Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 7, March 2015 ISSN 1531-7714

Item Response Theory Using Hierarchical Generalized Linear Models

Hamdollah Ravand, *Vali-e-Asr University of Rafsanjan*

Multilevel models (MLMs) are flexible in that they can be employed to obtain item and person parameters, test for differential item functioning (DIF) and capture both local item and person dependence. Papers on the MLM analysis of item response data have focused mostly on theoretical issues where applications have been add-ons to simulation studies with a methodological focus. Although the methodological direction was necessary as a first step to show how MLMs can be utilized and extended to model item response data, the emphasis needs to be shifted towards providing evidence on how applications of MLMs in educational testing can provide the benefits that have been promised. The present study uses foreign language reading comprehension data to illustrate application of hierarchical generalized models to estimate person and item parameters, differential item functioning (DIF), and local person dependence in a three-level model.

Data in social sciences in general and educational measurement in particular have a hierarchical structure. In other words, students are nested in classes which are in turn nested in schools. Nested data are locally dependent. As a result, the average correlation between variables measured on students from the same school/class will be higher than the average correlation between variables measured on students from different schools/classes. The within-class correlations would be, for example, due to a common teacher, the same syllabus, or the same textbook, and within-school correlations, among other things, may be the result of a common set of administrative policies or the selection processes (for example, some schools may select highly talented students or some may attract students form either high or low social economic status levels). Due to these clustering effects, a fundamental assumption underlying a majority of parametric statistical tests is violated (Goldstein, 1995; Raudenbush & Bryk, 2002). *Local independence assumption* holds that there should be no relationship among individuals in the sample for the dependent variable once the effect of the independent variable has been taken into account.

Non-independence assumption is usually taken for granted in conventional statistical tests such as regression and ANOVA. Violation of this assumption leads to underestimation of standard errors (SE). Since statistical significance of a predictor variable is judged by the ratio of its size to its SE (a significant coefficient should be at least twice as big as its SE), an underestimated SE would result in obtaining a significant effect when it does not really exist (Hox, 2010; Raudenbush &Bryk, 2002). Multilevel models (MLM) have been designed to handle interdependencies among the data points. In what follows first MLMs in general and hierarchical generalized linear models (HGLM) in particular are described. Then the ways they have been and could be used in educational testing is reviewed. Finally, the application of the Rasch HGLM with the HLM software is demonstrated and the outputs are interpreted.

Multilevel Models

MLMs have been differently termed as linear mixed models (Littell, Milliken, Stroup, &Wolfinger, 1996), hierarchical linear models (Raudenbush & Bryk 1986), and random coefficient models (Longford 1993). MLMs are extensions of standard multiple regression. They specifically account for dependency in the data with simultaneous multiple regressions at different levels. To analyze relationships between a dependent and independent variables in a hierarchical data set where, for example, students are nested in classes which are in turn nested in schools, a three-level regression model is formulated. The first and the lowest level is the student level, the second level is class level, and the third level is school level. It goes without saying that this 3-level model can be extended to include a fourth level (e.g., neighborhood level). In all the MLMs there is a single outcome or response variable which is measured at the lowest level and there could be explanatory variables at all the levels. For example, imagine we want to explore the effect of factors (i.e., knowledge of vocabulary and knowledge of grammar) that might affect foreign language reading comprehension. Suppose further, the data have been collected from j universities (*j*=1…J) with n_j students in each university. The first level (i.e., student-level) regression equation can be set up as in Equation 1:

$$reading = \beta_{0j} + \beta_{1j} vocabulary_{ij} + \beta_{2j} grammar_{ij} + e_{ij} \quad (1)$$

where *reading* is the outcome, β_{0j} is the *intercept* (i.e., the mean reading comprehension of university j), β_{1j} and β_{2j} are the *slopes* (i.e., the mean effects of the person-explanatory variables of vocabulary and grammar, respectively, on reading comprehension in university j) and e_{ij} represents the deviation of reading comprehension of student i from the intercept (the mean reading comprehension of his/her respective university). Equation 1 is different from a standard multiple regression in that, unlike in standard multiple regression where we assume regression coefficients (i.e., intercepts and slopes) are constant[1] (i.e., *fixed*) across all the students regardless of the university they belong to, in a MLM each cluster (here university) can be assumed to have a different intercept coefficient β_{0j} (here mean reading comprehension) and different slope coefficients β_{1j} and β_{2j} (here mean impact of vocabulary and grammar, respectively, on reading comprehension). Put another way, the intercept and the slopes are assumed to be *random* (i.e., vary) across universities. The group-specific coefficients are indicated by subscript j attached to each coefficient. With MLMs researchers can test whether the coefficients i.e., mean reading comprehension (intercept) and mean impact of vocabulary and grammar, respectively, on reading comprehension (the slopes) vary significantly across universities.

The next step in MLM procedure is to explain randomness (i.e., variation) in the intercept and slopes across the higher level units (in this case universities). The level-1 coefficients which are assumed to vary across higher units are set up as outcome variables in the level-2 equations. In the present case, since the intercept (the mean reading comprehension) and the slopes (the mean impact of vocabulary and grammar) were assumed to vary across universities, university-level explanatory variables (i.e., covariates) can be added to account for the variations of these coefficients at the second level. For example, universities in Iran, depending on whether they select students through screening tests, are divided into two broad categories: state universities and non-state universities. The subscripts in Equation 1 show that the intercept (the mean reading comprehension in each university) and the slopes (the mean impact of vocabulary and grammar in each university) are random across universities. Therefore at second level we need to have three regression equations one for the intercept as the outcome variable and two for the slopes as outcome variables as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \quad (2)$$

$$\begin{aligned} \beta_{1j} &= \gamma_{10} + \gamma_{11} Z_j + u_{1j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21} Z_j + u_{2j} \end{aligned} \quad (3)$$

Where γ_{00} is the average reading comprehension across all the universities (grand mean), γ_{10} and γ_{20} are the mean effect of vocabulary and grammar,

[1] Of the different ways fixed and random effects have been conceptualized, I have adopted the distinction by Kreft and De Leeuw (1998, p. 12).

respectively, across all the universities, u_{0j} represents deviation of the mean of the university *j* from the grand mean, u_{1j} and u_{2j} represent the deviations of the mean impact of vocabulary and grammar, respectively, on reading in university *j* from the respective means across all the universities. Equation 2 predicts average reading comprehension in each university (the intercept β_{0j}) by university type (Z) (i.e., state vs. non-state). As university type is a binary explanatory variable, in this case coded as zero for non-state universities and one for state universities, a positive γ_{01} indicates that the average reading comprehension is higher in state universities. Equations in 3 state that the relationship between grammar and reading comprehension and vocabulary and reading comprehension depend on the university type. Negative values for γ_{11} and γ_{21} indicate that the effect of vocabulary and grammar, respectively, on reading comprehension are stronger for non-state universities. Conversely, positive values for γ_{11} and γ_{21} indicate that the effect of vocabulary and grammar, respectively, on reading comprehension are stronger for state universities. If the variances of the *u*-terms u_{0j} ,u_{1j}, and u_{2j} are significant, more university-level covariates should be added to capture the variations.

Alternatively, the slopes can be assumed *fixed*, that is the effect of vocabulary and grammar can be assumed as being the same across universities. In that case β_{1j} and β_{2j} should be included into the level-1 model without the subscripts *j*. Accordingly, we would not need β_{1j} and β_{2j} equations at Level 2.

Hierarchical Generalized Linear Models

MLMs assume a continuous dependent variable with a normal distribution. However, if the dependent variable is a dichotomous variable, both the continuous dependent variable and the normality assumptions are violated (Hox, 2010). For situations where the dependent variable is non-normal non-continuous and the relationship between the predictor variable and the dependent variable is not linear, a variant of MLMs called *hierarchical generalized linear model* (HGLM) is appropriate.

Standard Rasch model has been shown to be a special case of HGLM (e.g., McClellan & Donoghue, 2001; Kamata, 2001, 2002; Miyazaki, 2005; Raudenbush, Johnson, & Sampson, 2003; Williams & Beretvas, 2006). In the Rasch model formulation of the HGLM (Rasch HGLM) item response data are treated as repeated observations where each test taker responds to multiple items. Multiple responses from the same subject cannot be regarded as independent from each other. As a case, take the reading comprehension example above. Each person possesses a different level of reading comprehension which is going to affect all the responses from the same person thus rendering these different responses inter-dependent rather than independent. In the Rasch HGLM item response data are treated as hierarchical data, where items are nested within persons. Unlike MLMs where persons are level one, in the Rasch HGLM items are level one and persons are level two. In what follows it is shown how the standard Rasch model can be derived from HGLM.

Hierarchical generealized Rasch model

Kamata (2001) showed how Rasch model can be formulated within the framework of a hierarchical model. In his formulation the first level is an item level model and the level-2 model is a person level model.

The level-1 model, the item level model, for item *i* (*i* =1,…, *k*) and person *j* (j = , …, *n*) is

$$\log(\frac{p_{ij}}{1-p_{ij}}) = \beta_{0j} + \beta_{1j}X_{1ij} + ...\beta_{(k-1)j}X_{(k-1)ij}$$
$$= \beta_{0j} + \sum_{q=1}^{k-1}\beta_{qj}X_{qij} \qquad (4)$$

where X_{ij} is the i^{th} item indicator which is a dummy variable for person *j*, with values 1 when the observation is the i^{th} item and 0 otherwise.

$\log(\frac{p_{ij}}{1-p_{ij}})$ is the logit *link fucntion* whereby the log odds of getting item *i* correct for the person *j* is predicted. p_{ij} is the probability that person *j* succeeds on item *i* and $1-p_{ij}$ is the probability of failure on the item. In HGLMs the link function linearizes the relationship between the predictor variables and the dependent variable and restricts the range of the predicted values to match the distribution of the dependent variable.

