Explanatory Latent Trait Models: A Tale of Two Studies

Lesa Hoffman

Professor, Educational Measurement and Statistics Psychological and Quantitative Foundations University of Iowa College of Education

Presented to CASMA Colloquium Series, 5/6/2022

Prelude: The Hofflin Lego-Based View of Quantitative Methods



Big Picture Idea:

If you understand the elemental building blocks of statistical models, then you can build **anything**! Today I want to illustrate how thinking this way has shaped my research for the better.



The 4 Lego Building Blocks

- 1. Linear models (for answering questions of prediction)
- 2. Estimation (for iterative ways of finding the answers)
- 3. Link functions (for predicting any type of outcome)
- 4. (a) Random effects /
 (b) Latent traits / factors / variables

(a) for modeling multivariate "correlation/dependency"(b) for modeling relations of "unobserved constructs"

How the Blocks Fit Together

- 1. Linear models answer research questions, and are the first building block of every more complex analysis
 - Is there an effect? Is this effect the same for everyone?
 Is the effect still there after considering something else?

To add more blocks, you need iterative **estimation** 2. Maximum likelihood or Bayesian (e.g., MCMC)

What other blocks you will need is determined by:

3. How your outcome is measured \rightarrow link functions

4. Your dimensions of sampling \rightarrow random/latent effects

From One to Many Outcomes...

- Most designs have more than one outcome per person...
 - > e.g., multiple outcomes, occasions, items, trials ... per person
 - > Multiple dimensions of **sampling** \rightarrow multiple kinds of **variability**



4. Random Effects / Latent Variables

- Random effects are for "handling dependency" that arises because multiple dimensions of sampling → multiple variances
 - Occasions within children (need 1+ random effect)
 - > Children within classrooms within schools (need 2+ random effects)
 - > aka, multilevel, mixed-effects, or hierarchical linear models
- Latent <traits/factors/variables> are for representing "error-free true construct variance" within observed variables
 - Normal outcomes + latent variables = factor analysis (CFA; SEM)
 - Categorical outcomes + latent variables = item response theory (IRT)
- Random effects / latent variables are **mechanisms** by which:
 - > Make best use of all the data; avoid list-wise deletion of incomplete data
 - > Quantify and predict distinct sources of variation... *cue story-time*...

The Curse of Non-Exchangeable Items

Jim Bovaird, University of Nebraska-Lincoln



Larry Locker, Georgia Southern University





- Psycholinguistic research (items are words and non-words)
 - Common persons, common items designs
 - Contentious fights with reviewers about adequacy of experimental control when using real words as stimuli
 - Long history of debate as to how data should be analyzed:
 F1 ANOVA, F2 ANOVA, or both?

Larry's Kinds of ANOVAs

Original Data per Person

| | B1 | B2 | | |
|----|--------------------------------------|--------------------------------------|--|--|
| A1 | Item 001 Item 002 Item 100 | Item 101 Item 102 Item 200 | | |
| A2 | Item 201 Item 202 Item 300 | Item 301 Item 302 Item 400 | | |

Person Summary Data

| | B1 | B2 |
|----|------------------|------------------|
| A1 | Mean (A1, B1) | Mean (A1, B2) |
| A2 | Mean (A2, B1) | Mean (A2, B2) |

"F1" Within-Persons ANOVA on N persons: $RT_{cp} = \gamma_0 + \gamma_1 A_c + \gamma_2 B_c + \gamma_3 A_c B_c + U_{0p} + e_{cp}$

"F2" Between-Items ANOVA on / items: $RT_i = \gamma_0 + \gamma_1 A_i + \gamma_2 B_i + \gamma_3 A_i B_i + e_i$

Item Summary Data

| | B1 |
|--------|--|
| A1, B1 | Item 001 = Mean(Person 1, Person 2, Person N) Item 002 = Mean(Person 1, Person 2, Person N) Item 100 |
| A1, B2 | Item 101 = Mean(Person 1, Person 2, Person N) Item 102 = Mean(Person 1, Person 2, Person N) Item 200 |
| A2, B1 | Item 201 = Mean(Person 1, Person 2, Person N) Item 202 = Mean(Person 1, Person 2, Person N) Item 300 |
| A2, B2 | Item 301 = Mean(Person 1, Person 2, Person N) Item 302 = Mean(Person 1, Person 2, Person N) Item 400 |

