

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**” = external 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**” = internal 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

Testing Construct Representation

- To understand the ability being measured by an instrument, one should understand what item features lead to differences in item difficulty
- One way to incorporate such hypotheses into an IRT model is via a Linear Logistic Latent Trait Model (LLTM):
 - **Rasch**: $P(Y_{is}=1|\theta_s, b_i) = \frac{\exp(\theta_s - b_i)}{1 + \exp(\theta_s - b_i)}$
 - **LLTM**: $P(Y_{is}=1|\theta_s, \tau_k, q_{ik}) = \frac{\exp(\theta_s - [\text{constant} + \sum_k(\tau_k q_{ik})])}{1 + \exp(\theta_s - [\text{constant} + \sum_k(\tau_k q_{ik})])}$
 - 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 a linear function of a constant (e.g., an intercept) + the weighted combination of item features

LLTM Approach

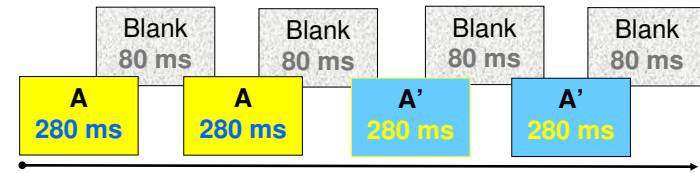
- LLTM:** $P(Y_{is}=1|\theta_s, \tau_k, q_{ik}) = \frac{\exp(\theta_s - [\text{constant} + \sum_k(\tau_k q_{ik})])}{1 + \exp(\theta_s - [\text{constant} + \sum_k(\tau_k q_{ik})])}$
- Older model, but can do polytomous versions (LPCM)
 - Specify b_i as a **deterministic** function of item features
 - Note no error term – that means b_i is a direct function of $\tau_k q_{ik}$
 - Item feature weights (τ_k) can be tested for significance
 - Model fit is judged by correlation between b_i 's from a Rasch model (i.e., a 'saturated model') and calculated from the LLTM (or similarly via an item-level regression model predicting b_i 's)
 - 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 & 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

Change detection task via the “flicker paradigm”



Presentation continues until 45 seconds or observer response.

Pilot Study: Rated Item Design Features

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: measurement model for persons and items

Precision of Measurement:

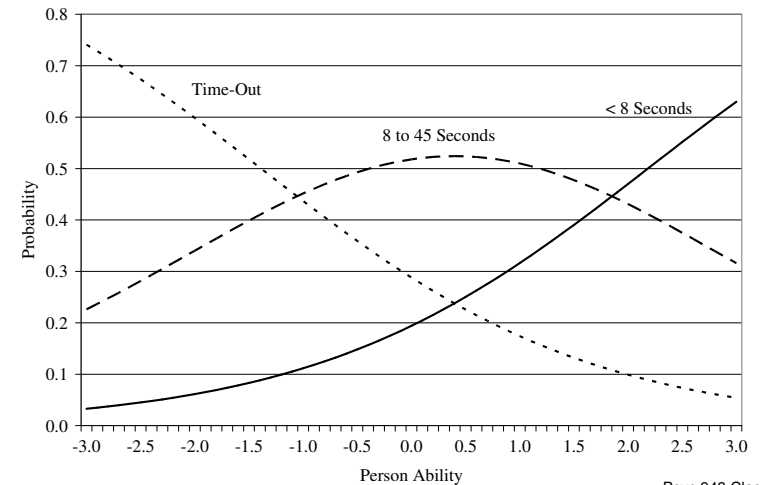
1. Items cover the range of ability
2. Reliability across ability levels

Construct Validity:

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

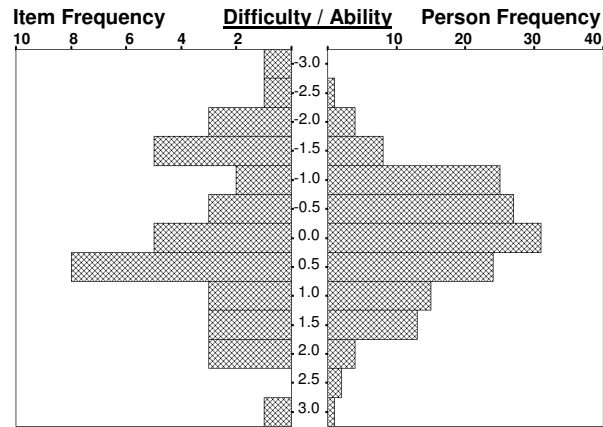
Hoffman, Yang, Bovaird, & Embretson (2006)

