Measurement models in longitudinal data analysis

MEASUREMENT INVARIANCE

Measurement invariance assumptions

- In all models we implicitly made an assumption that our tests measured the same construct(s) across the time points
  - We constrained the factor loadings, thresholds and reliabilities to stay the same
- Is this a fair assumption to make?
- Does this assumption hold?

Example: measuring self-esteem

- Measuring self-esteem longitudinally (from Horn, 1991)
- Suppose our measure is:
  - Do you feel you are as good looking as the average person?
  - Do you feel you are every bit as smart as the average person?
  - Do you feel you are liked by others as much as the average person is liked?
- Would the concept "self-esteem" have the same meaning (construct validity) for 20-year olds and for 60-year olds?

Example – continued

- Factor patterns might be like this

  X = .6*(looks) + .3*(smart) + .4*(likable)

  X = .0*(looks) + .8*(smart) + .4*(likable)
- Guess which one was found in a sample of 20-year olds?
- Qualitative difference in what is being measured
- Cannot use these items as valid indicators of self-esteem

Appropriate measures for each time point

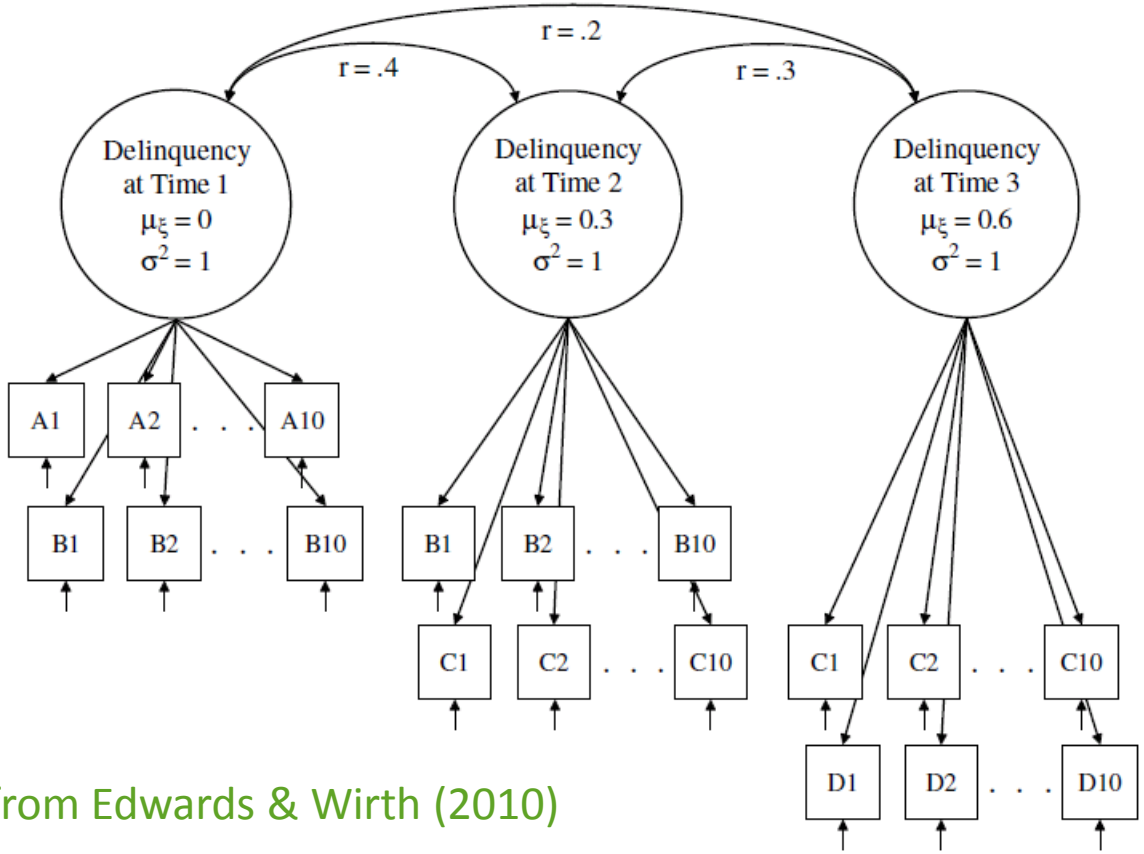


Illustration from Edwards & Wirth (2010)