Choosing Amongst ANOVA Models

- F1 Within-Persons ANOVA on person summary data:
 - > Within-condition *item* variability is gone, so items assumed fixed
- F2 Between-Items ANOVA on item summary data:
 - > Within-item *person* variability is gone, so persons assumed fixed
- Historical proposed ANOVA-based resolutions:
 - F' → quasi-F test with random effects for both persons and items (Clark, 1973), but requires complete data (uses least squares)
 - > Min F' → lower-bound of F' derived from F1 and F2 results, which does not require complete data, but is too conservative
 - F1 x F2 criterion → effects are only "real" if they are significant in both F1 and F2 models (*aka*, death knell for psycholinguists)
 - > But neither model is complete (two wrongs don't make a right)...

Multilevel Models: A New Way of Life?

Original Data per Person

| | B1 | B2 |
|----|--------------------------------------|--------------------------------------|
| A1 | Item 001 Item 002 | Item 101 Item 102 |
| A2 | Item 201 Item 202 Item 300 | Item 301 Item 302 Item 400 |

Pros:

- Use all original data, not summaries
- Responses can be missing at random
- Can include continuous predictors

Cons:

• Is still wrong (is ~F1 ANOVA)

Level 1:
$$y_{ip} = \beta_{0p} + \beta_{1p}A_{ip} + \beta_{2p}B_{ip} + \beta_{3p}A_{ip}B_{ip} + e_{ip}$$

Level 2: $\beta_{0p} = \gamma_{00} + U_{0p}$ Level 2: $\beta_{1p} = \gamma_{10}$ $\beta_{2p} = \gamma_{20}$ Level $\beta_{3p} = \gamma_{30}$

Level 1 = Within-Person Variation (Across Items)

Level 2 = Between-Person Variation

Multilevel Models: A New Way of Life?



A Better Way of (Multilevel) Life



• Multilevel Model with Crossed Random Effects: $RT_{tpi} = \gamma_{000} + \gamma_{001}A_i + \gamma_{002}B_i + \gamma_{003}A_iB_i$ $+ U_{0p0} + U_{00i} + e_{tpi}$ t trial p person i item

- Explicitly test **persons** and **items** as random effects:
 - > Person predictors capture between-person mean variation: τ_{0P0}^2
 - > Item predictors capture between-item mean variation: τ_{00I}^2
 - > Trial predictors capture trial-specific residual variation: σ_e^2

Larry's Story: Example Data

- Crossed design: 38 persons by 39 items (words or nonwords)
- Lexical decision task: Response Time to decide if word or nonword
- 2 word-specific predictors of interest:
 - > A: Low/High Phonological Neighborhood Frequency
 - » B: Small/Large Semantic Neighborhood Size





Locker, Hoffman, & Bovaird (2007, Behavior Research Methods)

Not Just in Larry's Example Data...

 Generality of results examined via simulation study of Type I error rates for person or item predictor effects

Testing person effects in common persons design?

- > Need to include **person** variance in model (so not **F2 ANOVA**)
- Need random effect for **persons** (in **MLM** or in **F1 ANOVA**), so that **person** predictors can explain that **person** variance

• Testing item effects in common items design?

- Need to include item variance in model (so not F1 ANOVA)
- Need random effect for items (in MLM or in F2 ANOVA), so that item predictors can explain that item variance

Nested vs. Crossed Multilevel Designs

- When should items be a separate level-2 random effect?
 - Items are clearly nested within persons if the model fixed effects
 explain all item differences (so no item variation remains)
 - e.g., via item-specific indicators (CFA, IRT; stay tuned)
 - e.g., by item design features given only one item per condition
 - > Items are clearly nested within persons if they are **endogenous**
 - e.g., autobiographical memories, eye movements, speech utterances
 - > More ambiguous if items are **randomly generated** per person
 - If items are truly unique per person, then there are no common items... but items are usually constructed systematically
 - Modeling items as **nested (no variance) assumes exchangeability**
- When does this matter?
 When turning experiments into instruments...

