Framing Item Response Models as Hierarchical Linear Models

Measurement Incorporated
Hierarchical Linear Models Workshop

Overview

- Nonlinear Item Response Theory (IRT) models.
- Conceptualizing IRT models as hierarchical generalized linear models.
- Comments on estimation for such models.

Item Response Theory Models

- To facilitate our discussion today, let me start by introducing two common IRT models: the one- and two-parameter logistic model:
  - For brevity, we omit the scaling constant from both of these.

- 1PL (or Rasch Model):

$$P(Y_{ij} = 1) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$

- 2PL:

$$P(Y_{ij} = 1) = \frac{\exp(a_i(\theta_j - b_i))}{1 + \exp(a_i(\theta_j - b_i))}$$

- With θ_j as the ability for examinee j, b_i the difficulty for item i, and a_i the discrimination for item i.

Rephrasing IRT Models

- To show how IRT models fit into the HGLM framework, some reorganization must first take place:
  - Work only with logits (η) rather than probabilities.
  - Move from traditional parameterization to slope/intercept (similar to logistic regression):

- 1PL (or Rasch Model):

$$P(Y_{ij} = 1) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$

Here $\beta_j = -b_j$

Now model has an error term which is distributed $N(0,\pi^2/3)$

$$Logit[P(Y_{ij} = 1)] = \eta_{ij} = \theta_j - b_i + \varepsilon_{ij} = \beta_i + \theta_j + \varepsilon_{ij}$$

Rephrasing IRT Models

- To show how IRT models fit into the HGLM framework, some reorganization must first take place:
  - Work only with logits (η) rather than probabilities.
  - Move from traditional parameterization to slope/intercept (similar to logistic regression):
- 2PL:

$$P(Y_{ij}=1)=\frac{\exp(a_i(\theta_j-b_i))}{1+\exp(a_i(\theta_j-b_i))}$$

$$Logit[P(Y_{ij}=1)]=\eta_{ij}=a_i(\theta_j-b_i)+\varepsilon_{ij}=\beta_i+\lambda_i\theta_j+\varepsilon_{ij}$$

Here $\beta_i = -a_i b_i$

Here $\lambda_i = a_i$

Now model has an error term which is distributed $N(0,\pi^2/3)$

Nonlinear Item Response Models

- Now we have reshaped IRT models, we will map them onto HGLMs
  - First, we will use the notation from Raudenbush and Byrk.
    - Accomplished by referent items and dummy codes.
- I am going to only go over the most basic case where we have a one parameter item response model.
- However, you should know that these models can be more difficult and in staying true with the Rasch type models it is simply a matter of developing dummy coded variables.

Nonlinear Item Response Models

- So why might we want to use HLM for something like this?

- The book actually gives 6 reasons:

  - Facilitates the study of multidimensional assessment.
  - Naturally incorporates variability between social settings.
  - Incorporates explanatory variables at several levels.
  - Provides a natural framework for studying measurement error.
  - Latent variables can be studied as explanatory variables.
  - Provides a natural way to deal with nonresponses.

Nonlinear Item Response Models

- So to start, in this case we assume that we have dichotomous responses to items:
  - Items are coded as either correct or incorrect (1/0).

- So there are *I* items (indexed by i) and *J* examinees (indexed by j)

- We assume that the probability (or log-odds) of a response to an item is a function of a persons ability and the difficulty of that item.

Level 1 Model

- For our level-1 model, we would like to predict the probability an examinee j answers an item i correctly.
  - We will use the logit representation to accomplish this.

- Let's start with a level-1 model where items are nested within person.

- Level-1 Model: $$\eta_{ij} = \pi_{0i} + \varepsilon_{ij}$$

Level-1 HGLM for IRT

$$\eta_{ij} = \pi_{0i} + \varepsilon_{ij}$$

- π_{0i} is the intercept (different from R & B's notation, meant to be consistent with 1- and 2-PL models).
- Level-1 error is distributed $N(0, \pi^2/3)$.
  - This comes from the logistic distribution for η_{ij}.

