	8. Our product is more stylish than our competitors'.
___	9. Our product is more comfortable than our competitors'.
___	10. Our product is more convenient than our competitors'.
___	11. Our product is more affordable than our competitors'.
___	12. Our product is more innovative than our competitors'.
___	13. Our product is more popular than our competitors'.
___	14. Our product is more trustworthy than our competitors'.

Category Scale

Types of Scale



Evaluate our brand in terms of the following statements:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I would recommend this brand to others.	☐	☐	☐	☐	☐
I have had a positive experience with this brand's customer service.	☐	☐	☐	☐	☐
I would purchase from this brand again.	☐	☐	☐	☐	☐

SUBMIT

Likert Scale

able

2 3 4 5 6

Acc

Co

Unsy

2 3 4 5 6

Semantic differential

2. Constant sum scale - Example

Que.: Below are five attributes that you might consider while buying a 4-wheeler. Please allocate a total of 100 points to them. More the points, more important is the attribute to you. You can allocate 0 point if the attribute is not at all important to you.

Attribute	Points
Mileage	
Seating capacity	
Luggage space	
Proximity of dealer	
Servicing network	
TOTAL	100

Constant sum scale

Types of Scale



+5	+5
+4 X	+4
+3	+3 X
+2	+2
+1	+1
gh Food Quality	Poor Crew Servi
-1	-1
-2	-2
-3	-3
-4	-4

Stapel Scale

Dull Ache	Slight Pain	More Sight Pain	Pain- ful	Very Painful	Unb Pain
hic Rating Scale					
Unb Pain					

Graphic Scale

Graphic rating scales

EXHIBIT 8.8 Graphic rating scale with picture response categories that stress visual communication

Happy Face Scale

3
Very Good

2

1
Very Poor

Graphic scale-
picture response



Summary of Scales

Rating Scale	Subject Must	Advantages	Disadvantages
Guttman Scale	Respond to a series of statements about a topic (Cumulative scoring)	Implemented to gain insights for multiple queries Produces data in a ranked manner:	Guttman scale is that respondents may feel overly committed to questions.
Thurstone Scale	Respond (agree / disagree) to a number of statements on a topic. Score is averaged out.	Equal appearing intervals.	Complex scoring and development.
Category scale	Indicate a response category	Flexible, easy to respond	Ambiguous items, few categories, only gross distinction.
Likert scale	Evaluate statements on a 5- or 7-point scale	Easiest scale to construct	Hard to judge what a single score means
Semantic differential and numerical scales	Choose points between bipolar adjectives on relative dimensions	Easy to construct, norms exist for comparison, e.g. profile analysis	Bipolar adjectives must be found, data may be ordinal, not interval
Constant sum scale	Divide a construct sum among response alternatives	Scale approximates an interval measure	Difficult for respondents with low education levels
Stapel scale	Choose point on scale with 1 centre adjective	Easier to construct than semantic differential	Endpoints are numerical, not verbal.
Graphic scale	Choose a point on a continuum	Visual impact, unlimited scale points	No standard answers
Graphic scale-picture response	Choose a visual picture	Visual impact, easy for poor readers	Hard to attach a verbal explanation to response



Assumptions of attitude scales

It is possible to express attitudes using verbal statements

A verbal statement will have the same meaning for all participants

Attitudes expressed as verbal statements can be measured or quantified

Likert Scale



This scale consists of a series of statements where the respondent provides answers in the form of agreement or disagreement.

The respondent selects a numerical score for each statement to indicate the degree of agreement or otherwise.

Each such score is finally added up to measure the respondents attitude.

Eg. of Likert Scale for evaluating the attitude of customers who have not used Vacuum cleaner, but who have aware of its existence:

	Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
The Product is costlier	1	2	3	4	5
I don't find time to use it	1	2	3	4	5
Advt. is not convincing	1	2	3	4	5
Never use a VC	1	2	3	4	5
I am satisfied with the present way of cleaning	1	2	3	4	5
Its use is cumbersome	1	2	3	4	5



Steps to creating a likert scale



1. Decide on the topic of your investigation

2. Collect a panel of volunteers

3. Ask them to generate statements about the topic

4. Collect all the statements and discard those which are repetitive or badly worded.

5. Put the statements into a list.

6. Ask a second panel to evaluate each statement as positive or negative.

7. Discard ambiguous statements, keeping only clear ones.

8. Select positive and negative statements and arrange them in random order.

9. Allocate a 5-point, 7-point or 9-point, scale to each statement.

10. Combine your statements into a full Likert-based questionnaire

Remember



Good questions **reduce error**

- By increasing the respondent's willingness to answer
- Increases reliability and validity.
 - Less error = better data