1. Example DriverScan Item Characteristic Curve ($\beta = .40$)

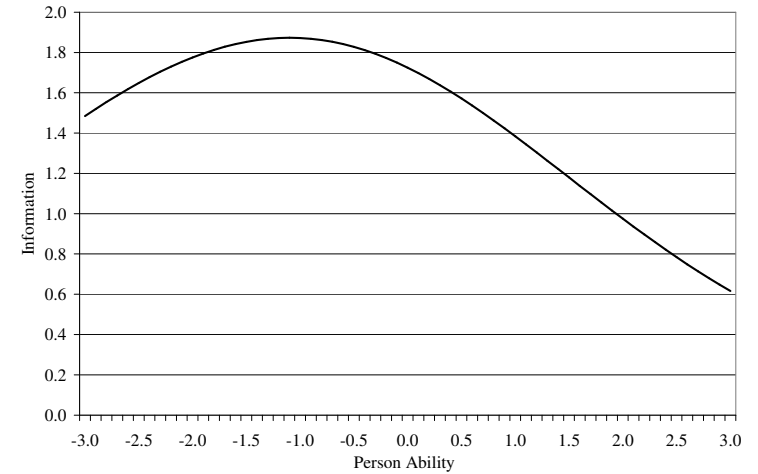


1. Distribution of Item Difficulty and Person Ability:

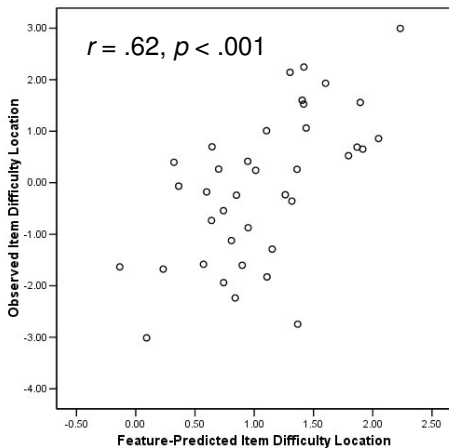
Constrained Graded Response Model (38 items, N = 155)



2. Distribution of Reliability: DriverScan Test Information Curve



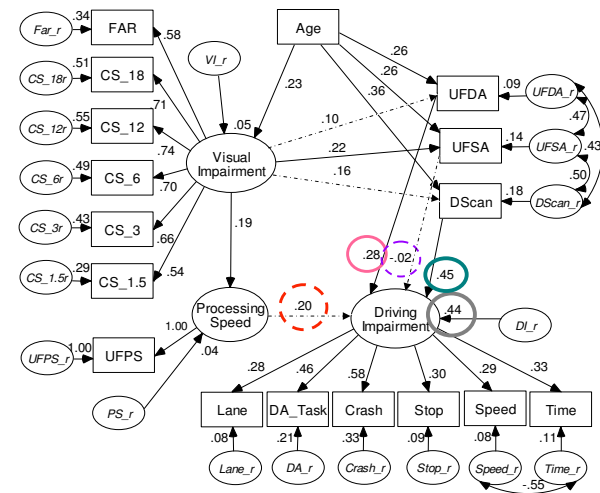
3. Construct Representation: Observed vs. Predicted Item Difficulties



- Faster change detection (less difficult items) WHEN:
 - Less clutter in the scene
 - More change relevance
 - More change brightness
 - Change to legible sign
- DriverScan abilities correlate with attention-related UFOV subtests ($r \approx .5$)

4. Individual Differences: Nomothetic Span

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



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 (suboptimal)
- More recently, new families of **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 NLMIXED

