# THE APPLICATION OF ORDINAL LOGISTIC HEIRARCHICAL LINEAR MODELING IN ITEM RESPONSE THEORY FOR THE PURPOSES OF DIFFERENTIAL ITEM FUNCTIONING DETECTION

Timothy Olsen

HLM II – Dr. Gagne

## ABSTRACT

Recent advances in multilevel modeling software have made it easier to investigate potential bias in test items. By modeling responses to questions as level-1 model and student manifest variables and latent constructs as level-2 models, I explain how ordinal logistic HLM techniques can be applied to identify questions with such bias.

**Keywords:** ordinal logistic HLM, item response theory, DIF analysis, research methods, hierarchical linear modeling

# I. INTRODUCTION

## ITEM RESPONSE THEORY (IRT)

Tests taken in testing centers such as the GRE and GMAT are based on the idea that the probability of getting an item correct is a function of a latent trait or ability. In theory, a student possessing higher intelligence would be more likely to respond correctly to a given item on an intelligence test. In practice, the same student for various reasons may not respond correctly, although he does possess higher intelligence. IRT analysis can be helpful in identifying such problem questions, and aiding in the development of tests.

Item response theory seeks to apply mathematical models to specify the probability of a certain response. These mathematical models are functions of both person and test item characteristics. Person characteristics may represent latent characteristics such as motivation or attitude or more easily measured characteristics such as ethnicity or gender. The purpose of IRT analysis is to provide a metric for evaluating how well assessments work and how well individual items perform.

## BIAS AND DIFFERENTIAL ITEM FUNCTIONING (DIF)

An item is considered to be “biased” if it unfairly favors one group over another. Specifically, an item is considered biased if two conditions are met. First, performance on the item must be influenced by factors other than differences on the construct of interest (intelligence, in an intelligence test). Second, the influence of non-group factors

must result in different performance across subgroups of examinees (Jensen, 1980; Vaughn, 2006).

One characteristic of bias is differential item functioning (DIF). An item shows DIF when people from different groups (i.e. gender or ethnicity) of the ***same true ability*** have a different probability to give a certain response on a test. (Scheuneman, 1979). On an item showing no DIF, subjects having the same ability would have the same probability of getting the item correct.

DIF is a necessary but insufficient condition for item bias (Williams, 1997). In order to show bias, one must show that groups differ by first controlling for their ability levels. If an item is biased, DIF is present. However, the presence of DIF does not imply bias, only a *potentially* biased item (Kamata & Vaughn, 2004).

DIF analysis typically compares a focal and a reference group. Uniform DIF refers to the scenario where the magnitude of group difference is the same across ability levels (in other words, there is no interaction between the groups). An example of uniform DIF would be a test where women uniformly outperform men by two points at all ability levels. Non-uniform DIF exists when the magnitude of group difference is not consistent across ability levels. An example of non-uniform DIF would be a test where high ability women outperform high ability men, but low ability women underperform low ability men. This is the equivalent of an interaction (Vaughn, 2006).

## HEIRARCHICAL LINEAR MODELING (HLM)

Hierarchical Linear Models, also known as multi-level models, are a statistical tool used for analyzing the variance of outcome variables in a nested structure. Examples of such

nested data include, students nested within schools, or students' responses to test items nested within students. Generally, within the domain of Item Response Theory, responses to items on tests are treated as a first level model nested within a student, which is modeled as a second level.

## II. APPLICATION AND EXAMPLE

### DESCRIPTION OF DATA

A hypothetical ninth-grade class consisting of 50 students (even number of Asian and Hispanic) was asked to take a survey which would measure their level of awareness of current world affairs which have commonly been in the news media. They were asked to rate the prominence of 15 different news items using a 3-point likert scale. For example, the first question on the instrument reads "Please rate your agreement with the following statement: Twitter has recently been discussed in the news".