Reliability & Validity



► Reliability

“...the degree to which a test or measure produces the same scores when applied in the same circumstances...” (Nelson 1997)

Validity

“Degree to which a test or instrument measures what it purports to measure” (Thomas & Nelson 1996)

Objectivity

“...the degree to which different observers agree on measurements...” (Atkinson & Nevill 1998)

Reliability



- ▶ Reliability refers to the consistency of a test/survey/questionnaire/measure
- ▶ Is the survey measuring something consistently?
 - ▶ Typically measured using Cronbach's alpha or split-half reliability (estimate of the internal consistency associated with the scores that can be derived from a scale)

For example, if a person weighs themselves during the course of a day they would expect to see a similar reading. Scales which measured weight differently each time would be of little use.



A correlation coefficient can be used to assess the degree of reliability. If a test is reliable it should show a high positive correlation.

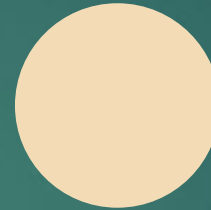
Reliability



There are two types of reliability – internal and external reliability.



Internal reliability assesses the consistency of results across items within a test.



External reliability refers to the extent to which a measure varies from one use to another.

Types of Reliability

INTERNAL

(extent to which a measure is consistent within itself.)

split-half method:

measures the extent to which all parts of the test contribute equally to what is being measured.

EXTERNAL

(the extent to which a measure varies from one use to another.)

test re-test: measures the stability of a test over time.

Inter-rater: to the degree to which different raters give consistent estimates of the same behavior

Split-half

TEST-RETEST INTER-RATER



Assesses the internal consistency of a test, such as psychometric tests and questionnaires.

It measures the extent to which all parts/items of the test contribute equally to what is being measured.

This is done by comparing the results of one half of a test with the results from the other half.

If the two halves of the test provide similar results this would suggest that the test has internal reliability.

It measures the stability of a test over time.

A typical assessment would involve giving participants the same test on two separate occasions.

If the same or similar results are obtained then external reliability is established.

The disadvantages of the test-retest method are that it takes a long time for results to be obtained.

Assesses the external consistency of a test...the degree to which different raters give consistent estimates of the same behaviour.

Note, it can also be called inter-observer reliability when referring to observational research.

Here researcher when observe the same behaviour independently (to avoided bias) and compare their data.

If the data is similar then it is reliable.

Internal Reliability



Cronbach Alpha

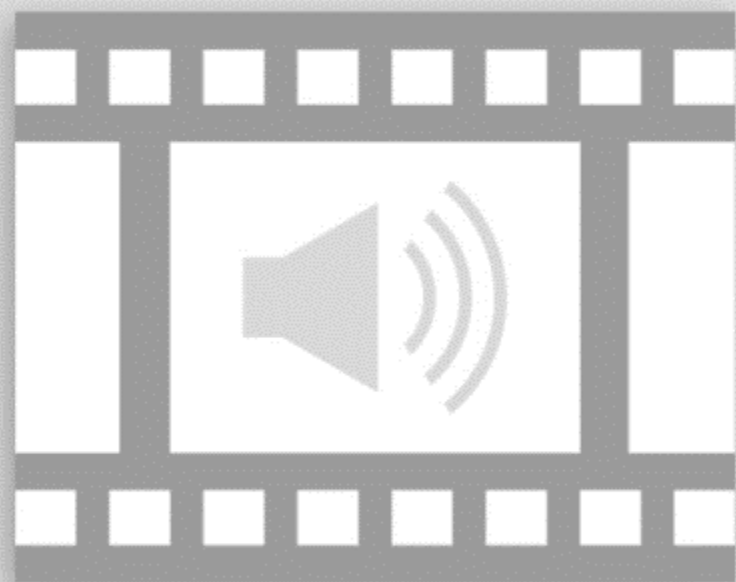
- ▶ Cronbach's alpha is a measure of internal consistency
- ▶ Measures all items – (or the items you wish to select) together.
- ▶ Typically used to see if all items of a likert scale are reliable.
- ▶ Careful if covering many sub areas (subscales) within a scale – should run separate Cronbach Alpha for these.
 - ▶ Emotion Scale with Subscales for Anger and Happiness – The Cronbach Alpha for the whole scale (all questions likely to be low).