To *identify* the item-level model, some constraints need to be imposed. Kamata (2001) suggested that identification be carried out by using one of the items as the reference item thereby dropping the dummy variable for that item form the model and including an intercept term which is the effect of the reference item. In Equation 4, β_{0j} is an intercept term which is the expected effect of the reference item for person *j*, and β_{1j} to $\beta_{(k-1)j}$ are coefficients associated with the effect of Item 1 to Item *k*-1 (one less than the number of items since the dummy variable for the reference item is dropped).

In the Rasch HGLM at Level 2, the intercept β_{0j} is assummed to be varying (random) across persons (the random intercept is introduced to take inter-dependencies among items answered by the same person) and the other items' effects are assummed to be constant (fixed) across persons as in Equation 5:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + r_{0j}\\ \beta_{1j} &= \gamma_{10}\\ &.\\ &.\\ &.\\ \beta_{(k-1)j} &= \gamma_{(k-1)0}\end{aligned} \qquad (5)$$

As a result, level-1 and level-2 models can be combined so that the probability that pereson *j* answers item *i* correctly is expressed as:

$$p_{ij=}\frac{1}{1+\exp[-\{r_{0j}-(\gamma_{i0}-\gamma_{00})\}} \qquad (6)$$

Equation 7 is algebrically equall to the Rasch model:

$$\frac{1}{1+\exp[-(\theta_j-\delta_i]}, \qquad (7)$$

where θ_j, person ability in Equation 7, is equivalent to r_{0j}, the random effect of the intercept and δ_i item difficulty in Equation 7 is equivalent to $\gamma_{i0-}\gamma_{00}$ (the effect of each item subtracted from that of the reference item).

Kamata's formulation can be extended to three- and four-level HGLMs.

HGLM vs. Conventional Item Response Theory Models

Educational measurement data can be used to fulfill two broad purposes (De Boeck & Wilson, 2004): (a) to describe performance of individual test takers on a test and (b) to explain item responses in terms of other explanatory variables. Accordingly, the two purposes lead to two approaches: *descriptive measurement* approach and *explanatory* approach. The explanatory approach is broader and can be seen as complementary to the descriptive approach. Conventional IRT models are suitable for the fulfillment of the descriptive purpose. Although they can be used to serve the explanatory purpose, this can only be carried out in a two-step procedure (De Boeck & Wilson, 2004), with measurement as the first step and correlating the derived test scores with external variables as the next step. Unlike conventional IRT models which estimate person abilities first and then investigate the effect of person-varying explanatory variables on ability estimates, HGLMs take a one-step approach to investigating the effect of person variables on ability estimates (Kamata, 1998). Therefore, estimates of item and person parameters are expected to be more precise (Mislevey, 1987).

Most existing IRT models are special cases of HGLMs. A HGLM perspective on item response data broadens the domain of IRT models and facilitates their explanatory uses beyond their standard descriptive uses (De Boeck & Wilson, 2004). The HGLM perspective has additional advantages over the standard IRT approach (De Boeck & Wilson, 2004, p.7): (a) The approach is a general one and therefore also flexible, and (b) The approach connects psychometrics strongly to the field of statistics, so that a broader knowledge basis and literature become available and (c) The availability of generalized statistical software makes the implementation of new models developed for specific situations much more straightforward than it has been in the past, where specific-purpose programs could be used only under less general circumstances. More specifically, HGLMs are flexible in that they can be conveniently used to estimate item and person parameters, differential item functioning (DIF), and the effect of person-level predictors on ability measures and item-level predictors on item difficulty, simultaneously. They can also be used to investigate local item dependence (i.e., testlet effect), local person dependence and differential testlet functioning.

Rasch HGLM extensions and Previous Applications

Papers on the Rasch HGLM have focused mostly on theoretical issues where applications have been add-ons to studies with a methodological focus. Kamata (2001) showed that HGLM is equivalent to the Rasch model. He also showed how the two-level HGLM can be extended to a three-level latent regression model which allows investigation of students across groups. Jiao, Wang, and Kamata (2005) extended Kamata's two-level model to capture item local dependence. Their three-level model can be used to estimate item difficulties, person abilities, DIF, and testlet effect. Jiao, Kamata, Wang, and Jin (2012) extended Jiao et al.'s (2005) model to a four-level model which permits simultaneous modeling of both item and local dependence. Beretvas and Walker (2012) developed a cross-classified multilevel model to handle testlet-based dependencies. Their model allowed simultaneous investigation of DIF and differential testlet functioning. Ravand (in press) employed Beretvas and Walkers' model to estimate LID, DIF, and differential testlet functioning in a high stakes reading comprehension test. Beretvas, Cawthon, Lockhart, and Kaye (2012) in a pedagogical paper explained similarities and difference of the two-level cross-classified model and the conventional two-level model. They applied the two models to investigate DIF and differential facet functioning (Meulders & Xie, 2004) in accommodated item scores. Van den Noortgate, De Boeck, and Meulders (2003) demonstrated how several common IRT models can be derived from the multilevel logistic model. Weirich, Hecht, and Bohme (2014) showed how item position effects can be modeled using the linear logistic test model within the framework of HGLMs. Debeer and Janssen (2013) proposed a general framework for detecting and modeling item position bias effects using explanatory IRT models in the framework of HGLM. Randall, Cheong, and Engelhard (2011) described how HGLMs can be used to investigate measurement invariance, specifically DIF, within the context of assessing students with disabilities. Albano (2013) also demonstrated how HGLMs can be employed to model item position effect. He argued that the HGLM approach is advantageous over the previously used models for this purpose in that it can estimate position effects simultaneously with item and person parameters.

Although the methodological direction was necessary as a first step to show how MLMs can be utilized and extended to model item response data, the emphasis needs to be shifted towards providing evidence on how applications of MLMs in educational testing can provide the benefits that have been promised.

Data Analysis

To illustrate the analysis procedures involved in the Rasch HGLM, university entrance examiantion (UEE) data of the applicants into the Masters' English programs at the Iranian state univerisites in 2012 are used. There were 21640 (71.3 % female and 26.8 % male) participants who took the test in this year. The participants received their Bachlors' degrees mostly from four univerisity types in Iran: (a) state universities which do not charge any tuition fees, (b) Azad universities which charge tuition fees (c) Non-profit Non-government universities which charge tuition fees, half as much as those of Azad universities, and (d) Payam-e-Noor universities which charge tuition fees as much as those of Non-profit Non-government universities but do not offer regualar classes. UEE is composed of two main sections namely general English (GE) section and content knowledge section. For the purpose of the present study the data for the reading section of the GE part of the UEE is analyzed. From among the 21640 participants 1298 persons were excluded from the analysis since they attended a university at Bachelor level which had less than 10 participants taking the test. The remaining 20342 students were from 227 universities.

Data analysis illustration is carried out using HLM 7.01 (Raudenbush, Bryk, Cheong, Fai, Congdon, & du Toit, 2013). HLM is a commercial software, two free versions of which are also available: (a) the student edition which is restricted in the size of the model one wants to analyze, (b) The 15-day tiral version which has got all the capabilities of the full version and can be requrested from the publishing company at http://www.ssicentral.com/hlm/downloads.html.

Data file preparation

Researchers usually input their data into SPSS in 'wide' format for conventional analysis purposes. For the purpose of Rasch HGLM analysis, the 'wide' data format should be restructured into the 'long' format. Moreover, for each item a dummy indicator should be created so that X_{qij} is the q^{th} dummy variable for person *j*, with values 1 when $q = i$, and 0 when $q \neq 1$, for item *i*. To convert the data into 'long' format, go to **Data** tab in the SPSS file which includes the data and then to **Restructure** as shown in Figure 1.

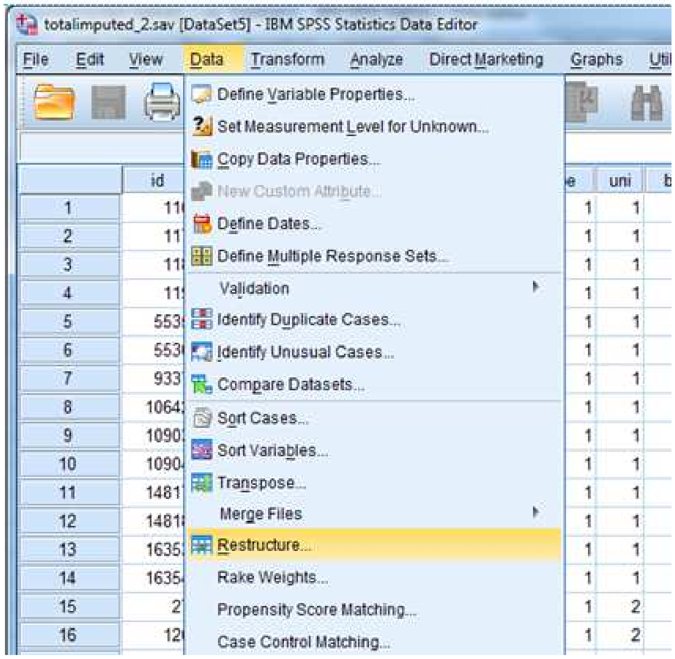
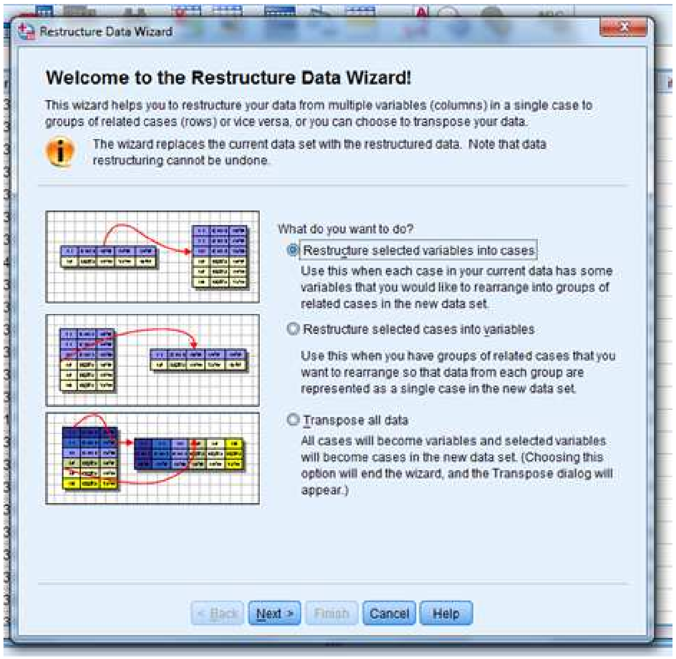


Figure 1. Restructuring 1.

