Multilevel Modeling of Item Parameter Drift

Anthony D. Albano & Michael C. Rodriguez
University of Minnesota

Minnesota Youth Development Research Group

April 2012

Paper presented at the annual meeting of the
National Council on Measurement in Education, Vancouver, BC.

Citation:

Albano, A.D., & Rodriguez, M.C. (2012, April). *Multilevel modeling of item parameter drift*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.

Multilevel Modeling of Item Parameter Drift

Item response theory (IRT) models are based, in part, on the assumption that the model parameters are invariant over examinee group. This assumption may be violated when examinees of the same ability but of different group membership (e.g., ethnicity, gender) differ in performance across one or more items. Such differential item functioning (DIF) across cohorts, groups of people categorized by time of administration, is referred to as item parameter drift (IPD; Goldstein, 1983). IPD reflects variability in item parameters over time, a variability which can lead to bias in item and person parameter estimates and instability in a measurement scale (Babcock & Albano, 2011).

IPD may occur for a variety of reasons. For example, difficulty estimates may vary over time as item content becomes more or less relevant to the construct measured (Bock, Muraki, & Pfeiffenberger, 1988). Item difficulty may also change with increased item exposure, where an item becomes easier as it is administered more frequently. As a result, IPD can be especially problematic for testing programs which rely on IRT anchor-item equating to create a single measurement scale that spans multiple test forms and years. A variety of methods for detecting and assessing IPD in such situations have been demonstrated in the literature. Two studies are reviewed here in terms of the IPD modeling techniques used.

Wu, Li, Ng, and Zumbo (2006) modeled IPD across three administrations of the Third International Mathematics and Science Study using a separate logistic regression for each item. Logistic regression was highlighted for its ability to include grouping variables with more than 2 categories and interactions between groups and ability, where models of increasing complexity could be tested for significance sequentially against one another. Though the majority of items were flagged for IPD, the logistic-regression based effect sizes were found to be negligible.

Bock, Muraki, and Pfeiffenberger (1988) proposed a system for managing IPD and thereby maintaining a stable IRT scale. Using data from five administrations of the College Board Physics Achievement Test, they estimated the effects of IPD using a series of time-dependent IRT models. These included a base model, with all item parameters constant across examinee groups; a linear drift model, with item difficulty changing linearly across time or examinee groups; a quadratic drift model, with a second-order polynomial term for the item difficulty by group term; a model with a separate item difficulty estimated for each group; and a model with separate item discriminations and difficulties estimated for each group. Based on likelihood ratio chi-square tests, the linear item difficulty drift model fit the data best.

These studies demonstrated two related modeling approaches that have proven to be useful in detecting and estimating the impact of IPD. The purpose of the present study is to build on this work by demonstrating a logistic regression model for estimating the impact of IPD within a multilevel framework. The model is formulated as a hierarchical generalized linear model (HGLM), one which is able to accommodate nested data structures while incorporating covariates at the item, person, and other grouping levels (e.g., Kamata, 2001; Pastor, 2003). The IPD HGLM is demonstrated using data from three administrations of a statewide survey of middle school and high school students.

Method

Data

Item-level data were obtained from the 2004, 2007, and 2010 administrations of the Minnesota Student Survey. Table 1 contains sample sizes for each cohort (i.e., year). Percentages across grades are for each year's cohort; thus, percentages sum to 100 across rows. Gender was split roughly evenly between females and males, and the mean age at each year was 14 years.

As the survey was administered to 6th, 9th, and 12th graders, many of the participants were likely present across more than one administration; however, the data were deidentified, making it impossible to link across time by people. Similarly, many of the items were either dropped or revised across administrations as the survey was updated, making it impossible to link the majority of scales across time by items.