Level-2 Equations

- Next, we will assume that a person's ability varies and an items difficulty is the same for all people.

- So using HLM that is: $\pi_{0i} = \beta_{0i} + \theta_{0j}$
  - β_{0i} is the intercept (item difficulty) for item i.
  - θ is the Level-2 error term.
    - The random intercept for examinee j.
  - We will discuss the distribution of θ on the next slide.
- Putting the level-1 and level-2 models together we get, for an item j, the original 1PL (or Rasch) model:

$$\eta_{ij} = \pi_{0i} + \varepsilon_{ij} = \beta_{0i} + \theta_{0j} + \varepsilon_{ij}$$

HLM Versus IRT Distinctions

- The key distinction between IRT and HLM comes from the distributional assumptions placed on θ.
- In HLM, level-2 error is typically said to be $N(0, \tau)$.
- In IRT, θ typically is said to be $N(0,1)$.
- In both, β_i is a fixed parameter called item difficulty (or the intercept).

Item Response Models

- So in this we can see that in HLM the differences across people are summarized in τ.

- Also, we should note that because we are using the logit link everything is in the log-odds scale

- What makes this nice is that now we could see how by adding a third level (school) we could start to model:
  - Students nested within classroom/school/district/county/state.
  - Student growth over time.

- By adding other level-2 variables, we can start to "explain" the difficulty of an item.
  - See de Boeck and Wilson's "Explanatory IRT Models" book.

Mapping the 2PL onto HGLM

- Because of the discrimination parameter, mapping the 2PL onto an HGLM is a bit more complicated.
  - We now need an additional structure for the covariance of the error terms.
- The basic two model equations still apply (in mixed form):

$$\eta_{ij} = \pi_{0i} + \varepsilon_{ij} = \beta_{0i} + \theta_{0j} + \varepsilon_{ij}$$

Covariance of Error Terms

- Now we say that across items, the covariance matrix of ε_{ij} is below.
  - Here, λ is the item slope for item i.
- This is a heterogeneous error model.

$$\mathrm{cov}(\varepsilon_{ij}) = \begin{bmatrix} \sigma_\varepsilon^2 & \lambda_1\lambda_2 & \cdots & \lambda_1\lambda_I \\ \lambda_1\lambda_2 & \sigma_\varepsilon^2 & \cdots & \lambda_2\lambda_I \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1\lambda_I & \lambda_2\lambda_I & \cdots & \sigma_\varepsilon^2 \end{bmatrix}$$

IRT in HLM Example

- To show how to estimate an IRT model in the HLM package, we present an example.
- Data are a set of 10 items from an 8^{th} grade End-Of-Grade reading assessment.
  - From a small Midwestern state.
  - Total of 5573 students taking a pencil-and-paper form.
    - Mainstream students – without IEP or ESL.