Since the intent is to convert the 'long' form into 'wide' (i.e., convert variables into cases), in the next dialog box go with the SPSS's default (i.e. "Restructure selected variables into cases"), as shown in Figure 2 and then click **Next**.



Welcome to the Restructure Data Wizard!

This wizard helps you to restructure your data from multiple variables (columns) in a single case to groups of related cases (rows) or vice versa, or you can choose to transpose your data.

The wizard replaces the current data set with the restructured data. Note that data restructuring cannot be undone.

What do you want to do?

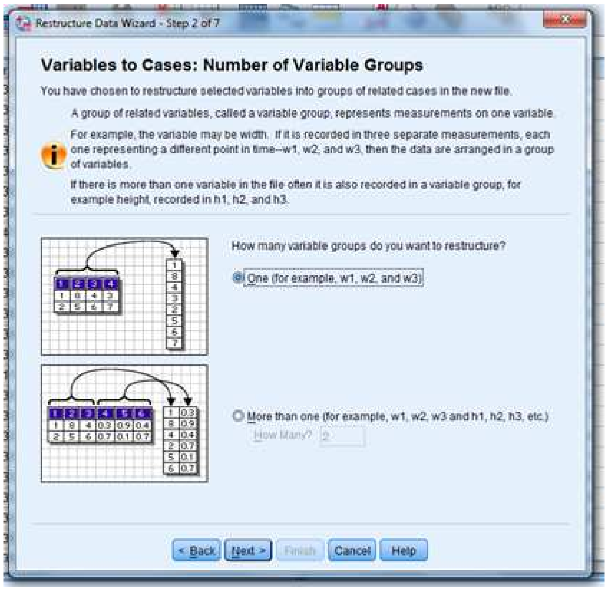
● Restructure selected variables into cases
Use this when each case in your current data has some variables that you would like to rearrange into groups of related cases in the new data set.

○ Restructure selected cases into variables
Use this when you have groups of related cases that you want to rearrange so that data from each group are represented as a single case in the new data set.

○ Transpose all data
All cases will become variables and selected variables will become cases in the new data set. (Choosing this option will end the wizard, and the Transpose dialog will appear.)

Figure 2. Restructuring 2.

Since the intention is to create just one new variable (i.e., item) from a set of columns in the old data file (i.e., Item1 through Item60), in the next dialog box leave the SPSS's default unchanged because it serves our purpose right and click **Next** as shown in Figure 3.



Variables to Cases: Number of Variable Groups

You have chosen to restructure selected variables into groups of related cases in the new file.

A group of related variables, called a variable group, represents measurements on one variable.

For example, the variable may be width. If it is recorded in three separate measurements, each one representing a different point in time--w1, w2, and w3, then the data are arranged in a group of variables.

If there is more than one variable in the file often it is also recorded in a variable group, for example height, recorded in h1, h2, and h3.

How many variable groups do you want to restructure?

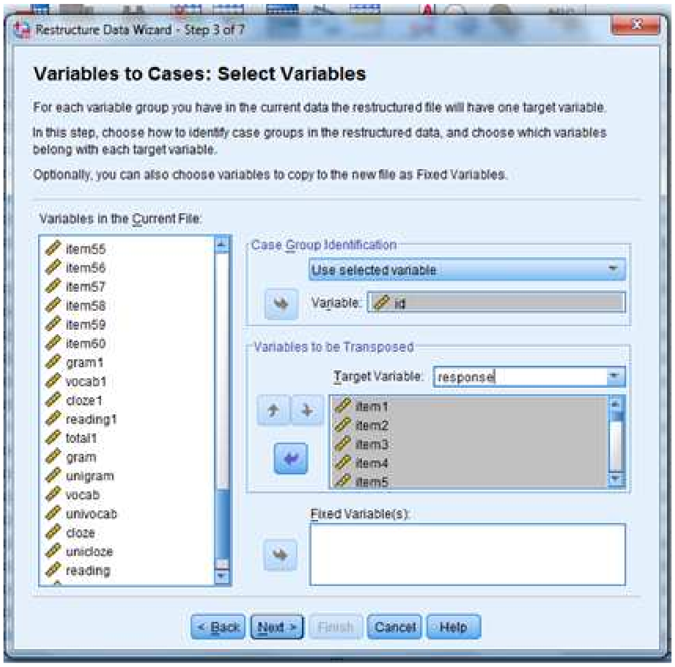
● One (for example, w1, w2, and w3)

○ More than one (for example, w1, w2, w3 and h1, h2, h3, etc.)
How Many? 2

Figure 3. Restructuring 3.

In the next window, from the Case Group Identification section choose **Use selected variable** and then select the **'id'** variable from the **Variables in the Current File** in the left box and move it into the box in front of the **Variable** in the right. Then choose

Item1 through Item60 variables from the **Variables in the Current File** box and move them into the box under the **Variables to be Transposed.** One can optionally change the name of the **Target variable** from trans1 (which is the default of SPSS) as shown in Figure 4.



Variables to Cases: Select Variables

For each variable group you have in the current data the restructured file will have one target variable.

In this step, choose how to identify case groups in the restructured data, and choose which variables belong with each target variable.

Optionally, you can also choose variables to copy to the new file as Fixed Variables.

Variables in the Current File: item55, item56, item57, item58, item59, item60, gram1, vocab1, cloze1, reading1, total1, gram, unigram, vocab, univocab, cloze, unicloze, reading

Case Group Identification: Use selected variable

Variable: id

Variables to be Transposed

Target Variable: response

item1, item2, item3, item4, item5

Fixed Variable(s):

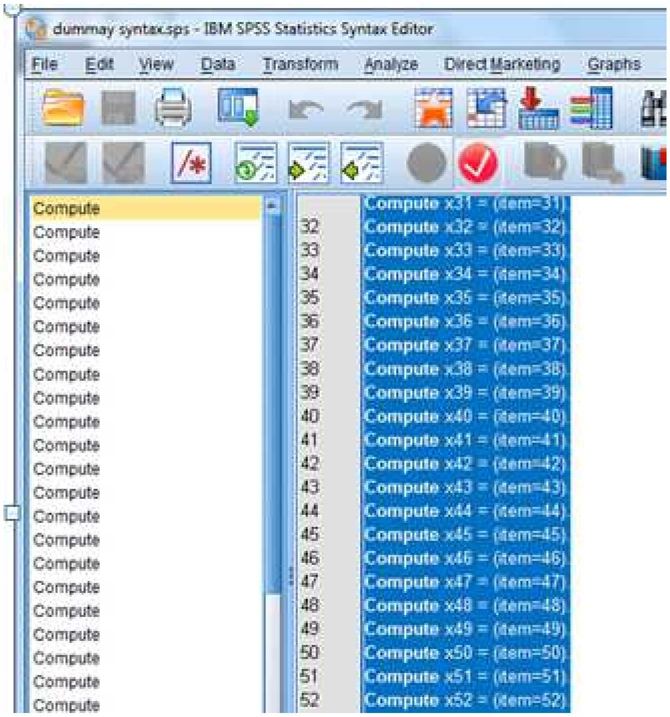
Figure 4. Restructuring 4.

In the next few dialog boxes, go with the SPSS's defaults and just click **Next** in each dialog box and you will come up with the 'long' data file.

In the next step, we need to create dummy variables for the items. It can be much more conveniently carried out through writing a set of commands in the SPSS Syntax Editor. From the File menu select **New** and then **syntax**. In the Syntax Editor window type the following commands as shown in Figure 5:

```
Compute x1 = (item=1).
Compute x2 = (item=2).
Compute x3 = (item=3).
Compute x4 = (item=4).
Compute x5 = (item=5).
.
.
.
Compute x60 = (item=60).
;EXECUTE.
```

Then select all the commands and from the menu above click on the **Run Selection** button, as shown in Figure 5. Running this command will return dummy variables X1 through X60 for Items 1 through 60.