Fixed vs. Random Effects

- **Fixed effects:** Goal is to compare specific levels of an IV
 - e.g., interested in dose amount 1 vs. 10 vs. 100
 - No inference is intended beyond those specific factor values
 - In RM ANOVA, this is typically how IV effects are estimated
→ **fixed effects comparisons** among means for each IV level
- **Random effects:** Goal is to compare across many possible IV levels, not just those you happen to sample
 - e.g., interested in whether “**dosage matters**” (1 thru 100)
 - Intend inference to population of possible IV values
 - In RM ANOVA, we estimate a **random (subject) intercept variance**, not fixed effects (we don’t care how much higher subject 1 is than subject 2, just how much between-person intercept variance is taken out of residual within-person error variance)

Items as Fixed vs. Random Effects

- Prob: $P(Y_{is}=1|\theta_s, b_i) = \frac{\exp(\theta_s - b_i)}{1 + \exp(\theta_s - b_i)}$
- Logit(Y_{is}): $\text{LN}(p/1-p) = \theta_s - b_i$
- Another way of viewing Rasch models:
 - As a logistic mixed model where items are crossed with persons
 - Estimate **fixed effects** for each separate item (no overall fixed intercept) → these become item difficulties
 - Estimate a **random effect** as theta (i.e., an intercept U_{0s})
→ fix mean=0, variance=1 (or can be estimated in some models)
- $\text{Logit}(Y_{is}) = \theta_s - [b_1I_1 + b_2I_2 + b_3I_3 + \dots + b_nI_n]$
 - Where the I’s are dummy codes that identify which item, and b’s are item difficulties for each item

Items as Fixed vs. Random Effects

- The traditional Rasch model treats items as fixed effects
 - Useful to get the difficulty (intercept) for each item specifically
 - Not useful if we want to **predict difficulty**, because having the item dummy codes saturates the item part of the model
 - It would be like putting in N-1 dummy codes to ‘control’ for subject... effective, but not parsimonious and not informative
- We can estimate items as a random effect instead
 - We estimate a fixed intercept (grand mean response), a random effect for theta (differences in ability across subjects), and a random effect for item (differences in difficulty over items)
 - $\text{Logit}(Y_{is}) = \gamma_0 + \theta_{0s} - b_i$

Items as Random Effects

- We can estimate **items as a random effect** instead
 - $\text{Logit}(Y_{is}) = \gamma_{00} + \theta_{0s} - b_i$
- This allows us to put in item features (X 's) as predictors of variance in difficulty across items to assess **construct representation**:
 - $\text{Logit}(Y_{is}) = \gamma_{00} + \gamma_{10}(X_{1i}) + \gamma_{20}(X_{2i}) + \theta_{0s} - b_i$
 - Predicted response = intercept + effect of item feature 1 + effect of item feature 2 + which person – which item
 - The extent to which variance across items is reduced as a function of the predictors can be directly evaluated in the model
 - Thus, this is now an 'explanatory' model with respect to the items, because differences between items in difficulty is due to theoretical factors, not just which item it is (which would be a descriptive model)

Persons as Random Effects

- Similarly, we can put in predictors of ability to see to what extent variance in theta can be reduced by person characteristics
 - This is useful for evaluating **nomothetic span** types of questions
- This model is now '**explanatory**' on both the item side (X item features predict item difficulty variance) and on the person side (Z person features predict person theta variance):
 - $\text{Logit}(Y_{is}) = \gamma_{00} + \gamma_{10}(X_{1i}) + \gamma_{20}(X_{2i}) + \gamma_{01}(Z_{1s}) + \gamma_{02}(Z_{2s}) + \theta_{0s} - b_i$
 - Note that this assumes measurement invariance over the Z 's
 - Can put in $X*Z$ interactions to test for "facet invariance" instead
- We could also add additional random effects over persons for differential effects of the item features → **multidimensional** thetas:
 - $\text{Logit}(Y_{is}) = \gamma_{00} + \gamma_{10}(X_{1i}) + \gamma_{20}(X_{2i}) + \gamma_{01}(Z_{1s}) + \gamma_{02}(Z_{2s}) + \theta_{1s}(X_{1i}) + \theta_{2s}(X_{2i}) + \theta_{0s} - b_i$

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 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: $\text{Logit}(Y_{is}) = 1.7a_i(\theta_s - b_i)$
 - This becomes: $\text{Logit}(Y_{is}) = 1.7a_i\theta_s - 1.7a_ib_i$
 - 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