Table 1

Sample Sizes Across Grades for Each Year (Cohort)

Year	N	Grade 6	Grade 9	Grade 12
2004	19,471	34%	38%	27%
2007	20,666	36%	37%	27%
2010	19,863	35%	36%	29%

Five survey scales contained the same item sets across all three time points. These were each examined in terms of relevance to educational outcomes and likelihood of containing items with drifting parameters. A set of eight items assessing safe/unsafe experiences at school and perceptions of school safety was chosen for further analysis. The eight survey items are included in Table 2. Students responded to the first four items using a 4-point rating scale, containing the options strongly agree, agree, disagree, and strongly disagree. A dichotomous yes/no response was given for the next three items. The eighth item included a 5-point scale, with the following options: 0 times, 1 time, 2 or 3 times, 4 or 5 times, and 6 or more times. The polytomous responses were dichotomized to a 0/1 scale, with items 1 through 4 collapsed as 1, 2 = 0 and 3, 4 = 1, and item 8 as 1 = 0 and 2, 3, 4, 5 = 1. Item responses were then recoded so that positive values indicated higher safety ratings, and lower values indicated lower safety ratings.

Table 2

Eight School Safety Items

Item #	Item content
	How much do you agree or disagree with the following statements?
1	I feel safe going to and from school
2	I feel safe at school
3	Bathrooms in this school are a safe place to be
4	Illegal gang activity is a problem at this school
	During the last 12 months, which of the following has happened to you on school property? Has a student…
5	threatened you?
6	pushed, shoved, or grabbed you?
7	kicked, bitten, or hit you?
8	During the last 12 months, how many times has someone stolen or deliberately damaged your property such as your car, clothing, or books on school property?

HGLM

The traditional Rasch (1960) model, written in terms of the probability of correct response to item i for person j,

$$P(y_{ij} = 1|\gamma_i, u_j) = \frac{1}{1 + \mathrm{e}^{-(u_j - \gamma_i)}}, \quad (1)$$

can also be described as a logistic regression model, in terms of the log-odds of correct response:

$$\log \frac{P}{1 - P} = \eta_{ij} = \gamma_i + u_j. \quad (2)$$

Here, η_{ij} represents the log-odds that $y_{ij} = 1$, γ_i represents the difficulty of item i, and u_j represents the ability or trait level for person j. η_{ij} is modeled as a summation of item difficulty and person ability, rather than a difference, which means that the item effect is expressed as an item easiness, where a higher value indicates an easier item. For the survey items analyzed in this

study, γ_i would be interpreted as an average log-odds of endorsement for item i, where higher values indicate that the item was easier to endorse.

The HGLM extends (2) to a hierarchical framework which is applied to all $N \times J$ item responses at level 1 across J people at level 2:

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij}. \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} \\ &\vdots \\ \beta_{(N-1)j} &= \gamma_{(N-1)0}. \end{aligned} \tag{3}$$

In this base model (M0), the intercept γ_{00} is the easiness parameter for a selected reference item, here item *N*, and the terms γ_{q0} are parameters for the remaining items expressed as differences from the reference, where the item indicator variable $X_{qij} = 1$ when $q = i$ and $X_{qij} = 0$ otherwise (for additional details see Kamata, 2001; Kamata, Bauer, & Miyazaki, 2008).

To center the scale at the mean item difficulty, or easiness, rather than the easiness of the reference item, the item indicators X could be grand-mean centered (Raudenbush & Bryk, 2002; Cheong, 2006). Since all items were seen by all people, this would be equivalent to coding $X_{qij} = (N-1)/N$ when $q = i$ and $X_{qij} = -1/N$ otherwise. Reduced to a single item q, model M0 then becomes

$$\eta_{ij} = \gamma_{00} + \gamma_{q0} + u_{0j}, \tag{4}$$

which is equivalent to the Rasch model, where the item difficulty is expressed as $\gamma_{00} + \gamma_{q0}$, the mean item easiness parameter and the item q deviation from the mean. As in (3), an indicator for the reference item is not included in the model. However, with mean-centering, the coefficient

γ_{00} no longer represents the easiness for the reference item. The reference item parameter must be obtained indirectly, either by combining the remaining item effects and the mean as described below, or by rerunning the model with a different item as the reference.