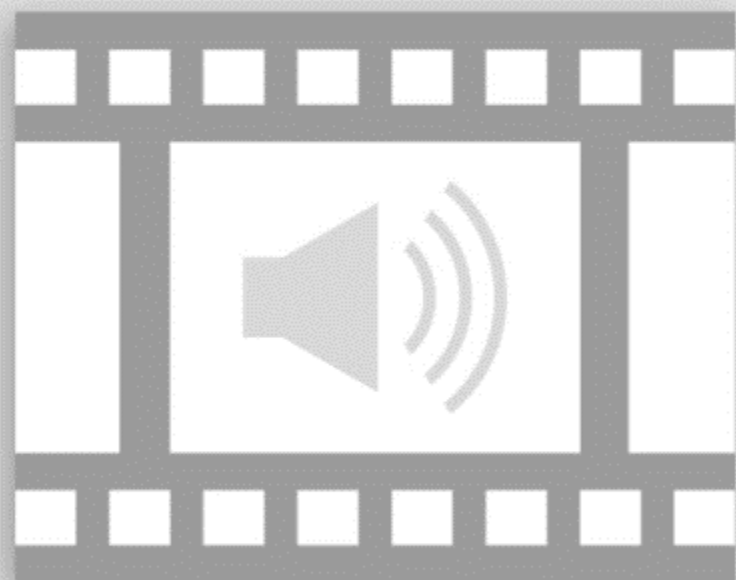
Split Half Reliability (Spearman Brown)

- ▶ Used in a variety of areas and in multiple trial scenarios.
- ▶ Splits the measures / items from the scale (or procedure) in half. Calculates reliability for each half separately.
- ▶ Then Spearman Brown Technique calculates a combined score.
- ▶ Not as frequently reported as Cronbach Alpha – but useful in multiple trial scenarios or two subscales of equal size.



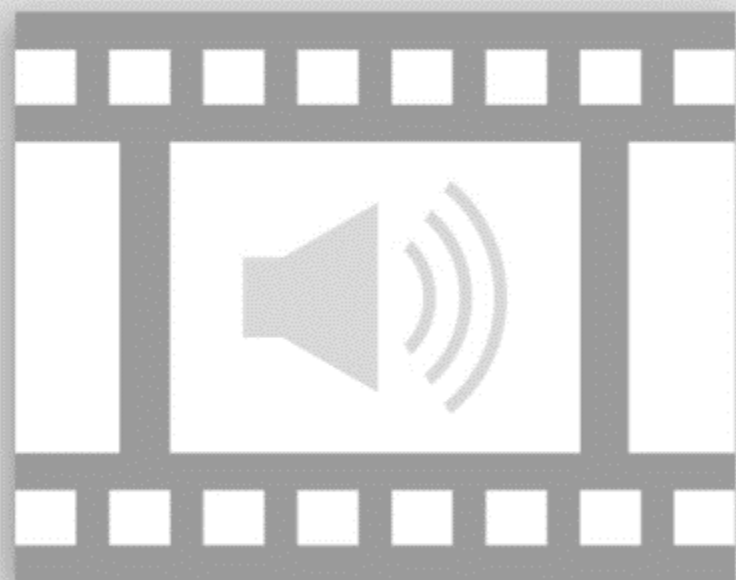
Running Cronbach Alpha on SPSS.







Running Split Half Reliability (Spearman Brown) on SPSS



Interpreting scale reliability



The higher the score, the more reliable the generated scale is

A score of .70 or greater is generally considered to be acceptable

lower thresholds are sometimes used in the literature.