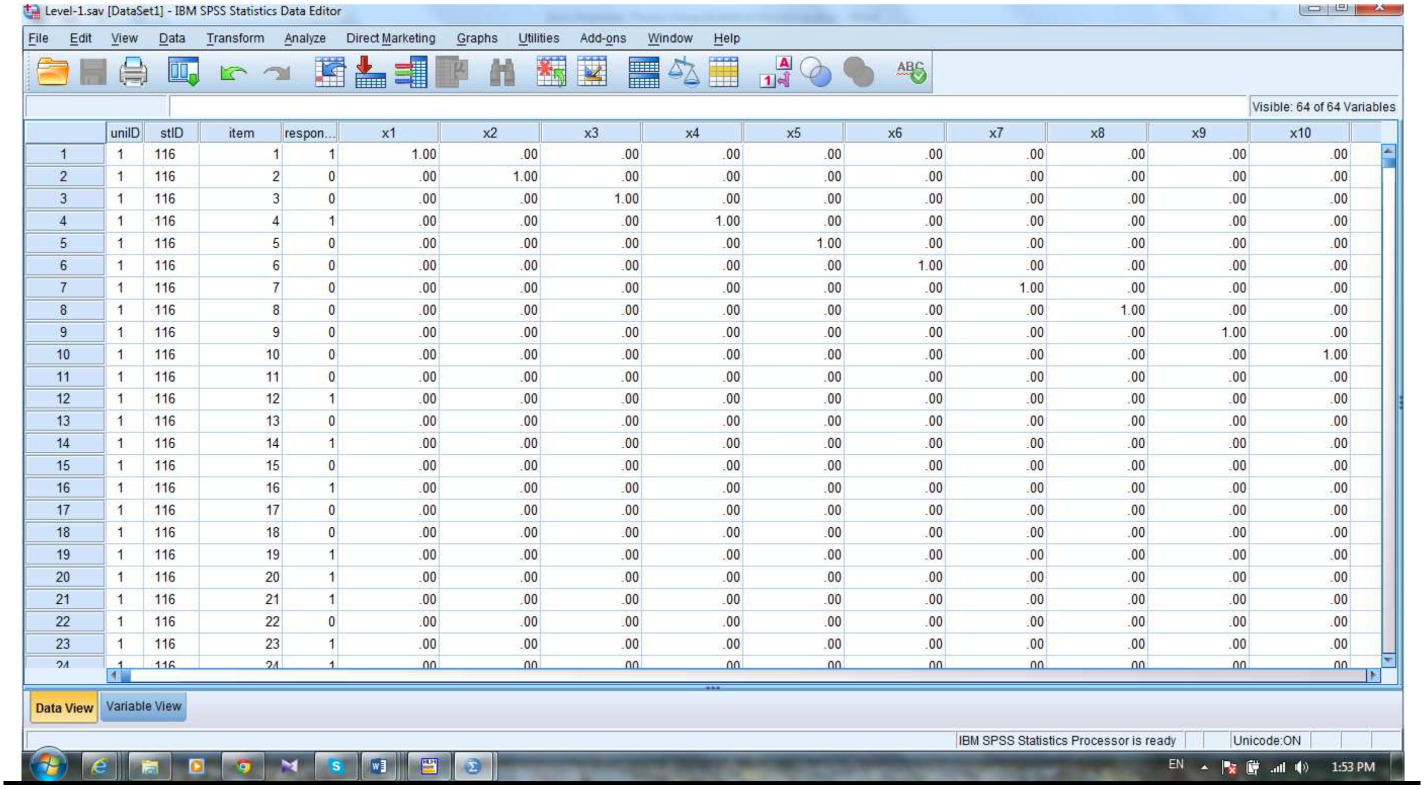


```
Compute x31 = (item=31).
Compute x32 = (item=32).
Compute x33 = (item=33).
Compute x34 = (item=34).
Compute x35 = (item=35).
Compute x36 = (item=36).
Compute x37 = (item=37).
Compute x38 = (item=38).
Compute x39 = (item=39).
Compute x40 = (item=40).
Compute x41 = (item=41).
Compute x42 = (item=42).
Compute x43 = (item=43).
Compute x44 = (item=44).
Compute x45 = (item=45).
Compute x46 = (item=46).
Compute x47 = (item=47).
Compute x48 = (item=48).
Compute x49 = (item=49).
Compute x50 = (item=50).
Compute x51 = (item=51).
Compute x52 = (item=52).
```

Figure 5. Dummy item coding.

HLM works with either separate data files for each level or a single data file which contains information on all the levels. For the purpose of illustration, information for each level was saved in a separate SPSS file. It is worthy of mention that HLM requires that data in all the data files should include an ID variable for the respective level and IDs for all the levels above it and they should be sorted according to the ID of the highest level. In a two-level model, for example, where the first level is the item level and the second level is the person level, both the item and person files should be sorted according to person ID. And in a three-level model where the third level is class or university, all the three data files should be sorted according to class/university ID. Figure 6 shows an excerpt of the item-level (level-1) data file.

As one can see the first level data file has been sorted according to the third-level IDs (i.e., uniID). The first column is university ID (uniID), the second column represents data on student ID (stID), the third column display item id (item) information, and the fourth column is a vector of item responses where 1


Level-1.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 64 of 64 Variables
Data View Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON
EN
1:53 PM

	uniID	stID	item	respon...	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	1	116	1	1	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
2	1	116	2	0	.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00
3	1	116	3	0	.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00
4	1	116	4	1	.00	.00	.00	1.00	.00	.00	.00	.00	.00	.00
5	1	116	5	0	.00	.00	.00	.00	1.00	.00	.00	.00	.00	.00
6	1	116	6	0	.00	.00	.00	.00	.00	1.00	.00	.00	.00	.00
7	1	116	7	0	.00	.00	.00	.00	.00	.00	1.00	.00	.00	.00
8	1	116	8	0	.00	.00	.00	.00	.00	.00	.00	1.00	.00	.00
9	1	116	9	0	.00	.00	.00	.00	.00	.00	.00	.00	1.00	.00
10	1	116	10	0	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.00
11	1	116	11	0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
12	1	116	12	1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
13	1	116	13	0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
14	1	116	14	1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
15	1	116	15	0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
16	1	116	16	1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
17	1	116	17	0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
18	1	116	18	0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
19	1	116	19	1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
20	1	116	20	1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
21	1	116	21	1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
22	1	116	22	0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
23	1	116	23	1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
24	1	116	24	1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

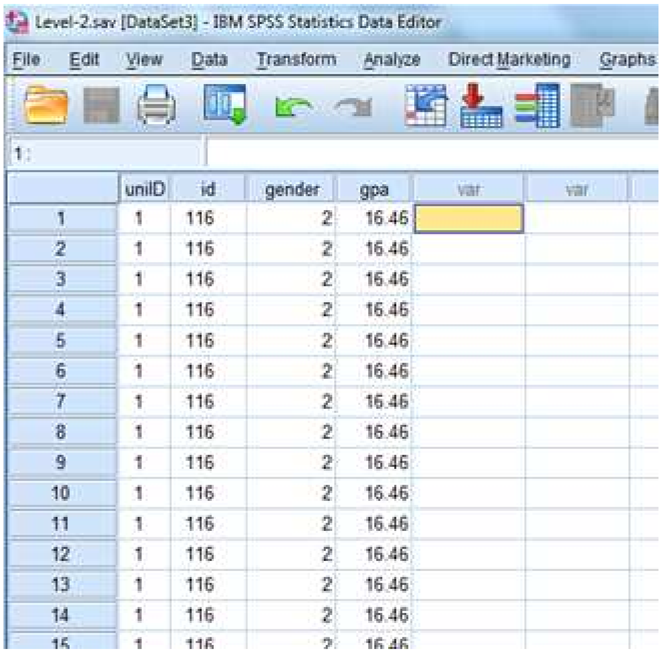
Figure 6. Item-level data file.

indicates a correct answer and 0 indicates an incorrect answer. Since item responses are arrayed in a column and there are 60 items on the test, data for each test taker consists of 60 rows. Put another way, ID for each student is repeated 60 times which results in $60 \times 20342 = 1220520$ rows of data. From the fifth variable onward, the dummy variables for each item (X1 to X60) are displayed.

Figure 7 displays the person-level data file. As the reader might note, the second-level data file has also been sorted according to the highest level ID variable (uniID). This data file includes uniID, stID, and second-level related variables such as gender and grade point average (gpa).

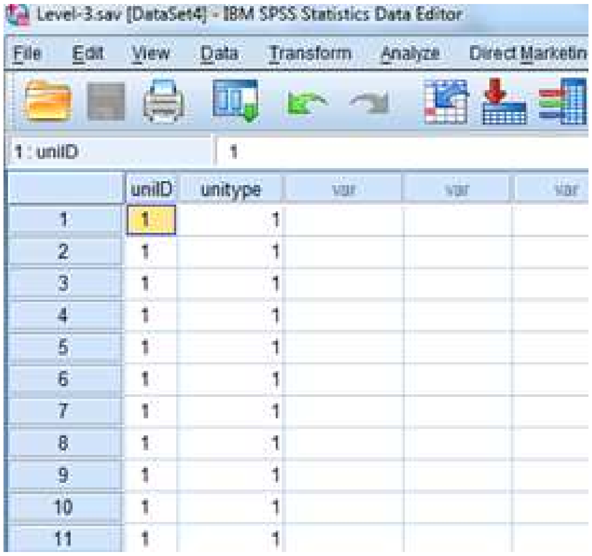
Finally, the third-level data file includes university ID (uniID) and university-related variable of university type (unitype). A slice of the third-level data (university in this case) is displayed in Figure 8.

Therefore, an important point to note about the data files is that the highest-level ID variable (here uniID) should appear in the data files for all the three levels and exactly in the same order.


Level-2.sav [DataSet3] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs
1:

	uniID	id	gender	gpa	var	var
1	1	116	2	16.46		
2	1	116	2	16.46		
3	1	116	2	16.46		
4	1	116	2	16.46		
5	1	116	2	16.46		
6	1	116	2	16.46		
7	1	116	2	16.46		
8	1	116	2	16.46		
9	1	116	2	16.46		
10	1	116	2	16.46		
11	1	116	2	16.46		
12	1	116	2	16.46		
13	1	116	2	16.46		
14	1	116	2	16.46		
15	1	116	2	16.46		

Figure 7. Person-level data file.


