Validity Evidence via Explanatory Latent Trait Models

- Today's Topics:
  - Construct Validity
  - LLTM for Item Decomposition
  - Example of LLTM Approach: DriverScan
  - Items as Fixed vs. Random effects
  - Item Decomposition
  - Person Decomposition

2 Types of Construct Validity (Embretson, 1983)

- **"Nomothetic Span" = <u>external</u> evidence for validity**
  - What is usually targeted in validity studies
  - Individual differences in your test show expected relationships with other constructs (i.e., convergent and discriminant validity)
  - But what happens if expected relations are not found? Then what?
- **"Construct Representation" = <u>internal</u> evidence for validity**
  - If you understand your construct, you should know what processes, strategies, and knowledge are involved in item responding
  - Construct representation is operationalized by specifying item features as predictors/components of item difficulty
  - Essentially, you are predicting the ordering of items on the construct map as a function of their item stimulus characteristics (gettting difficulty right = validity)

Testing Construct Representation

- To understand the ability measured by a test is to understand which item features lead to differences in item difficulty
- One way to incorporate such hypotheses into an IRT model is via a **Linear Logistic Test Model** (LLTM; Fischer, 1973):

- **Rasch:** $p(\mathrm{y_{is}} = 1 |\, \theta_s) = \frac{\exp(\theta_s - \mathrm{b_i})}{1+\exp(\theta_s - \mathrm{b_i})}$
- **LLTM:** $p(\mathrm{y_{is}} = 1 |\, \theta_s) = \frac{\exp(\theta_s - [\mathrm{constant_i} + \sum(\tau_k \mathrm{q_{ik}})])}{1+\exp(\theta_s - [\mathrm{constant_i} + \sum(\tau_k \mathrm{q_{ik}})])}$

  - τ_k = weight of item feature k (same across items)
  - q_{ik} = value of item feature k (varies across items)
  - So each b_i is now created from a linear model of a constant (e.g., an intercept) + the weighted combination of item features

LLTM Approach

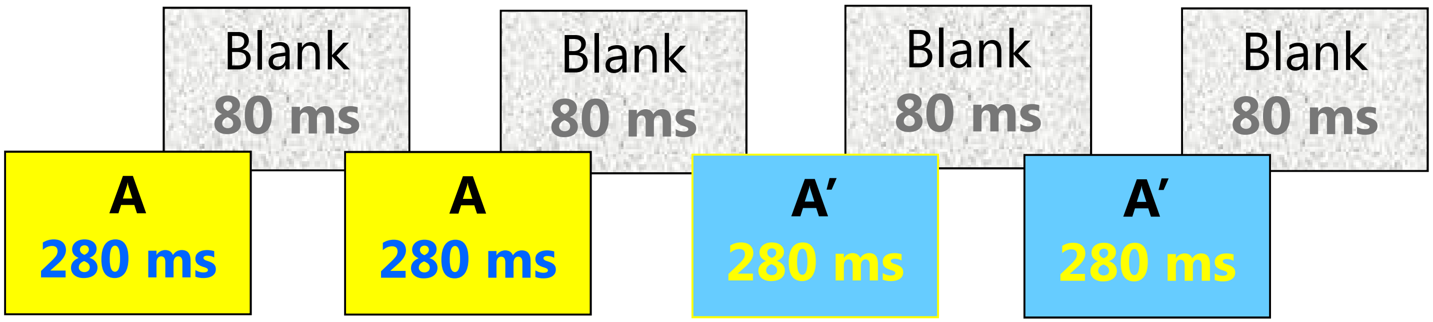
- **LLTM:** $p(\mathrm{y_{is}} = 1|\, \theta_s) = \dfrac{\exp(\theta_s - [\mathrm{constant_i} + \sum(\tau_k q_{ik})])}{1 + \exp(\theta_s - [\mathrm{constant_i} + \sum(\tau_k q_{ik})])}$
- Can also have polytomous versions (LPCM)
- Specify b_i as a **deterministic** function of item features
  - No residual term—that means b_i is a perfect function of $\tau_k q_{ik}$ (i.e., items are fixed effects, are interchangeable after controlling for item features)
  - Item feature weights (τ_k) can be tested for significance
  - Model fit is judged by correlation between b_i values from a Rasch model (i.e., a 'saturated difficulty' model) and calculated from the LLTM (or similarly via an item-level regression model predicting b_i terms)
- If you can reliably predict item difficulty from the features of the items, then such information has many advantages:
  - Create items of targeted difficulty levels where needed
  - Create items 'on the fly'

Example using LLTM for Construct Representation

- **DriverScan Instrument Design:**
- **Visual Clutter of Scene**
  - Greater amount and similarity of distractors hampers performance
- **Relevance of the Change to Driving**
  - Goal-directed orienting; effective compensatory strategy
- **Brightness of the Change**
  - Contrast sensitivity and retinal illumination declines
  - Attentional processing → quality of representation

Development of DriverScan: A Measure of Search Efficiency

<u>Change detection task via the "flicker paradigm"</u>



Presentation continues until 45 seconds or observer response.