The data are collected with the following labels and characteristics:

RESPONSE = (ordinal likert values of 0, 1, 2 for disagree – neutral - agree)

ETHNICITY = (0 = Asian, 1 = Hispanic)

ITEM = (1-15, for the 15 test questions)

From these data we can build Level-1 (item) and Level-2 (student) models. The Independent or Outcome variable is the ordinal response variable consisting of 1 of 3 Likert values.

**LEVEL-1 MODEL**

A cumulative probability model is used to consider DIF for ordinal items. For each ordered response, the probability of making each unique response is established. These probabilities can be seen in the formulas created by the HLM software (Raudenbush, Bryk, & Congdon, 2004), and are listed below.

P(0) = Prob[RESPONSE(0) = 1|B] = P("disagree")

P(1) = Prob[RESPONSE (1) = 1|B] = P("neutral")

P(2) = Prob[RESPONSE (2) = 1|B] = P("agree")

Thus, each different probability is the probability of responding with each successive ordinal response. Unlike logistic models in which there are binomial (0 or 1) outcomes and one probability function, ordinal logistic models have more than two outcomes and three probability functions (as you can see above). Because of this, a cumulative probability model is incorporated. For a 3-point likert model, these would be represented by:

Prob[R = 0|B] = P'(0) = P(0)

Prob[R <= 1|B] = P'(1) = P(0) + P(1)

Prob[R <= 2|B] = 1.0

Thus, the first cumulative probability would represent the likelihood of a "disagree" response. The second cumulative probability would represent the likelihood of a "neutral" or "disagree" response. The last cumulative probability is not usually considered because it is 1, which represents the probability of a response in any category.

Cumulative probabilities give rise to use of thresholds. A threshold represents the unique intercept for the probability functions of each category. Thus, the threshold for the first equation represents the unique intercept for the "disagree" response, and the threshold for the second equation represents the unique intercept for the "disagree" ***or*** "neutral" response. Thus, a common intercept can be introduced into the model by considering the difference between the thresholds. The *threshold difference* is defined as the difference in thresholds for the first two logits. This value is reflected in the coefficient $d_{(2)}$ below. Because the probabilities are cumulative, the threshold difference values for entire sample will always be positive.

Using the above probability equations, a Level-1 (item) ordinal logistic HLM model can be expressed as:

$$\log[P'(0)/(1 - P'(0))] = B_{0j} + B_{qj}*(ITEM)_{qij}$$

$$\log[P'(1)/(1 - P'(1))] = B_{0j} + B_{qj}*(ITEM)_{qij} + d_{(2)}$$

In this model, $B_{0j}$ represents the expected effect of the reference item for subject *j*. $B_{qj}$, represents the expected effect of the reference item for subject *j*, for a particular item *i*, when q=i, and a value of 0 otherwise. The $d_{(2)}$ represents the difference between the thresholds. Adding the threshold difference to the prediction equation will yield the cumulative probability for the first two ordinal responses.

For the sake of simplicity, we create an expression to see the effects for one item q:

$$\log[P'(1)/(1 - P'(1))] = B_{0j} + B_{qj} + d_{(2)}$$

**LEVEL-2 MODEL**

$$b_{0j} = \gamma_{00} + \gamma_{01}\cdot ETHNICITY_j + u_{0j}$$

$$b_{1j} = \gamma_{10} + \gamma_{11} \cdot ETHNICITY_j + u_{0j}$$

$$d_{(2)}$$

In the level-2 model, the group indicator variable (ETHNICITY) is used to indicate the ethnic affiliation for the subject *j*. (0 = reference group, 1 = focal group). In considering the effect of ethnicity, the items are portioned into two effects: the difficulty of an item *q* for the reference group ($\gamma_{00}$ and $\gamma_{10}$ ), and contrast between focal and reference groups for a particular test question (item) which represents the amount of DIF between the first and second ordinal response for item *q*. A positive DIF value would indicate a potential bias against the focal group (Vaughn, 2006).