Level-3.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketin
1 : uniID 1

	uniID	unitype	var	var	var
1	1	1			
2	1	1			
3	1	1			
4	1	1			
5	1	1			
6	1	1			
7	1	1			
8	1	1			
9	1	1			
10	1	1			
11	1	1			

Figure 8. University-level data file.

Working with HLM

In the next step, the models for different levels should be specified. HLM stores data in its own MDM format which can be created from SPSS, Stata, SAS, and SYSTAT. To create the MDM file, after you run the HLM 7, select "Make the new MDM file" and then "Stat package input" from the file menu, as shown in Figure 9. The latter is selected since our data are stored in a statistical package (in this case SPSS).

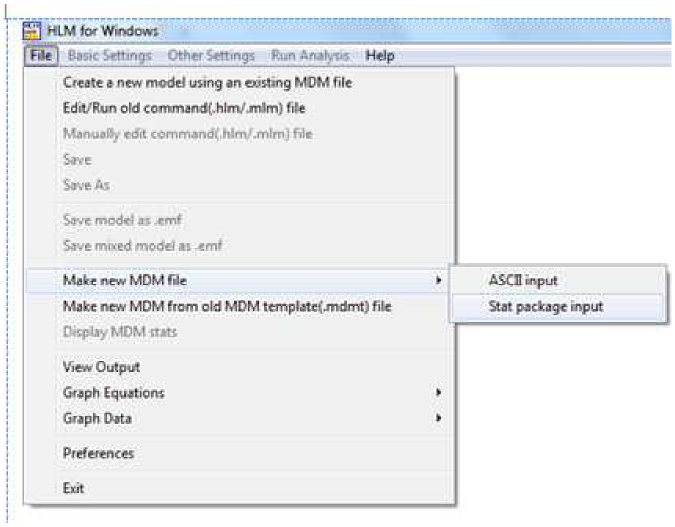

HLM for Windows
File Basic Settings Other Settings Run Analysis Help
Create a new model using an existing MDM file
Edit/Run old command(.hlm/.mlm) file
Manually edit command(.hlm/.mlm) file
Save
Save As
Save model as .emf
Save mixed model as .emf
Make new MDM file
Make new MDM from old MDM template(.mdmt) file
Display MDM stats
View Output
Graph Equations
Graph Data
Preferences
Exit
ASCII input
Stat package input

Figure 9. MDM construction 1.

In the next dialog box the type of the MLM should be selected. For simple two-, three-, and four-level MLMs the respective model should be selected from the upper section of the dialog box. For the purpose of the present study select HLM3 option because there are three levels: items (level 1) nested within students (level 2) and students in turn nested within universities (level 3). Then click OK (Figure 10).

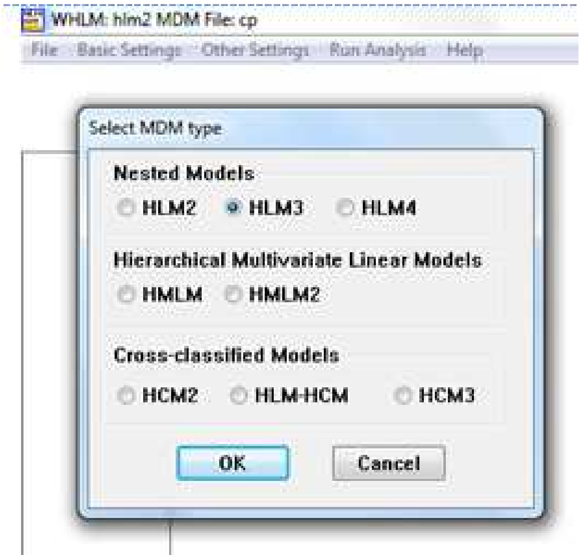


Figure 10. MDM construction 2.

In the next dialog box, click the **Browse** button for each respected level, as shown in Figure 11, and go to the directory where each of the data files have been saved. As soon as the data files are located, the **Choose variable** buttons for each respective level are activated. Click the buttons and specify the variables related to each level as shown in Figure 11. For the level-1 model in the present study the ID variables for all the three levels, the first-level-specific variables such as "response" and the dummy variables for the items should be specified. For Level 2, the ID variables for Levels 2 and 3 and second-level-related variables of gender and 'gpa' should be specified. Finally, for the third level, university ID (uniID) and university type

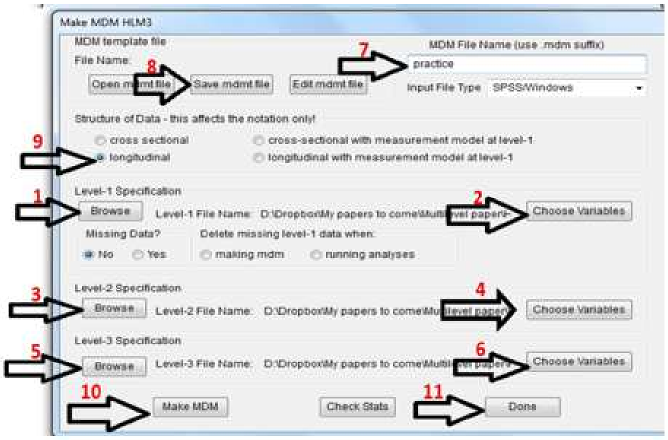

Make MDM HLM3
MDM template file
File Name:
8
Open mdmt file
Save mdmt file
Edit mdmt file
7
MDM File Name (use .mdm suffix)
practice
Input File Type SPSS/Windows
Structure of Data - this affects the notation only!
9
○ cross sectional
● longitudinal
○ cross-sectional with measurement model at level-1
○ longitudinal with measurement model at level-1
Level-1 Specification
1
Browse
Level-1 File Name: D:\Dropbox\My papers to come\Multilevel paper\
2
Choose Variables
Missing Data?
● No ○ Yes
Delete missing level-1 data when:
○ making mdm ○ running analyses
Level-2 Specification
3
Browse
Level-2 File Name: D:\Dropbox\My papers to come\Multilevel paper\
4
Choose Variables
Level-3 Specification
5
Browse
Level-3 File Name: D:\Dropbox\My papers to come\Multilevel paper\
6
Choose Variables
10
Make MDM
Check Stats
11
Done

Figure 11. MDM construction steps.

(unitype) should specified. Save the MDM template by giving it a name and clicking "save mdmt file", as specified in Figure 11. To complete the model construction process, click **Make MDM**. When the MDM is created click **Done** to exit the screen.

Specifying an intercept-only (null) model

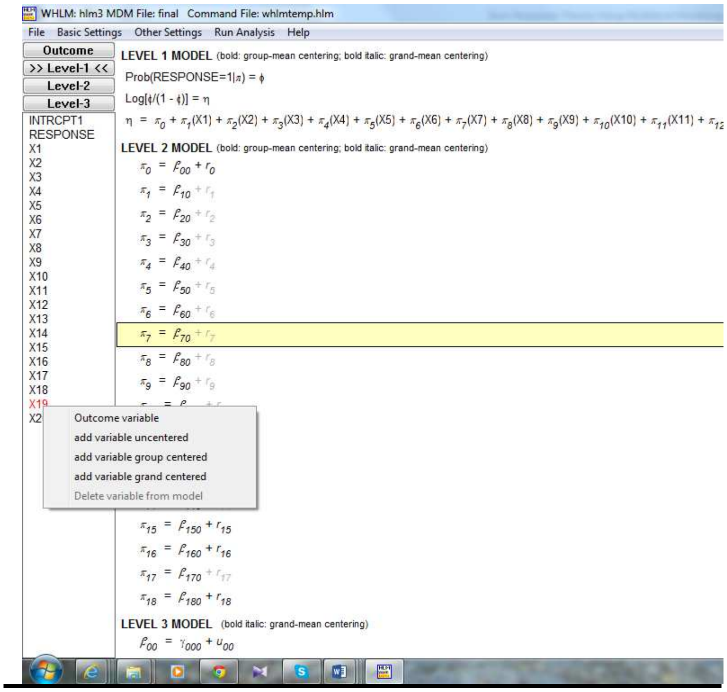
An important step before embarking upon multilevel analysis is the inspection of the amount of dependence between observations (i.e., data points). Intraclass correlation (ICC) is an index of the amount of within-cluster dependency. A high ICC indicates that the average correlations between data points (i.e., scores or item responses) from the same cluster (here university) are higher than the average correlation between scores obtained from different cluster. The more similar within-group observations to each other are and the more different they are than observations from other groups, the more inappropriate the application of the traditional statistical tests to the data will be. If the amount of variance in the person-level (Level 1 in MLMs and Level 2 in HGLMs) outcome variable attributable to the cluster level is negligible, multilevel modeling is not appropriate for the data. If ICC is +1 it indicates that there is no variation within the groups but the groups are very different from each other. If it is negative or approaches 0 multilevel analysis is not needed. To calculate ICC an intercept-only (null) model should be used.

How to specify a model

To add variables at any level, first activate the level by pressing the respective button in the left panel; all the variables related to the level will appear in the lower part of the left panel. Click on the related variable and in the drop-down menu choose "add variable uncentered".

To add a random term to any equation, click on the equation then click on the random term (r_{0jk} for Level 2 and u_{00k} for Level 3) it will be activated. Clicking once again on the same random term, will delete it from the equation. To delete any covariate already added, click on the relevant equation the variables related to the respective level appear in the left panel. In the panel click on the variable you intend to delete. In the drop-down menu click on the only active option: "Delete variable from model".