Data File Setup

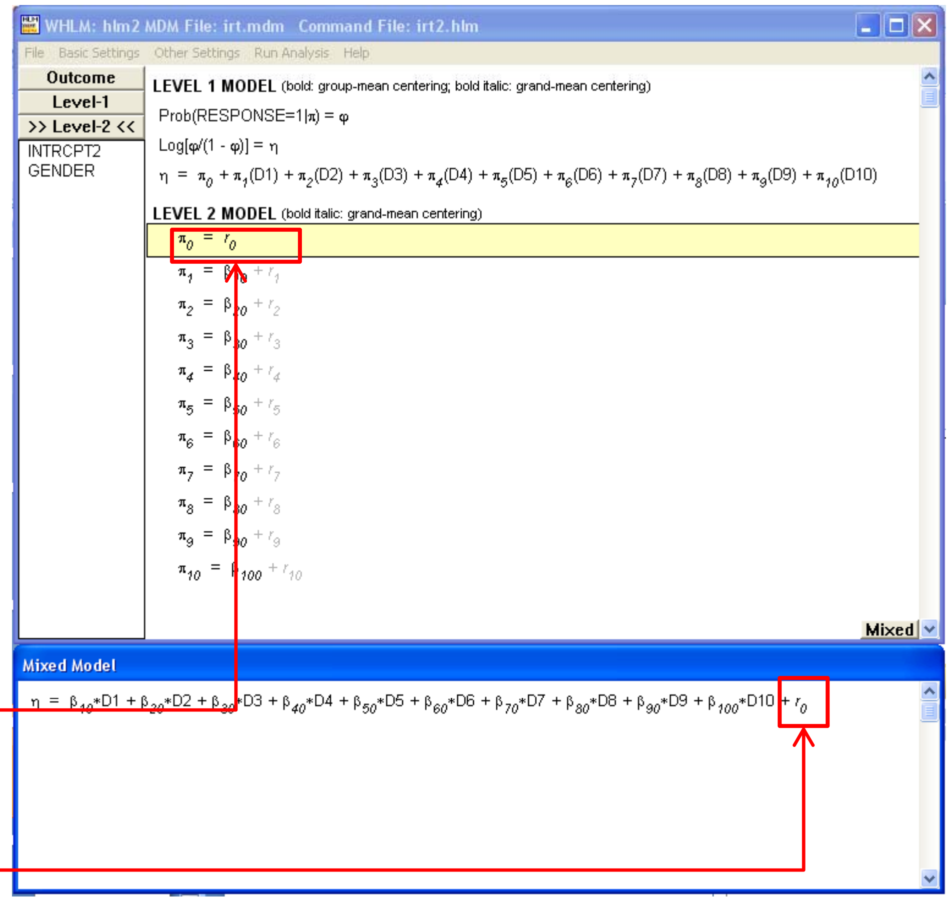
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	state_id	usd	bldg	gender	total	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	item	response
2	1003770355	305	3022	0	6	1	0	0	0	0	0	0	0	0	0	1	1
3	1003770355	305	3022	0	6	0	1	0	0	0	0	0	0	0	0	2	1
4	1003770355	305	3022	0	6	0	0	1	0	0	0	0	0	0	0	3	1
5	1003770355	305	3022	0	6	0	0	0	1	0	0	0	0	0	0	4	1
6	1003770355	305	3022	0	6	0	0	0	0	1	0	0	0	0	0	5	0
7	1003770355	305	3022	0	6	0	0	0	0	0	1	0	0	0	0	6	0
8	1003770355	305	3022	0	6	0	0	0	0	0	0	1	0	0	0	7	1
9	1003770355	305	3022	0	6	0	0	0	0	0	0	0	1	0	0	8	0
10	1003770355	305	3022	0	6	0	0	0	0	0	0	0	0	1	0	9	1
11	1003770355	305	3022	0	6	0	0	0	0	0	0	0	0	0	1	10	0

- Data is in “long” format – each item response has it’s own row.
  - Variable “response” stores item response (0/1).

- Dummy variables (d1-d10) indicate which item “response” is the response.

HLM Setup

- Because of the setup of the HLM program, we have to be somewhat selective when entering our data.
  - Enter all dummy variables into the level-1 equation.
  - Remove the level-2 intercept fixed effect term (β_{00}).
  - Ability parameter is level-2 error (r_0).



HLM IRT Model Output: Variance Components

- First, we can look at the results for our Level-2 variance (τ_{00}).
  - This is the variance of the latent trait (θ).

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	R0	0.55302	0.30583	5574	9608.61581	0.000

  - τ_{00} = 0.306, meaning $\theta \sim N(0, 0.306)$.

HLM IRT Model Output: Fixed Effects

- The fixed effects give us the item difficulty values.
  - Recall, difficulty from HLM is really -1 times the actual difficulty.
    - Item easiness parameterization (higher values mean easier items).