Paradigms in Studying Cognition

Experimental Designs

- Goal is inference about processes or architecture of cognitive ability
- Create meaningfully different items through **specific** manipulations
- Many items given to **few** people
- Multiple aspects of construct represented within a single task
- ANOVA \rightarrow Ability represented by:
 - Mean performance (e.g., RT, # correct)
 - Mean differences between conditions
- MLM \rightarrow Ability represented by:
 - Random intercept
 - Random slopes for item effects

Psychometric Measures

- Goal is to measure individual differences in cognitive ability
- Create equivalent items to reflect general ability being measured
- Fewer items given to **more** people
- **Multiple measures** given to better represent the ability construct
- CTT \rightarrow Ability represented by:
 - Mean performance (e.g., # correct)
 - Mean/component of multiple measures
 - CFA/IRT \rightarrow Ability represented by:
 - > Random intercept (≈ factor, theta)
 - Multidimensional ability model

Combining Paradigms

- The fine-grained task decomposition found in experimental designs can be combined with latent trait models to more rigorously quantify and predict individual differences
 - > **Synergy** of experimental and individual differences research
 - Theoretical models of cognitive processes inform test construction; research using these instruments then informs theoretical models
- Long-term goal: construct measures of cognition that are theoretically meaningful and psychometrically viable
 - > Short-term goal: build instruments to individual **visual attention**
 - > Individual differences matter in aging and in real-world situations
 - > Lack of *psychometric* **instruments** to measure attention
 - > **Visual search** tasks are well-understood \rightarrow recipe for item creation

Why Measure Selective Attention?

- Attention is...
 - > "A system for routing information and for control of priorities" (Posner, 1980)
 - "The capacity or energy to support cognitive processing" (Plude & Hoyer, 1985)
- Lifespan changes in attentional abilities matter:
 - Significant real-world consequences of attentional deficits with age (that can't be fixed by glasses or heading aids)
 - Difficulty with specific aspects of modulating attention is a marker of some non-normative aging processes
- Measuring visual search in particular:
 - > Task difficulty is well-understood \rightarrow recipe for item creation
 - > Current lack of *psychometric* instruments to measure attention
 - Attention is rarely measured in individual differences studies, so little is known about how it relates to other abilities (nomothetic span)

Measuring Visual Search Ability: Take 1



Change detection task using the "flicker paradigm"

cycle continues until response for max of 45 sec

Rated Item Design Features:

- Visual clutter of the scene
- Relevance of the change to driving
- Brightness of the change
- Change made to legible sign
- 155 persons, 46 items retained,
 DV = response time (if < 45 sec)



Measuring Visual Search Ability: Take 1

- How to fit a censored response time into an "IRT" model?
 - > Cut up RT, fit tau-equivalent graded response model (GRM)
 - "1. immediate" = RT < 8 sec, "2. delayed" = 8-45 sec, "3. time out"</p>
- LLTM version of GRM to examine predictors of item difficulty $prob(y_{pi} > c | \theta_p) = \frac{exp(\theta_p - \beta_{ic})}{1 + exp(\theta_p - \beta_{ic})} \qquad \begin{bmatrix} i = \text{item} \\ c = \text{category (threshold)} \end{bmatrix}$
- Where each item threshold is:

 $\begin{aligned} \beta_{ic} &= \gamma_{c0} + \gamma_1 Clutter_i + \gamma_2 Relevance_i + \gamma_3 Brightness_i \\ &+ \gamma_4 Sign_i \end{aligned} \ (difference of category intercepts modeled directly) \end{aligned}$

p = person

• r = .62 of model-predicted and observed item difficulty

Predicting Driving Impairment*

- 155 current drivers age 63-87; 56% women
- Predictors:
 - Vision (distance acuity, contrast sensitivity)
 - > Visual Attention (Useful Field of View subtests, DriverScan)
- Driving Simulator Task Outcome:
 - Easy curves, divided attention, passing, stoplights, obeying speed limits, weaving, narrow radius turns, overtaking vehicles
 - Nothing predicted self-reported and state-recorded accidents
- * Like a good neighbor, State Farm was there (2002 Dissertation Grant)