<u>Pilot Study: Rated Item Design Features</u>

Visual Clutter of the Scene

Relevance of the Change to Driving

Brightness of the Change

Hoffman, Yang, Bovaird, & Embretson (2006)

Psychometric Evaluation of DriverScan via Item Response Theory

IRT: nonlinear latent trait measurement model that differentiates characteristics of both persons and items

Precision of Measurement:

1. Items cover the range of ability needing to be measured?
2. Reliability (information) across ability levels?

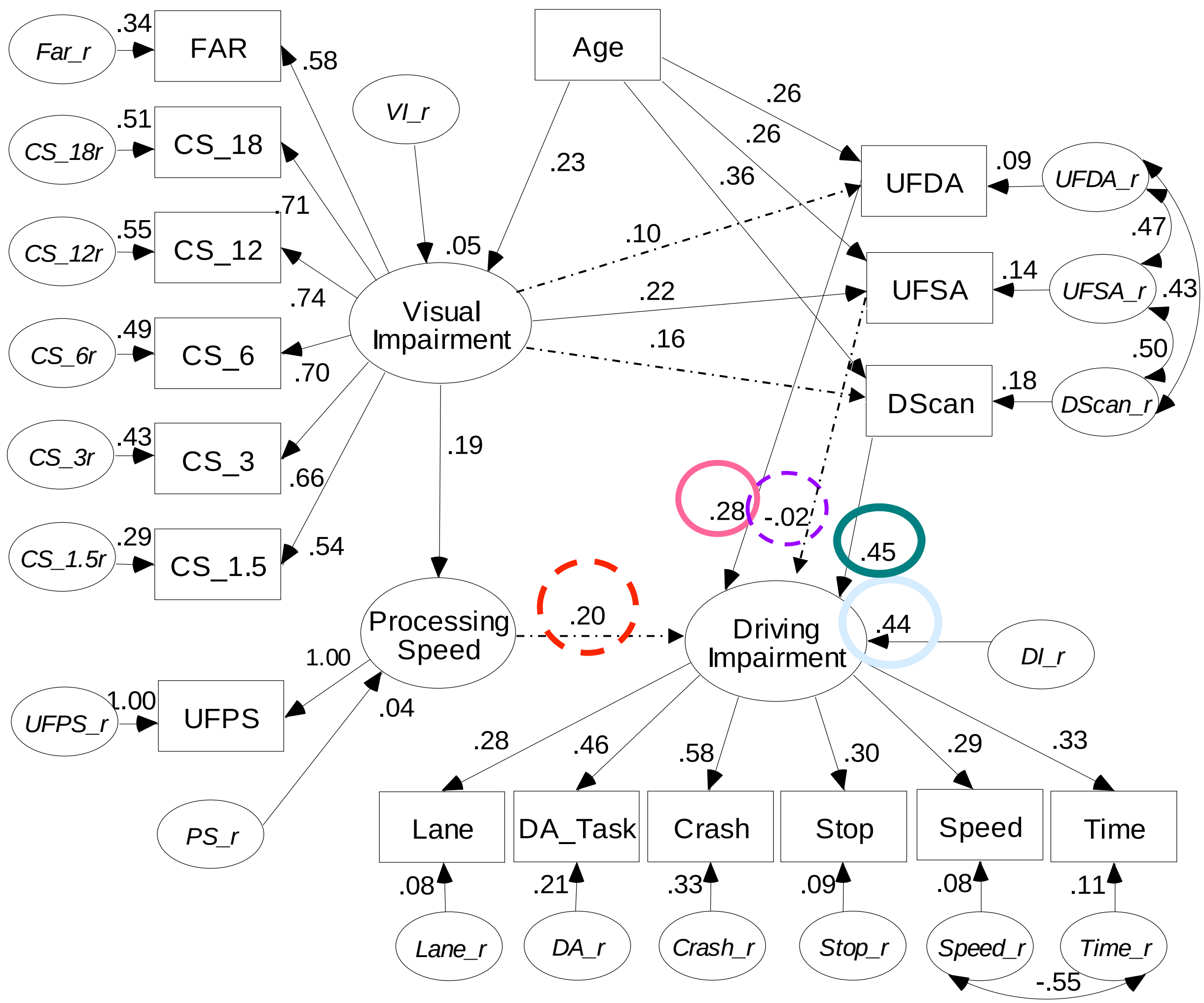
Construct Validity:

3. **Design features predict item difficulty?**
4. **Expected relationships with other constructs?**

Hoffman, Yang, Bovaird, & Embretson (2006)

4. Individual Differences: Nomothetic Span

Model Fit: $\chi^2(108) = 142$, CFI = .94, RMSEA = .05



Hoffman, McDowd, Atchley, & Dubinsky (2005)

Explanatory IRT Models

- Although LLTM is useful for testing hypotheses about construct representation, it has a few drawbacks:
  - Assumes perfect prediction of item difficulty (no residual term)
  - Model fit assessed via a two-stage procedure (usually suboptimal)