An intercept-only model is a baseline model against which more extended models can be compared. A null Rasch HGLM model is a model which includes only item dummy variables at Level-1 and no variable at other levels but intercepts. Generally, to add a variable at any level select the level you intend to add the variable to from the upper part of the left panel of the model specification dialog box. When a level is selected, the level name is embraced by double "less than" and "greater than" symbols (>> Level-1 <<). At Level 1 first, the outcome variable (in this case "RESPONSE") and then all the item dummy variables except the one for the reference item (in this case the last item) should be included. To add the outcome variable, when the Level-1 button is activated, click on "RESPONSE" in the left panel and from the drop-down menu click on "add as dependent variable". To add the dummy variables, in the left panel click on them one by one and from the drop-down menu click on "add variable uncentered", as shown in Figure 12.



LEVEL 1 MODEL (bold: group-mean centering; bold italic: grand-mean centering)

Prob(RESPONSE=1|π) = ϕ

Log[ϕ/(1 - ϕ)] = η

$\eta = \pi_0 + \pi_1(X1) + \pi_2(X2) + \pi_3(X3) + \pi_4(X4) + \pi_5(X5) + \pi_6(X6) + \pi_7(X7) + \pi_8(X8) + \pi_9(X9) + \pi_{10}(X10) + \pi_{11}(X11) + \pi_{12}$

LEVEL 2 MODEL (bold: group-mean centering; bold italic: grand-mean centering)

$\pi_0 = \beta_{00} + r_0$

$\pi_1 = \beta_{10} + r_1$

$\pi_2 = \beta_{20} + r_2$

$\pi_3 = \beta_{30} + r_3$

$\pi_4 = \beta_{40} + r_4$

$\pi_5 = \beta_{50} + r_5$

$\pi_6 = \beta_{60} + r_6$

$\pi_7 = \beta_{70} + r_7$

$\pi_8 = \beta_{80} + r_8$

$\pi_9 = \beta_{90} + r_9$

$\pi_{15} = \beta_{150} + r_{15}$

$\pi_{16} = \beta_{160} + r_{16}$

$\pi_{17} = \beta_{170} + r_{17}$

$\pi_{18} = \beta_{180} + r_{18}$

LEVEL 3 MODEL (bold italic: grand-mean centering)

$\beta_{00} = \gamma_{000} + u_{00}$

Figure 12. Model Specification.

Since we have already specified a 3-level model (see Figure 10), as soon as we include the Level 1 variables the respective Level 2 and Level 3 equations are automatically created. For the purpose of an intercept-only model leave Level-2 and 3 models intact. Next, press the "Run Analysis" button from the top menu.

In the context of the present study, ICC is the proportion of the university-level variance compared to the total variance. ICC can be computed through Equation 8.

$$ICC = \frac{\sigma_{u0}}{\sigma_{u0} + \sigma_r} \tag{8}$$

Where σ_r and σ_{u0} (in our case) are the variances of the person-level errors r_{ij} and the university-level error u_{u0}. As Tables 1 and 2 show, the variance component associated with the random term in the second level (here person level) is about 0.20 and that associated with the third level (here university level) is about 0.06. Plugging in the respective values into Equation 8, we will have

$$ICC = \frac{0.06}{0.06 + 0.20} = 0.23$$

Thus 23% of the RESPONSE variance is at university level and 77% (1-23%) is at person level. Thus addition of a cluster level to the model is warranted.

Table 1. Final estimation of level-2 variance components

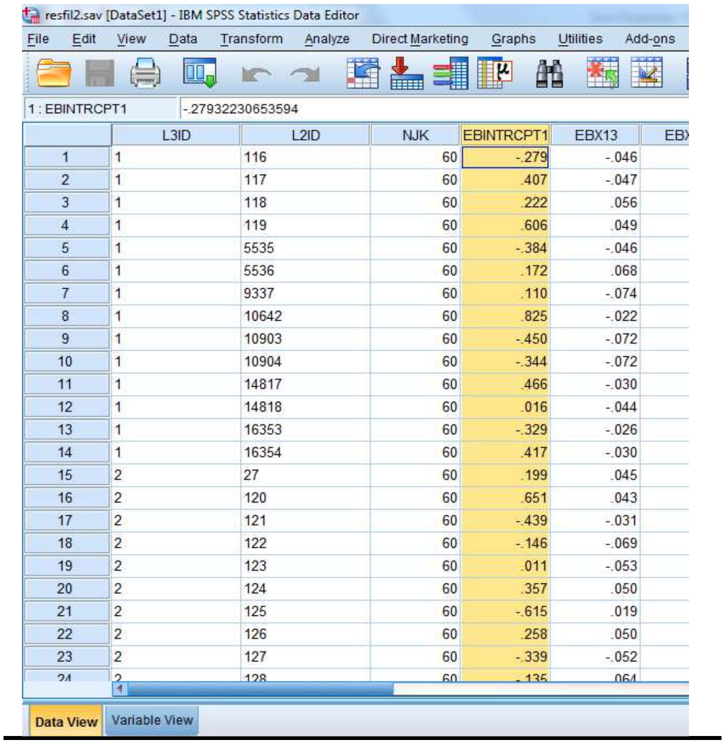
Random Effect	sd	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1,r_0	0.45081	0.20323	20116	75802.42875	<0.001

Table 2. Final estimation of level-3 variance components

Random Effect	sd	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1/ INTRCPT2,u_{00}	0.24111	0.05814	225	4660.66298	<0.001

Obtaining Item and Person parameters

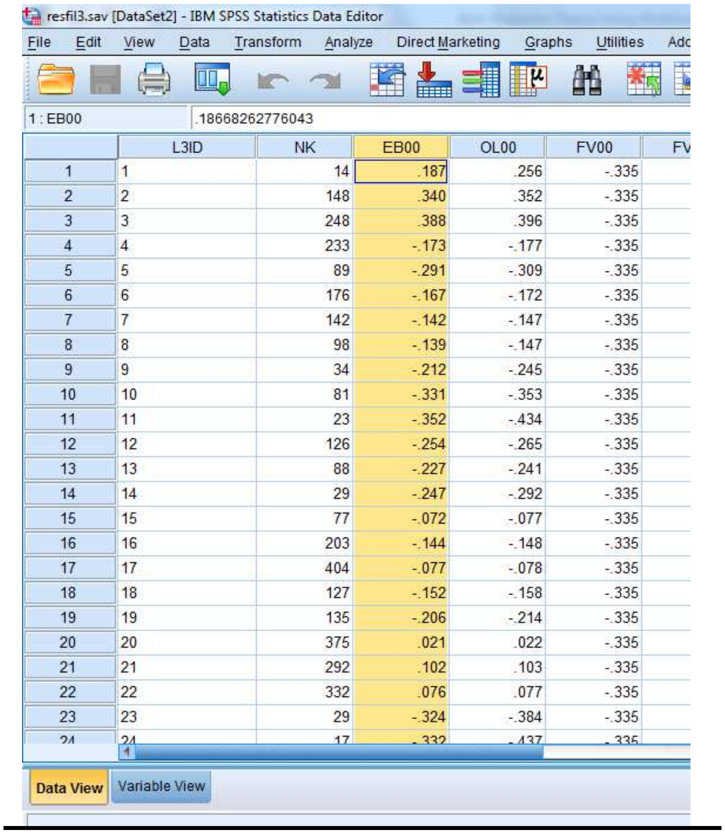
As it was mentioned above, one of the advantages of using the Rasch HGLM is that it circumvents the inconsistency associated with simultaneous estimations of person and item parameters (Kamata, 2001). In the Rasch HGLM, person parameters vary across people (are random) and fixed across items. That's why in Equation 5 (see above) there is a random term (r_0) only for the intercept not the items. To go with the conventional practice in MLM, we will start with the simplest possible model: a model with no explanatory variables or an intercept-only model. As it was said before, person parameters in Rasch HGLM are not estimated they are the residuals of the intercept component. In a two-level Rasch HGLM, person abilities are level-2 residuals, which HLM generates on demand. In a three-level model person abilities are sum of level-2 and 3 residuals. To get the residuals, go to the "Basic Settings" menu and click "Level-2 Residual File" and "Level-3 Residual File" and in the new dialog boxes specify the format you want HLM to produce the file in and give the files a name and then click "OK". One more thing needs to be specified in this dialog box: the distribution of the dependent variable(s). Since we are working with binary variables with one trial (i.e., each test taker tries each item once) the "Bernoulli" distribution should be specified from the "Distribution of the Outcome Variable" section. Then click on the "Run the Analysis" button to run the specified model. In the interest of space, only a slice of the level-2 and 3 residual files are displayed in Figures 13 and 14.



1 : EBINTRCPT1 -.27932230653594

	L3ID	L2ID	NJK	EBINTRCPT1	EBX13	EBX
1	1	116	60	-.279	-.046	
2	1	117	60	.407	-.047	
3	1	118	60	.222	.056	
4	1	119	60	.606	.049	
5	1	5535	60	-.384	-.046	
6	1	5536	60	.172	.068	
7	1	9337	60	.110	-.074	
8	1	10642	60	.825	-.022	
9	1	10903	60	-.450	-.072	
10	1	10904	60	-.344	-.072	
11	1	14817	60	.466	-.030	
12	1	14818	60	.016	-.044	
13	1	16353	60	-.329	-.026	
14	1	16354	60	.417	-.030	
15	2	27	60	.199	.045	
16	2	120	60	.651	.043	
17	2	121	60	-.439	-.031	
18	2	122	60	-.146	-.069	
19	2	123	60	.011	-.053	
20	2	124	60	.357	.050	
21	2	125	60	-.615	.019	
22	2	126	60	.258	.050	
23	2	127	60	-.339	-.052	
24	2	128	60	-.135	.064	

Figure 13. Level-2 Residuals.



