Validity and Reliability Assessment

Harry Ganzeboom

RESMA Data Analysis & Report #2

February 5 2009

Measurement

- Measurement = expressing observations in numbers
- Our measures are always an imperfect reflection of the underlying phenomenon:
  - Incompleteness
  - Inaccurate
- How do we assess and improve quality of measurement?

Levels of measurement

- Levels
  - Nominal, categorical
  - Ordinal
  - Interval - metric
  - Ratio - metric
- To some extent all measurement is nominal; higher levels of measurement are only assumptions / hypotheses on patterns in data.
- Classical test theory assumes interval level at a minimum. Reliability is a complicated thing in ordinal and nominal data.

True score model

- Measure = true score + error
- x = X + e
- Error terms may be random:
  - Average to 0
  - Uncorrelated with X.
- Error terms may also be systematic
  - Do not average to 0 (bias)
  - Are correlated with X (heteroskedasticity)
  - Are correlated with other e or X variables in the system

Random error terms

- Random errors lead to unreliability: the observations are unstable in an unpredictable way.
- Unreliability will decrease with repeated measurement.
- Unreliability can be assessed with repeated measurement.
- Unreliability can be repaired.

Repeated measures

- Let x1 = X + e1 and x2 = X + e2.
- Then x1 en x2 will (only) be correlated due to the common (confounding!) influence of true score X.
- Size of correlation depends upon variance of e1 and e2 (error variances).
- Var(x) = var(T) +var (e)
- Reliability of x = var(T) / var(x) = R(x1,x2)

Reliability of composite measures

- Reliability can be defined at the level of individual indicators, but is more commonly a question about the average of multiple indicators.
- An assessment of reliability can be made using the (mean) correlation R among indicators.
- Cronbach's alpha: = N*R / (1+(N-1)*R)
  - N: Number of items
  - R: Average correlation among items

Lessons from Cronbach's alpha

- Reliability increases:
  - With higher R, i.e. stronger correlations among indicators
  - With larger N, i.e. using more indicators
- These two tendencies usually go in opposite directions: alpha can often be optimized.
- Alpha is an estimate of the correlation of a composite index with itself; its square root is the correlation between index and true score.

Other forms of reliability assessment

- Test-retest reliability: repeating the whole instrument after a little while
- Split-half reliability: correlating two random halves of one instrument; correct for test length.
- Both are special cases of internal consistency reliability – as measured by Cronbach's alpha. And have no particular advantage.

Validity

- Invalidity (bias) = lack of systematic error, i.e. we can predict the size of the error.
- Systematic under / overestimation.
- Correlated error: there is another variable that influences the error terms. E.g. ‘social desirability’.
- Systematic (correlated) bias can be estimated and repaired, but we need to measure the source of the bias.
- Multiple measurement can also be a big help to fight systematic bias. Bias in x1 can be repaired if we have an x2 that does contain another bias!

Types of validity

- Content validity: degree to which instruments cover aspects / facets / dimensions of underlying construct.
- Criterion validity:
  - Construct validity: degree to which instruments relates with other constructs.
  - Predictive validity: degree to which instrument predicts criterion.

Validity assessment

- The types of validity (there are more…) are not very helpful in assessing validity.
- Face validity: plausible arguments about whether instrument measures construct. Most obviously related to content validity.
- Content validity can be assessed using dimensional analysis (component analyses, factor analysis).
- Construct and predictive validity seem to overlap: they are on relationships with other variables.

Relationships between reliability and validity

- Reliability is a necessary condition for validity. An unreliable measure can never be valid, but an invalid measure can be reliably measured.
- Babbie stresses ***tension*** between validity and reliability: if you increase reliability, validity seems to decrease. I do not know whether this is true. It certainly does not imply that you can have validity without reliability – as Babbie implies.