In ordinal response questions, it is possible that DIF may be present between certain categorical responses within a particular item, and not in other responses. A negative DIF value for $d_{(2)}$ for a specific item q would indicate a decrease in DIF effect (as compared to the first ordinal response) and less potential bias against the focal group responding to the last ordinal response on item *q*. A positive value would indicate an increase in DIF effect and a potential bias for the focal group at the 2nd ordered response at item *q*.

**COMBINED MODELS**

When the models are combined, the specific predictor model for a certain item *q* becomes:

$$\text{Log Odds} = \gamma_{00} + \gamma_{q0} + \gamma_{01} \cdot ETHNICITY_j + \gamma_{q1} \cdot ETHNICITY_j + u_{0j} + d(1)$$

This specific item effect,

$$\gamma_{00} + \gamma_{q0} + d_{(1)}$$

Corresponds to the item difficulties for the reference group, and refers to the difficulty associated with responding to the first two ordinal categories on item *q*. For the focal group, the probability of responding to the first two categories on item *q* is denoted by:

$$\gamma_{00} + \gamma_{q0} + \gamma_{01} + \gamma_{q1} + d_{(2)}$$

$\gamma_{01}$ represents the contrast across all items between the focal (Asian) and reference (Hispanic) group, and an item-specific contrast between the focal and reference group ($\gamma_{q1}$). The magnitude of $\gamma_{q1}$ indicates the presence and effect of an overall DIF between the first and second ordinal response for a particular item *q*.

As we have an ordinal response, there is a possibility that DIF may be present between certain categorical responses within a particular item. The magnitude of the threshold value indicates the change in DIF effect on particular consecutive ordinal response items (Vaughn, 2006).

## III. METHODOLOGICAL ISSUES

The use of manifest groups, as this paper does (Asian and Hispanic students), does not accurately reflect the actual causes of DIF (Allan S. Cohen & Bolt, 2005). Samuelson (Samuelsen, 2005) notes problems with detecting DIF by simply using a manifest variable. First, manifest grouping variables do not represent homogenous populations. The Hispanic population in the United States is diverse in origin and ethnicity, as such, this classification does not yield a group which is homogenous on a dimension related to DIF (Schmitt, Holland, & Dorans, 1993). Second, manifest groups used for DIF

comparisons such as gender and ethnicity are not related to the issues of learning about which educators care, but are really proxies for something else. Thus the potential lack of overlap that may exist between the manifest groups and latent classes may obscure the true magnitude of DIF (A. S. Cohen, Gregg, & Deng, 2005).

As this explanatory paper applies DIF detection methods based on manifest group variables it must assume homogeneity among group members which may not be justified. This paper does not model these interesting latent constructs and their effects on DIF, and thus recognizes this shortcoming. We refer the reader to Cho (Cho, 2007) for more information on multi-level IRT modeling with latent constructs.

## REFERENCES

Cho, S. (2007). A Multilevel Mixture IRT Model for DIF Analysis.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research and Practice*, *20*(4), 225.

Cohen, A. S., & Bolt, D. M. (2005). A Mixture Model Analysis of Differential Item Functioning. *Journal of Educational Measurement*, *42*(2), 133-148. doi: 10.1111/j.1745-3984.2005.00007.

Jensen, A. R. (1980). *Bias in mental testing*. 1980.

Kamata, A., & Vaughn, B. K. (2004). An Introduction to Differential Item Functioning Analysis. *Learning Disabilities: A Contemporary Journal*, *2*(2), 21.

Raudenbush, S., Bryk, A., & Congdon, R. (2004). *HLM for Windows*. Version 6.06.

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective*. University of Maryland.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, 143-152.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. *Differential item functioning*, 281-315.

Vaughn, B. K. (2006). *A Hierarchical Generalized Linear Model of Random Differential Item Functioning for Polytomous Items: A Bayesian Multilevel Approach*.

Williams, V. S. L. (1997). The" Unbiased" Anchor: Bridging the Gap between DIF and Item Bias. *Applied Measurement in Education*, *10*(3), 253-67.