1 : EB00 .18668262776043

	L3ID	NK	EB00	OL00	FV00	FV
1	1	14	.187	.256	-.335	
2	2	148	.340	.352	-.335	
3	3	248	.388	.396	-.335	
4	4	233	-.173	-.177	-.335	
5	5	89	-.291	-.309	-.335	
6	6	176	-.167	-.172	-.335	
7	7	142	-.142	-.147	-.335	
8	8	98	-.139	-.147	-.335	
9	9	34	-.212	-.245	-.335	
10	10	81	-.331	-.353	-.335	
11	11	23	-.352	-.434	-.335	
12	12	126	-.254	-.265	-.335	
13	13	88	-.227	-.241	-.335	
14	14	29	-.247	-.292	-.335	
15	15	77	-.072	-.077	-.335	
16	16	203	-.144	-.148	-.335	
17	17	404	-.077	-.078	-.335	
18	18	127	-.152	-.158	-.335	
19	19	135	-.206	-.214	-.335	
20	20	375	.021	.022	-.335	
21	21	292	.102	.103	-.335	
22	22	332	.076	.077	-.335	
23	23	29	-.324	-.384	-.335	
24	24	17	-.332	-.437	-.335	

Figure 14. Level-3 Residuals.

If we were working with a two-level model person abilities could be directly read from the "olintrcp" or the "ebintrcp" columns in Figure 13. However, in a three-level model, as shown in Equations 13 and 14 below, there are two random terms: one for the level-2 intercept (r_{0jk} in Equation 13 below) and one for level-3 intercept (u_{00k} in Equation 14 Below). The random term for the level-2 intercept (r_{0jk}) indicates the degree to which person *j* in university *m* is deviated from the mean of the university *m* whereas the random term for Level 3 (u_{00k}) indicates how much the mean ability in university *m* deviates from the grand mean (i.e., the mean of all universities' means). In a three-level model, ability of persons can be obtained by aggregating level-2 and level-3 random terms (i.e., $r_{00m} + u_{0jm}$).

Level-3 residuals can be read from the "olintrcp" or the "ebintrcp" columns in Figure 14. One can manually compute person parameters by summing level-2 and level-3 residuals. In Figure 13 the first column (i.e., L3D column) represents the university each person belongs to and the second column (i.e., L2D column) represents the person IDs. As one can read from the figure, the first 14 students belonged to university No. 1. As Figure 13 shows, ability of Person No. 116 which belongs to University No. 1 is -.279 logits and according to Figure 14 the mean university ability for the university he/she attended (university No. 1) is .187. Therefore his/her ability can be computed by summing his/her person-level ability and the mean university ability as follows:-.279+.187=0.092

In a three-level model, item difficulties are computed based on the item effects in the third level. In the interest of space the effects for Items 1 to 3 are presented in Table 3.

Table 3. Item effects

Fixed Effect	Coefficient	Stand. error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	-0.334462	0.017469	-19.146	225	<0.001
For X1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	0.827132	0.014959	55.293	1199933	<0.001
For X2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	-0.814584	0.016640	-48.953	1199933	<0.001
For X3 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	-0.343377	0.015278	-22.476	1199933	<0.001

γ_{000} is the difficulty of the reference item (-0.334462) and according to the Rasch HGLM the difficulties of other items are computed by subtracting each item's effect from the reference item effect $(-\pi_{i00} - \pi_{000})$. The difficulty for Item 1, for example, is computed as follows: 0.827132 -0.334462=0.49267

Adding level-2 predictors

Impact

Beretvas, Cawthon, Lockhart, & Kay (2012) define impact as "difference in person abilities as a funciton of some person-level predictors"(p. 6). Take a simple case where a researcher intends to study the effect of test takers' gender on their perfomance in a reading comprehension test. As was explained formerly, the usual approach in conventional IRT models is to estimate person and item parameters first and then in the second step estimate the effect of explanatory variables such as gender on test perofmance. However, the two-step approach may not yeild accurate results. Approached form a Rasch HGLM one-step perspective, the level-1 mdoel remains the same, as in Equation 4, and the person-level predictor (in this case gender) is added to the level-2 model as in Equation 9.

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}(\text{gender})_j + r_{0j} \\ \beta_{1j} &= \gamma_{10} \\ &. \\ &. \\ &. \\ \beta_{(k-1)j} &= \gamma_{(k-1)0}, \end{aligned} \tag{9}$$

where Coefficient γ_{01} represents *impact*. Since β_{0j} is a parameter that is common to all items in the level-1 model and the intercept value affects every item's difficulty, statistically significant coefficient of γ_{01} would indicate that overall, males and females performed significantly differently on *all* of the items. Therefore coefficient $\gamma_{01}$is the difference in ability of a male versus a female test taker.

Generally, to add any covariate to level-2 and 3 equations, click on the relevant equation; all the variables related to the respective level appear in the left panel. Click on the relevant variable and from the drop-down menu select "add variable uncentered". To add gender impact, for example, to the model, in the **"model specification"** menu click on the first line of the level-2 equation (the intercept equation π_0) to activate the relevant variables in the left panel and then add the variable as explained above. A slice of the output is displayed in Table 4.

Table 4. Final estimation of fixed effects: (Population-average model)

Fixed Effect	Coefficient	Stand. error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	-1.854943	0.039772	-46.639	225	<0.001
For GENDER, β_{01}					
INTRCPT3, γ_{010}	-0.224154	0.008099	-27.678	20114	<0.001

According to Figure 18, the intercept for gender (γ_{01}=-0.22) is significant (*p*-value <0.001). The negative value implies that on average males had higher ability estimates than females by 0.22 logits (because the code assigned to males [i.e., 0] was lower than that of females [i.e., 1]).

How much variance was explained?

In traditional multiple regression R^2 is a gauge of the amount of variance explained by the predictor variables. In MLM the amount of variance explained should be examined for each level separately by calculating a statistic analogous to R^2. A straightforward approach is to compare the variance of the intercept in each level after addition of the explanatory variables with the variance component of the baseline model (i.e., intercept-only model). Raudenbush and Bryk (2002) suggested using Equation 10 to calculate R^2 for the person-level model:

$$R^2 = \left(\frac{\sigma^2_{e|b} - \sigma^2_{e|m}}{\sigma^2_{e|b}}\right) \tag{10}$$

where $\sigma^2_{e|b}$ is the person-level residual variance for the intercept-only model and $\sigma^2_{e|m}$is the person-level residual variance for the model with explanatory variable. As Table 5 shows, inclusion of gender into the person-level model reduced the variance component to about 0.17.

Table 5. Final estimation of level-1 and level-2 variance components

Random Effect	sd	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1,r_0	0.41352	0.17100	20114	65943.84306	<0.001

As Table 1 above showed the variance of the intercept in the baseline model was 0.20. Plugging in the respective values into Equation 10 we will have

$$R^2 = \frac{0.20 - 0.17}{0.20} = 0.15$$

The implication is that gender explain about 15% of the explainable variance at person level. The significance variance component in Table 5 suggests that more person characteristic variables should be added.

Estimating DIF

DIF occurs when test takers with the same ability level but from different observed groups have different probabilities of giving the correct answer to an item (Clauser and Mazor, 1998). In other words, DIF refers to significant difference in item difficulties across different groups in the same population, which are matched for ability. Differences in item difficulties and discriminations across subpopulations with equal latent trait abillity are refered to as *uniform* and *non-uniform* DIF, respectively. Rasch HGLM tests only for uniform DIF. To investigate DIF, a person covariate (here gender) can be added as to the level-2 equation as follows:

$$\begin{aligned}\beta_{0j} &= \gamma_{00}+\gamma_{01}(\text{gender})_j+u_{0j}\\ \beta_{1j} &= \gamma_{10}+\gamma_{11}(gender)_j\\ &\cdot\\ &\cdot\\ &\cdot\\ \beta_{(k-1)j} &= \gamma_{(k-1)0}+\gamma_{(k-1)1}(gender)_j,\end{aligned} \tag{11}$$

In Equation 11, γ_{01} represents impact and γ_{11} to $\gamma_{(k-1)1}$ represent DIF. If the coefficient γ_{q1} is positive, the item is, after controlling for ability, easier for females (since females were code as 1) and if the coefficient is negative, after controlling for ability, the item is easier for males (since 0 represented males here). To add gender DIF to the equation for any of the items, first select the respective equation and, as explained above, add gender. To save space, the results for Items 15, 16, and 18 are presented in Table 6.

Table 6. 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 X15 slope, π_{15}					
For INTRCPT2, β_{150}					
INTRCPT3, γ_{1500}	1.002693	0.088475	11.333	20107	<0.001
For GENDER, β_{151}					
INTRCPT3, γ_{1510}	-0.673751	0.043299	-15.560	20107	<0.001
For X16 slope, π_{16}					
For INTRCPT2, β_{160}					
INTRCPT3, γ_{1600}	0.416603	0.073470	5.670	20107	<0.001
For GENDER, β_{161}					
NTRCPT3, γ_{1610}	0.175237	0.038120	4.597	20107	<0.001
For X18 slope, π_{18}					
For INTRCPT2, β_{180}					
INTRCPT3, γ_{1800}	-0.346167	0.097776	-3.540	20107	<0.001
For GENDER, β_{181}					
INTRCPT3, γ_{1810}	-0.481891	0.049642	-9.707	20107	<0.001

According to Table 6, the effect of gender on the three items was statistically significant (*p*-value <0.001). Items 15 and 18 were easier for males as indicated by the negative signes(males were coded as 0) and Item 16 was easier for females as indicated by the positive sign(females were coded 1).