Hoffman, McDowd, Atchley, & Dubinsky (2005, Psychology and Aging)

Measuring Visual Search Ability Take 1: Lessons Learned

Response time is problematic as an outcome

- > Speed is contaminated with decision threshold
- > Physical limitations may prevent older adults from responding quickly
- Continuous distribution, but it's almost always very skewed
- > Limited utility in real-world assessment

Change detection task format is less than ideal

- > Other-rated item features don't generalize to new items
- > No basis for extrapolation for to create new items
- Fixed test items can't be used to measure change
- > What if search ability measured was specific to driving scenes?
- Time for Take 2 → use accuracy and standard search tasks
 → use legit explanatory IRT models

Measuring Visual Search Ability: Take 2 Project Goals:

Determine the dimensionality of visual attention ability within and between methods of assessment and its relationships with other constructs (~nomothetic span)

 Identify the factors that predict task difficulty commonly across both context-free, simple visual search tasks and context-specific, applied visual search tasks measuring selective visual attention (~construct representation)

Abilities Measured (by # tasks):

primary memory (3), working memory (3), comparison speed (3), visual search (4)





Context-Free Basic Search Tasks



Context-Specific Applied Search Tasks

Web Page Search: find link to either "Medical Center" or "Grayhawk Lab"



Grocery Shelf Search: find either can of corn or can of carrots



Measuring Visual Search Ability



and presentation time were counterbalanced.

Predictors of Accuracy:

- Item presentation time (short, medium, long)
- Target location in 6x6 grid (inner, middle, outer)
- # distractors (5 levels)
- % distractors similar to target (~20, 40, 60, 80, 100)
- Log of trial order

Sample:

- 329 adults (OA: age 62-88)
- 102 college students (YA)
- Shared medium time, and
 YA→ short, OA → long

Lego #3: Predicting Accuracy Instead of RT



Latent Variable Models of Ability

- **1PL model** predicts accuracy via fixed item effects and random person effects (i.e., *n* items are nested in persons)
- 1PL model:

> Probability
$$(y_{pi} = 1 | \boldsymbol{\theta}_{p}) = \frac{exp(\boldsymbol{\theta}_{p} - \boldsymbol{b}_{i})}{1 + exp(\boldsymbol{\theta}_{p} - \boldsymbol{b}_{i})}$$

>
$$Logit(y_{pi} = 1 | \boldsymbol{\theta}_p) = \boldsymbol{\theta}_p - \boldsymbol{b}_i$$

$$b_i$$
 is fixed effect of
difficulty per item
 θ_p is random person
ability (variance τ_p^2)

1PL is also a general<u>ized</u> multilevel model:

- > $Logit(y_{pi} = 1 | U_{p0}) = \gamma_{01}I_1 + \gamma_{02}I_2 + \dots + \gamma_{0n}I_n + U_{p0}$
- Because item difficulty/easiness is perfectly predicted by the *I* indicator variables, here items do not need a level-2 crossed random effect

 γ_{0i} is fixed effect of <u>easiness</u> per item

 U_{p0} is random person ability (variance τ_{P0}^2)

Adding Lego #1: Linear Models

- 1PL can be extended to **predict item difficulty** via the LLTM
- **LLTM** $\rightarrow k$ item features predict b_i ; random persons (θ_p):
 - > $Logit(y_{pi} = 1 | \boldsymbol{\theta}_{p}) = \boldsymbol{\theta}_{p} \boldsymbol{b}_{i}$
 - $\succ b_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \dots + \gamma_k X_{ki}$

Item difficulty = linear model of k item features (of $X^*\gamma$ fixed effects); θ_p is random person ability (variance τ_{θ}^2)