- More recently, **explanatory** IRT models have been developed within the estimation framework of "generalized linear mixed models" that can be used to assess both kinds of validity
  - "Generalized" → non-normal link functions (logit, probit, etc)
  - "Linear" → linear in the parameters (add weighted predictors)
  - "Mixed" → has both random and fixed effects
  - "Model" → prediction of data instead of description of data
  - De Boeck & Wilson (2004) show some of these via SAS NLMIXED, but in SAS 9.3 GLIMMIX can fit crossed random effects models as well

Measuring Ability: IRT Models

- **1PL model** predicts accuracy via fixed item effects and random person effects (i.e., n items are nested in persons)
- **1PL model**:
  - $\text{Probability}(y_{ip} = 1|\theta_p) = \frac{\exp(\theta_p - b_i)}{1+\exp(\theta_p - b_i)}$
  - $\text{Logit}(y_{ip} = 1|\theta_p) = \theta_p - b_i$

b_i is fixed effect of **difficulty** per item

θ_p is random person ability (variance τ_θ^2)

- **1PL can also be written as generalized multilevel model**:
  - $\text{Logit}(y_{ip} = 1|U_{0p}) = \gamma_{10}I_1 + \gamma_{20}I_2 + \cdots + \gamma_{n,0}I_n + U_{0p}$
  - Because item difficulty/easiness is perfectly predicted by the I indicator variables, items do not need a level-2 crossed random effect

γ_{i0} is fixed effect of **easiness** per item

U_{0p} is random person ability (variance τ_{0P}^2)

Measuring Ability: IRT Models

- 1PL can be extended to **predict item difficulty** via the LLTM
- **LLTM →** k item features predict $\mathrm{b_i}$, random persons (θ_p):
  - $\mathrm{Logit}(\mathrm{y_{ip}} = 1 | \theta_\mathrm{p}) = \theta_\mathrm{p} - \mathrm{b_i}$
  - $\mathrm{b_i} = \gamma_0 + \gamma_1 \mathrm{X_{1i}} + \gamma_2 \mathrm{X_{2i}} + \cdots + \gamma_k \mathrm{X_{ki}}$

Item difficulty is predicted via a linear model of X item features and γ fixed effects; **θ_p is random person ability (variance τ_θ^2)**

- **LLTM can also be written as generalized multilevel model**:
  - $\mathrm{Logit}(\mathrm{y_{ip}} = 1 | \mathrm{U_{0p}}) = \gamma_{00} + \gamma_{10} \mathrm{X_{1i}} + \gamma_{20} \mathrm{X_{2i}} + \cdots + \gamma_{k0} \mathrm{X_{ki}} + \mathrm{U_{0p}}$
  - Because there is no random item effect, the model says that items are still just nested within persons—that item difficulty or easiness is *perfectly* predicted by the X item features (no item differences remain)

Item easiness is predicted via a linear model of X item features and γ fixed effects
$\mathrm{U_{0p}}$ is random person ability (variance τ_{0P}^2)

Measuring Ability: IRT Models

- Experimental tasks can become psychometric instruments via **explanatory IRT (generalized multilevel) models** in which **items** and **persons** have crossed random effects at level 2

$$\mathrm{Logit}\left(y_{tip} = 1 \middle| U_{00p}, U_{0i0}\right) = \gamma_{00} + \gamma_{10}X_{1ip} + \gamma_{20}X_{2ip} + \cdots + \gamma_{k0}X_{kip} + U_{00p} + U_{0i0}$$

  - U_{0p} is person ability with variance of $\boldsymbol{\tau^2_{0P}}$
  - Item easiness is predicted via a linear model of X item features and γ fixed effects, with random (remaining) variance of $\boldsymbol{\tau^2_{0I}}$, so we can see directly how much item variance was predicted
  - Can also include person predictors to explain person random effects (so that the model can be explanatory on the person side as well)
  - Can examine random effects of X item features across persons (i.e., individual differences in effects of item features)

Explanatory IRT Model Extensions

- Testing for uniform DIF (group differences in difficulties)
  - Add group*item interaction terms for each item
  - Can test group*predictor DIF, too ("differential facet functioning")
- Many extensions are possible for polytomous data
  - Baseline category logit = nominal, adjacent category logit = partial credit, cumulative category logit = graded response
- Adding discrimination parameters is possible, but trickier:
  - 2PL: $\mathrm{Logit}(y_{is}) = a_i\,(\theta_s - b_i)$
  - This becomes: $\mathrm{Logit}(y_{is}) = a_i\theta_s - a_i b_i$
  - Because 2 parameters are multiplied together, this heads into truly "nonlinear" mixed models (nonlinear in the parameters)

Wrapping Up…

- Issues of construct validity primarily concern the question "How do I know I'm measuring what I think I am?"

- Two distinct ways of answering this:
  - **Construct representation** = internal evidence = able to predict differences across items in difficulty and/or discrimination
    - Test hypotheses about processes, strategies, and knowledge that are thought to contribute to the construct
  - **Nomothetic span** = external evidence = instrument's usefulness as a measure of individual differences
    - Test hypotheses about how other constructs should be related to it

- Both aspects of construct validity are important, and explanatory IRT models show promise as a means of assessing both within a single estimation framework