Rasch HGLM is flexible in that it can also test for *unobservable* (i.e., not-yet-measured) person characteristics as DIF source. Equation 11 above tested for gender DIF as an *obsevable* DIF source. Testing for unobservable sources of DIF can be carried out by adding a person-specific residual contributing to the overall difficulty of any given item through Equation 12. That is item difficulits should be modeled as random rather than fixed.

$$\begin{aligned}\beta_{0j} &= \gamma_{00}+\gamma_{01}(\text{gender})_j+r_{0j}\\ \beta_{1j} &= \gamma_{10}+\gamma_{11}(gender)_j+r_{1j}\\ &\cdot\\ &\cdot\\ &\cdot\\ \beta_{(k-1)j} &= \gamma_{(k-1)0}+\gamma_{(k-1)1}(gender)_j\end{aligned} \tag{12}$$

Signnificance of the random effect r_{1j} indicates that "the item's difficulty is affected by an , as-yet-*unmeasured* person charactirsic and cannot be assumed as fixed across people"(Beretvas, Cawthon, Lockhar, & Kaye,2012,p.760).

In what follows Items 15,16, and 18 are tested for unobsevable source of DIF. To do so, in the "model specification" menue select the equation for the relevant item by clicking once on the equation and then click on the random term (i.e., r_{qj}) for the respective equation and it will get activated (i.e., gets black). Running the model by clicking the "Run analysis" button you will get the following output (Table 7) regarding the random terms of Items 15,16, and 18. As one can read from the table, the variance component for Item 16 is still significant (*p*-value<0.001) which implies that other person covariates than gender should be added to the model for Item 16 to capture variations in test takers' performance on the item.

Table 7. Final estimation of level-1 and level-2 variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ²	*p*-value
X13 slope,r_{13}	0.32464	0.10539	20340	20054.83849	>.500
X15 slope,r_{15}	0.54028	0.29190	20340	19031.63363	>.500
X16 slope,r_{16}	0.52583	0.27649	20340	21145.05560	<0.001
X18 slope,r_{18}	0.55096	0.30356	20340	18286.01242	>.500

Adding Level-3 Predictors

Flexibility of the Rasch HGLM permits researchers to add a third level to capture the clustering of examinees nested within classes or schools. In a three-level model, at the second level, item difficulties are modeled as fixed across test takers as shown in Equaiton 13.

$$\begin{aligned} \beta_{0jk} &= \gamma_{00k} + r_{0jk} \\ \beta_{1jk} &= \gamma_{10k} \\ &. \\ &. \\ &. \\ \beta_{mjk} &= \gamma_{m0k} \end{aligned} \qquad (13)$$

where r_{0jk} represents the extent to which the ability of person *j* in school *k* deviates from the mean ability of school *k*. And at the third level, item difficulties can also be assumed fixed across schools:

$$\begin{aligned} \gamma_{00k} &= \pi_{000} + u_{00k} \\ \gamma_{10k} &= \pi_{100} \\ &. \\ &. \\ &. \\ \gamma_{(k-1)0m} &= \pi_{(k-1)00} \end{aligned} \qquad (14)$$

where u_{00k} is the residual for school *k*. In Equaiton 14, γ_{100} to γ_{m00}, represent the difficulty for item *q* but the ability is now decomposed into a person-specific ability , r_{0jk}, and the school-specific ability which is the average ability of students in school *k*. Thus the ability, θ_j in the standard Rasch model in Equation 7 above, corresponds to $r_{0jk} + u_{00k}$. The three-level model can also be extended by adding school-level predictors. The level-3 predictor in this study is university type (i.e., "unitype"). As it was explained before, subjects in the present study for their B.A. studied at four university types: state unversity, Azad unvrisity, Payam-e-Noor university, and nonprofit not-government university. The model at the third level is specified as in Equation 15.

$$\begin{aligned} \gamma_{00k} &= \pi_{000} + \pi_{001}(unitype)_k + u_{00k} \\ \gamma_{10k} &= \pi_{100} \\ &. \\ &. \\ &. \\ \gamma_{(k-1)0m} &= \pi_{(k-1)00} \end{aligned} \qquad (15)$$

To add "UNITYPE", which is the only predictor variable at Level 3, select the intercept equation (γ_{000}) and as explained above, in the left panel click on 'unitype' variable. Next, click on "add variable uncentered" and run the analysis. Part of the output for the three-level model is displayed in Table 8. According to this table, the type of university students attended had a significant effect on their performance (*P*<0.001).

Table 8. Final estimation of fixed effects: (Unit-specific model)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	0.142087	0.042568	3.338	224	<0.001
UNITYPE, γ_{001}	-0.102606	0.015305	-6.704	224	<0.001

The variance component for the intercept at the third level is still significant at *p*<0.001, as shown in Table 8. The implication is that more university-level covariates can be added to capture variations in the performance of the universities. According to Raudenbush and Bryk (2002), the amount of variance explained by the third-level predictor can be calculated using Equation 16 as follows:

$$R^2 = \left(\frac{\sigma^2_{\mathrm{u}_0|b} - \sigma^2_{\mathrm{u}_0|m}}{\sigma^2_{\mathrm{u}_0|b}}\right), \qquad (16)$$

where $\sigma^2_{\mathrm{u}_0|b}$ is the university-level residual variance for the intercept-only model and $\sigma^2_{\mathrm{u}_0|m}$ is the university-level residual variance for the model with "unitype" as the predictor.

Table 9. Final estimation of level-3 variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1/ INTRCPT2,u_{00}	0.21733	0.04723	224	4280.86603	<0.001

Plugging in the respective values from Tables 2 above and 9, we will have

$$R^2 = \frac{0.058 - 0.047}{0.058} = 0.19$$

The implication is that "unitype" explains about 19% of the explainable variance at the university level.

Summary

In the present paper I tried to introduce MLMs in general and HGLMs in particular in an easy-to-follow language and illustrate their application to language testing data. First the 'wide" data format was converted into "long" format to make it compatible with HGLM. Then I showed how person and item parameters can be estimated and items showing DIF can be detected. Finally, I explained how to add a third level and the related covariates. It was illustrated how to calculate the amount of variance explained after addition of the covariates at Levels 2 and 3.

References

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logic. In G. Engelhard & M. Wilson (Eds.),*Objective measurement: Theory and practice*(Vol. 3, pp. 143-166). Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47–76.

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, *50*(4), 408-426.

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L., & Kaye, A. D. (2012). Assessing Impact, DIF, and DFF in Accommodated Item Scores A Comparison of Multilevel Measurement Model Parameterizations. *Educational and Psychological Measurement*, *72*(5), 754-773.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, *72*(2), 200-223.

Clauser, E. B. & Mazor, M. K. (1998) Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice.* 17, 31-44.

Debeer, D., & Janssen, R. (2013). Modeling Item-Position Effects Within an IRT Framework. *Journal of Educational Measurement*, *50*(2), 164-185.

De Boeck, P., & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson(Eds.),*Explanatory item response models*(pp. 3-42). New York, NY: Springer.

Goldstein, H. (1995) *Multilevel statistical models.* Edward Arnold, London

Hox, J. J. (2010). Multilevel analysis. Techniques and applications. NY: Routledge.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*(1), 82-100.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of applied measurement*, *6*(3), 311.

Kamata, A. (1998). *Some generalizations of the Rasch model: An application of the hierarchical generalized linear model.* Doctoral dissertation, Michigan State University, East Lansing.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

Kamata, A. (2002). Procedure to perform item response analysis by hierarchical generalized linear model. In *annual meeting of the American Educational Research Association, New Orleans, LA.*

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). SAS system for mixed models. Cary, North Carolina: SAS Institute.

Longford, N. T. (1993) *Random Coefficient Models.*Clarendon Press, Oxford.

McClellan, C., & Donoghue, J. R. (2001, April).*Applying the hierarchical rater model to NAEP*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.),*Explanatory item response models A framework for item response models.* (pp. 213-240). New York, NY: Springer.

Mislevy, R. J. (1987). Exploiting information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*, 81-91

Miyazaki, Y. (2005). Some links between classical and modern test theory via the two-level hierarchical generalized linear model. *Journal of Applied Measurement, 6*(3), 289–310.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 1-32.

O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. J. (2008). Multilevel logistic models for dichotomous and ordinal data. *Multilevel modeling of educational data*, 199-242.

Randall, J., Cheong, Y. F., & Engelhard, G. (2011). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement*, *71*(1), 129-147.

Raudenbush, S. W., & A. S. Bryk. (1986). A hierarchical model for studying school effects. *Sociology of Education,* ***59,*** 1-17.

Raudenbush, S. W., & Bryk, A. S. (2002).Hierarchical linear models: Applications and data analysis methods (2nd ed.). Newbury Park, CA:Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & du Toit, M. (2013). HLM 7 for Windows [Computer software]. Skokie, IL: Scientific Software International.

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A multivariate, multilevel Rasch model with application to self–reported criminal behavior.*Sociological methodology*, *33*(1), 169-211.

Ravand, H. (in press). Assessing testlet effect, impact, differential testlet and item functioning using cross-classified multilevel measurement modeling. *SAGE Open*.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*(4), 369-386.

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling Item Position Effects Using Generalized Linear Mixed Models. *Applied Psychological Measurement*, *38*(7), 535-548.

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement, 30*, 22-42.

Citation:

Ravand, Hamdollah (2015). Item Response Theory Using Hierarchical Generalized Linear Models. *Practical Assessment, Research & Evaluation*, 20(7). Available online: http://pareonline.net/getvn.asp?v=20&n=7

Author:

Hamdollah Ravand
Department of English Language Teaching & Literature,
Faculty of Literature & Humanities,
Vali-e-Asr University of Rafsanjan
P.O. Box: 7713936417
Rafsanjan, Iran

Email: ravand [at]vru.ac.ir