- LLTM written as a generalized multilevel model:
 - > $Logit(y_{pi} = 1 | U_{p0}) = \gamma_{00} + \gamma_{01}X_{1i} + \gamma_{02}X_{2i} + \dots + \gamma_{0k}X_{ki} + U_{p0}$
 - > 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 \rightarrow **no item differences remain**

Item easiness = a linear model of k item features (of $X^*\gamma$ fixed effects); U_{0p} is random person ability (variance τ_{P0}^2)

Proof of Concept: Random Items Matters

Item re-analysis predicting accuracy in dissertation data using SAS PROC GLIMMIX (Laplace estimation)

| Effect | Items Treated as Fixed | | Items Treated as Random | | | |
|--------------|------------------------|-------|-------------------------|-------|---------|------------|
| | Est | SE | p < | Est | SE | p < |
| Intercept | 0.862 | 0.153 | .0001 | 1.31 | 1 0.635 | .0474 |
| Clutter | -0.268 | 0.055 | .0001 | -0.32 | 0.242 | .1809 |
| Relevance | 0.220 | 0.099 | .0266 | 0.03 | 0.426 | .9305 |
| Brightness | 0.474 | 0.113 | .0001 | 0.79 | 0.499 | .1136 |
| Legible Sign | 0.662 | 0.082 | .0001 | 0.73 | 0.337 | .0283 |
| | | | | | | |

Putting It All Together...

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

 $Logit(y_{tpi} = 1) = \gamma_{000} + \gamma_{001}X_{1i} + \gamma_{002}X_{2i} + \dots + \boldsymbol{U_{0p0}} + \boldsymbol{U_{00i}}$

- > U_{0p0} is person ability with random (unpredicted) variance of τ^2_{0P0}
- > U_{00i} is item easiness is predicted from a linear model of the X item features, with random (leftover) variance of τ_{00I}^2
- > Can add person predictors to explain τ_{0P0}^2
- Can examine random effects across persons of X item features (i.e., differential susceptibility to item manipulations)
- So how did we do? Let's examine **reliability for U_{0p0}** ...

Reliability of Individual Differences



From model controlling for level-1 presentation time only: $Logit(y_{tpi} = 1) = \gamma_{000} + \gamma_{100}Time_{tpi} + U_{0p0} + U_{00i}$

Improving Efficiency (Reducing Boredom)

- Can we give fewer items but still retain measurement precision?
 - > ANOVA/CTT: Ability is mean RT or # correct? Then no.
 - > MLM/IRT: Ability is estimated along with item properties? Then yes!
- Adaptive search tasks in 5 easy steps:
 - 1. **Decompose** item difficulty into effects of known features
 - 2. **Create** new, structurally equivalent items on the fly
 - 3. Estimate person ability between each item to determine what level of difficulty the next item should have to be the most informative
 - 4. **Test** younger and older adults via adaptive cognitive tests instead
 - 5. Change the world of cognition



My Partners in Crime (and Estimation)

How does it work?

- Display items and collect responses in Visual Basic
- Custom MCMC algorithm in Fortran to predict theta after each item given precalibrated item properties and presentation time
- Administer most relevant item from item bank next
- Increase or decrease to presentation time to fill in gaps in item difficulty
- So let's see how we did...



SE for Ability by Trial: n=34 OA



Hoffman UNL Faculty Seed Grant

Adaptive Ability by Trial: n=34 OA



Hoffman UNL Faculty Seed Grant

SE for Ability by Trial: *n*=175 YA



Hoffman UNL Faculty Seed Grant

Adaptive Ability by Trial: n=175 YA



When More is Not Better...



Adaptive Search: My Legos to Ponder

1. Linear models:

- "Explained" item variance in making new items by age
- > Is using presentation time to fill in the gaps ok to do?

2. Estimation:

Converging evidence across ML and MCMC methods

3. Link functions:

Need better match of forced-choice format (chance = 50%)

4. Random / latent effects for multidimensionality:

 Individual differences in effects of item features via additional latent variables (random slopes or latent attributes)

Thank you for your attention!

Questions or comments? Lesa-Hoffman@Ulowa.edu