Final estimation of fixed effects
(Unit-specific model with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For D1 slope, P1					
INTRCPT2, B10	0.384489	0.028065	13.700	55730	0.000
For D2 slope, P2					
INTRCPT2, B20	0.558127	0.028557	19.544	55730	0.000
For D3 slope, P3					
INTRCPT2, B30	-0.045281	0.027611	-1.640	55730	0.101
For D4 slope, P4					
INTRCPT2, B40	0.124324	0.027658	4.495	55730	0.000
For D5 slope, P5					
INTRCPT2, B50	0.090195	0.027639	3.263	55730	0.001
For D6 slope, P6					
INTRCPT2, B60	-1.161956	0.031887	-36.439	55730	0.000
For D7 slope, P7					
INTRCPT2, B70	0.259437	0.027818	9.326	55730	0.000
For D8 slope, P8					
INTRCPT2, B80	-0.261952	0.027824	-9.414	55730	0.000
For D9 slope, P9					
INTRCPT2, B90	-0.143912	0.027681	-5.199	55730	0.000
For D10 slope, P10					
INTRCPT2, B100	0.257183	0.027824	9.243	55730	0.000

HLM IRT Model Output: Fixed Effects

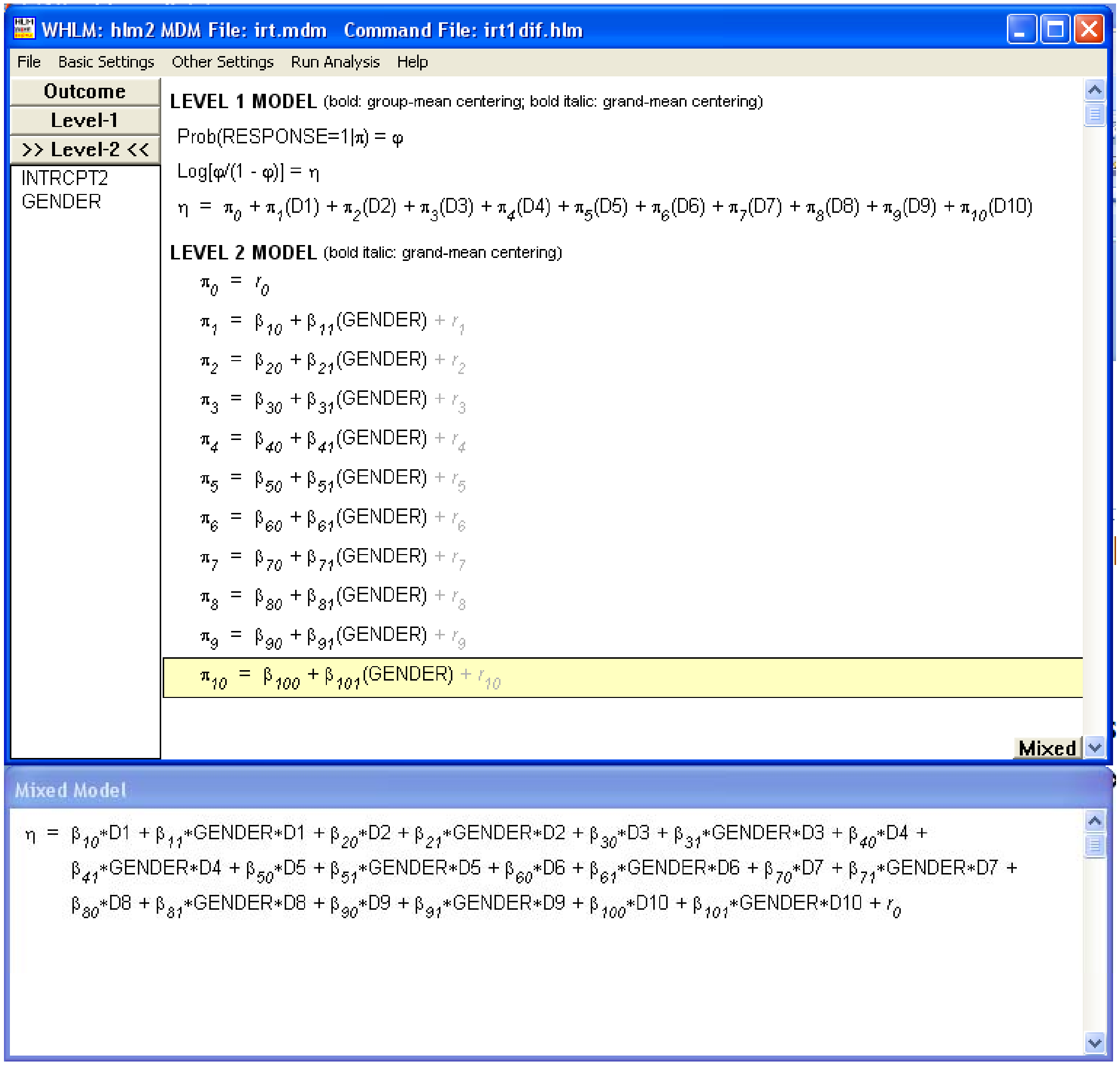
Item i	Easiness (β_{i0})
1	0.384
2	0.558
3	-0.453
4	0.124
5	0.090
6	-1.162
7	0.259
8	-0.262
9	-0.144
10	0.257

```
--------------------------------------------
  Fixed Effect                  Coefficient
--------------------------------------------
For        D1 slope, P1
   INTRCPT2, B10                   0.384489
For        D2 slope, P2
   INTRCPT2, B20                   0.558127
For        D3 slope, P3
   INTRCPT2, B30                  -0.045281
For        D4 slope, P4
   INTRCPT2, B40                   0.124324
For        D5 slope, P5
   INTRCPT2, B50                   0.090195
For        D6 slope, P6
   INTRCPT2, B60                  -1.161956
For        D7 slope, P7
   INTRCPT2, B70                   0.259437
For        D8 slope, P8
   INTRCPT2, B80                  -0.261952
For        D9 slope, P9
   INTRCPT2, B90                  -0.143912
For       D10 slope, P10
   INTRCPT2, B100                   0.257183
--------------------------------------------
```