.90 or > = high reliability

.80-.89 = good reliability

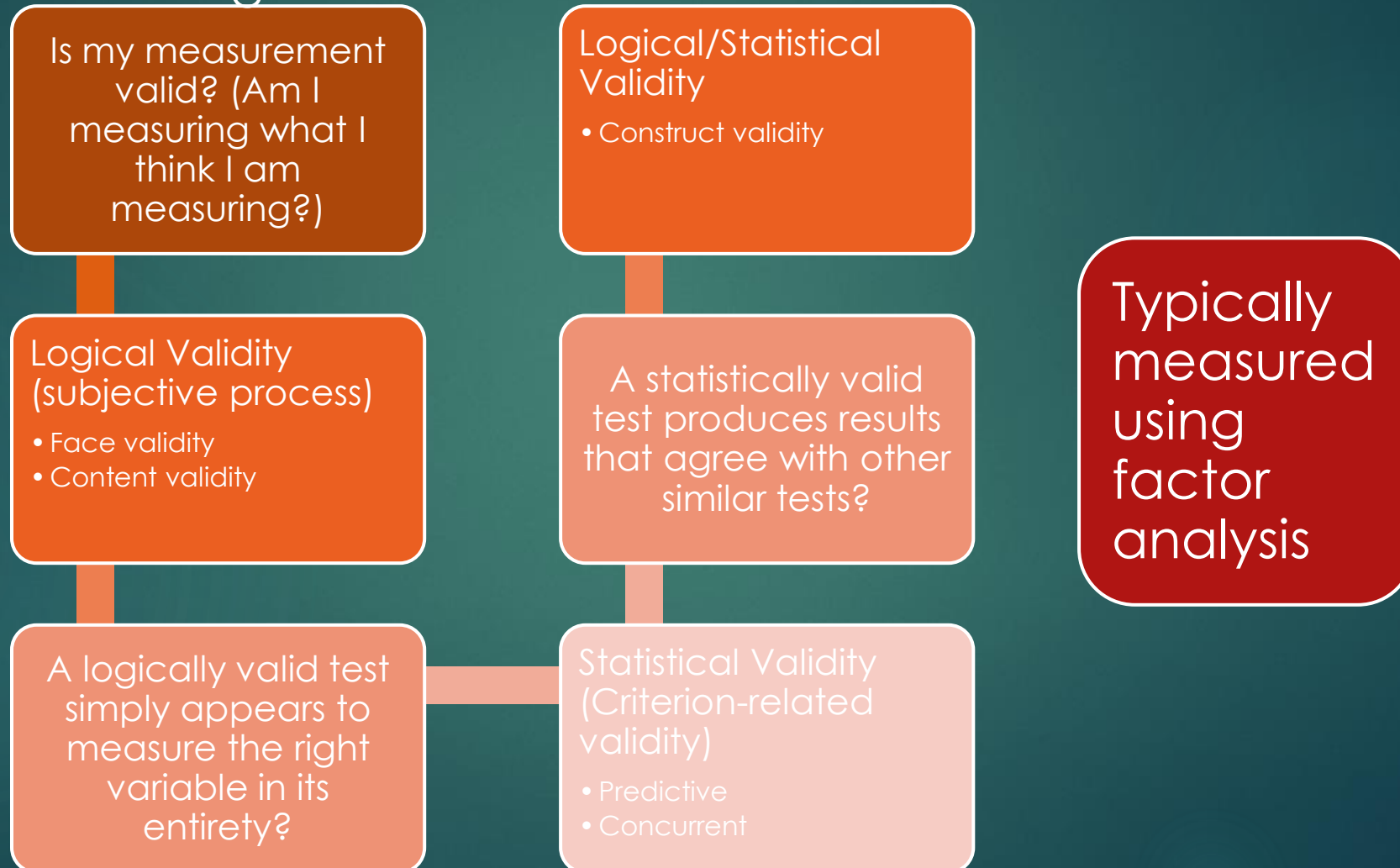
.70-.79 = acceptable reliability

.65-.69 = marginal reliability

Measuring validity



Is the survey measuring what it's supposed to be measuring?



Face Validity



Infers that a test is valid by definition

It is clear that the test measures what it is supposed to

e.g. if you want to assess reaction time, measuring how long it takes an individual to react to a given stimulus would have face validity



CONTENT VALIDITY



Infers that the test measures all aspects contributing to the variable of interest

e.g.

Who is the most physically fit?

- VO_2 max test?
- Wingate test (cycle)?
- 1 RM (repetition –weights)?

Concurrent Validity



Infers that the test produces similar results to a previously validated test



e.g.
 VO_2
max



Incremental Treadmill Protocol with
expired gas analysis

Multi-Stage Fitness (Beep) Test

PREDICTIVE VALIDITY

e.g. can performance during
test A be used to predict future
performance in test B?

Infers that the test
provides a valid
reflection of future
performance
using a similar test

A



B



Construct Validity



Infers not only that the test is measuring what it is supposed to, but also that it is capable of detecting what *should* exist, theoretically

Therefore relates to hypothetical or intangible constructs



e.g.
Team Rivalry



Sportsmanship.

This makes assessment difficult, i.e. if what *should* exist cannot be detected, this could mean:

a) Test Invalid? b) Theory Incorrect? c) Sensitivity/Specificity Issues?