Levels of equivalence

- Construct equivalence
  - The same psychological constructs are measured across time
- Measurement unit equivalence
  - The same measurement unit (individual differences found at time A can be compared with differences found at time B)
- Scalar / full score equivalence
  - The same measurement unit and the same origin (scores can be compared across time)

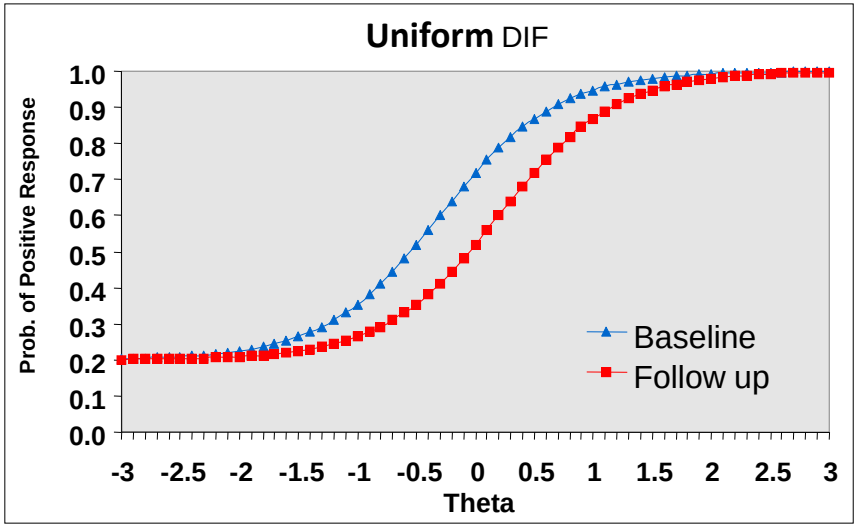
Same unit of measurement, same origin

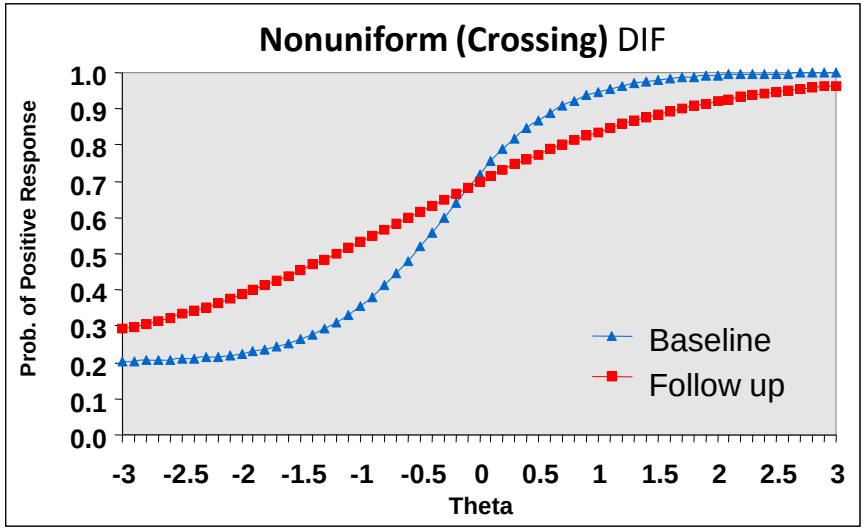
- Threatened by item bias
  - Some nuisance factor(s) that alter how the item response relates to the construct it measures at different times
  - The difference is not due to change in true score

- Differential Item Functioning (DIF)
  - Respondents show differing probabilities of endorsing the item from one time to another, *after matching on the level of construct* that the item is intended to measure

Types of DIF and levels of equivalence

- Uniform DIF
  - E.g. lower probability of endorsing the item at all trait levels
  - Affects origin of scale
- Non-uniform DIF
  - Higher probability of endorsing the item at low level of trait, but lower probability at high level (or vice versa)
  - Affects measurement unit and origin of scale





Why is DIF important?

- For example, latent difference score relies on measurement equivalence at the item level
  - Equal thresholds (no uniform DIF)
  - Equal loadings (no non-uniform DIF)
- If goes unnoticed, DIF distorts the model results
  - We can mistakenly take uniform DIF for real change in the construct level
  - Or non-uniform DIF for reduced stability of the construct

How to deal with DIF

- First any statistical findings must be interpreted by subject matter experts
- If confirmed as bias, it is advisable to adjust for this bias in the model
- For example, one can release equality constraints in M*plus*
  a) items without DIF have item parameters equal across time points (estimated at Time 1)
  b) items with DIF have parameters estimated separately at different time points