HLM IRT Example Extension

- Now, we will demonstrate how to assess DIF in an HLM context.
- Now we would like to check for differences in item difficulty as a function of the gender of an examinee.
- Gender is a level-2 variable.
  - We have ours dummy-coded (male = 1, female = 0).
- Hypothesis test for parameters will indicate DIF for each item.

HLM IRT DIF Setup



HLM IRT DIF Model Output: Variance Components

- First, we can look at the results for our Level-2 variance (τ_{00}).
  - This is the variance of the latent trait (θ).

```
Final estimation of variance components:
-----------------------------------------------------------------------------
Random Effect              Standard       Variance      df     Chi-square  P-value
                           Deviation      Component
-----------------------------------------------------------------------------
INTRCPT1,        R0          0.55474        0.30773   5574     9622.87365     0.000
-----------------------------------------------------------------------------
```

  - τ_{00} = 0.307, meaning $\theta \sim N(0, 0.307)$.
  - This is different by 0.001 from before.
    - A quirk of the estimation algorithm – more on that later.

HLM IRT Model Output: Fixed Effects

- The fixed effects give us the item difficulty values.
- The GENDER variable provides the difference in difficulty value for the males.
- The p-value of GENDER allows for the hypothesis test of DIF for each item.

Final estimation of fixed effects
(Unit-specific model with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For D1 slope, P1					
INTRCPT2, B10	0.447785	0.039296	11.395	55720	0.000
GENDER, B11	-0.129784	0.056184	-2.310	55720	0.021
For D2 slope, P2					
INTRCPT2, B20	0.526529	0.039601	13.296	55720	0.000
GENDER, B21	0.065834	0.057180	1.151	55720	0.250
For D3 slope, P3					
INTRCPT2, B30	0.010658	0.038425	0.277	55720	0.781
GENDER, B31	-0.115640	0.055287	-2.092	55720	0.036
For D4 slope, P4					
INTRCPT2, B40	0.014962	0.038435	0.389	55720	0.697
GENDER, B41	0.226762	0.055448	4.090	55720	0.000
For D5 slope, P5					
INTRCPT2, B50	0.000616	0.038435	0.016	55720	0.987
GENDER, B51	0.185484	0.055380	3.349	55720	0.001
For D6 slope, P6					
INTRCPT2, B60	-1.141047	0.044184	-25.825	55720	0.000
GENDER, B61	-0.043948	0.063845	-0.688	55720	0.491
For D7 slope, P7					
INTRCPT2, B70	0.212233	0.038626	5.495	55720	0.000
GENDER, B71	0.097950	0.055703	1.758	55720	0.078
For D8 slope, P8					
INTRCPT2, B80	-0.132979	0.038513	-3.453	55720	0.001
GENDER, B81	-0.268775	0.055838	-4.813	55720	0.000
For D9 slope, P9					
INTRCPT2, B90	-0.074014	0.038466	-1.924	55720	0.054
GENDER, B91	-0.144810	0.055444	-2.612	55720	0.009
For D10 slope, P10					
INTRCPT2, B100	0.410383	0.039167	10.478	55720	0.000
GENDER, B101	-0.313385	0.055820	-5.614	55720	0.000

HLM IRT Model Output: Fixed Effects

- Low p-values indicate significant differences in item difficulty for each gender.
- The effect size (in logits) is the estimate for GENDER.

Final estimation of fixed effects
(Unit-specific model with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For D1 slope, P1					
INTRCPT2, B10	0.447785	0.039296	11.395	55720	0.000
GENDER, B11	-0.129784	0.056184	-2.310	55720	0.021
For D2 slope, P2					
INTRCPT2, B20	0.526529	0.039601	13.296	55720	0.000
GENDER, B21	0.065834	0.057180	1.151	55720	0.250
For D3 slope, P3					
INTRCPT2, B30	0.010658	0.038425	0.277	55720	0.781
GENDER, B31	-0.115640	0.055287	-2.092	55720	0.036
For D4 slope, P4					
INTRCPT2, B40	0.014962	0.038435	0.389	55720	0.697
GENDER, B41	0.226762	0.055448	4.090	55720	0.000
For D5 slope, P5					
INTRCPT2, B50	0.000616	0.038435	0.016	55720	0.987
GENDER, B51	0.185484	0.055380	3.349	55720	0.001
For D6 slope, P6					
INTRCPT2, B60	-1.141047	0.044184	-25.825	55720	0.000
GENDER, B61	-0.043948	0.063845	-0.688	55720	0.491
For D7 slope, P7					
INTRCPT2, B70	0.212233	0.038626	5.495	55720	0.000
GENDER, B71	0.097950	0.055703	1.758	55720	0.078
For D8 slope, P8					
INTRCPT2, B80	-0.132979	0.038513	-3.453	55720	0.001
GENDER, B81	-0.268775	0.055838	-4.813	55720	0.000
For D9 slope, P9					
INTRCPT2, B90	-0.074014	0.038466	-1.924	55720	0.054
GENDER, B91	-0.144810	0.055444	-2.612	55720	0.009
For D10 slope, P10					
INTRCPT2, B100	0.410383	0.039167	10.478	55720	0.000
GENDER, B101	-0.313385	0.055820	-5.614	55720	0.000

Estimation Issues

- The HLM program uses an estimation method called Penalized Quasi-Likelihood (PQL).
  - Approximates the maximum likelihood function.
  - This method can produce biased results.
  - Can be very unstable because of complicated integral.

- For this reason, we recommend *not* using HLM to fit IRT models.

- Instead try the following:
  - Mplus
  - SAS proc nlmixed
  - Bayesian methods in R (i.e. glmmgibbs package).