In Equations (1) through (4) it is assumed that no other characteristics of item *i*, beside its difficulty, and no other trait or ability for person *j*, beside u_j, are necessary to describe the relationship between the two in terms of the probability that $y_{ij} = 1$. IPD violates this assumption by requiring an additional parameter, a cohort effect, in the linear component $\gamma_i + u_j$. In order to separate the IPD effect of cohort on an individual item from the overall change in safety by cohort, the intercept β_{0j} is first conditioned on the cohort covariate W:

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij}. \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} \\ &\vdots \\ \beta_{(N-1)j} &= \gamma_{(N-1)0}. \end{aligned} \tag{5}$$

Reduced to a single item q, this model (M1) becomes:

$$\eta_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{q0} + u_{0j}. \tag{6}$$

With W coded as 0, 1, 2, M1 models the log-odds as a function of the grand mean log-odds safety rating γ_{00} at cohort 0, the average change in log-odds for a 1 unit change in cohort γ_{01}, the additional effect γ_{q0} associated with item q, and the safety level u_{0j} for person j. Controlling for the average change in safety rating associated with cohort is equivalent to controlling for ability differences across groups in a DIF framework. The term γ_{01} is referred to as a main effect for cohort.

To estimate bias introduced by IPD, model M2 includes the cohort covariate W within the remaining level-2 models:

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij}. \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} W_j \\ &\vdots \\ \beta_{(N-1)j} &= \gamma_{(N-1)0} + \gamma_{(N-1)1} W_j. \end{aligned} \tag{7}$$

Reduced to a single item q, M2 becomes

$$\eta_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{q0} + \gamma_{q1} W_j + u_{0j}. \tag{8}$$

The term γ_{q1} estimates the expected linear change in item location γ_{q0}, on the logit metric, for a one point change in cohort W, after controlling for the average cohort effect γ_{01}. IPD for item q is thus expressed as a difference from the mean IPD effect, just as the item easiness for item q is expressed as a difference from the average easiness γ_{00}. A significant cohort effect suggests that an item parameter is not invariant over different cohorts (see Rupp & Zumbo, 2006, for further discussion of IPD and IRT parameter invariance).

Analysis

Models M0, M1, and M2 were estimated with the statistical software *HLM6*, using a Laplace approximation to maximum likelihood. Item 8 was set as the reference item. The models were first compared based on fit statistics AIC, BIC, and χ^2 likelihood ratio. These served as omnibus tests of the appropriateness of the main effect for cohort in M1, compared to M0, and the item-cohort interaction terms in model M2, compared to M1. Next, individual item cohort effects were examined for significance. Results are reported below for each model.

Results

Table 3 contains the model fit results for M0 versus M1 and M1 versus M2. The likelihood ratio tests were both statistically significant (M1, $\chi^2 = 265.42$, *p*-value < .0001; M2, $\chi^2 = 99.42$, *p*-value < .0001). AIC and BIC were smaller for the more complex models in both comparisons.

Table 3

Model Fit Results Comparing M0 and M1

Model	*df*	AIC	BIC	Deviance	logLik	χ^2	$\chi^2 df$	*p*-value
M0	9	1285818	1285918	1285800	-642900			
M1	10	1285555	1285665	1285535	-642767	265.42	1	< .0001
M2	17	1285469	1285658	1285435	-642718	99.42	7	< .0001

Table 4 contains the item and item-cohort IPD effects for models M0, M1, and M2. Because each item indicator *X* was mean-centered, the following equation was used to obtain the non-reference item effects, i.e., the non-reference item mean log-odds:

$$\eta_{ij} = \gamma_{00} + \frac{7}{8}\gamma_{q0} - \frac{1}{8}\sum_{all\ q'}\gamma_{q'0}, \tag{9}$$

where $q' \neq q$. The reference item effect was not directly estimated in M2, and was obtained by:

$$\eta_{ij} = \gamma_{00} - \frac{1}{8}\sum_{q=1}^{7}\gamma_{q0}. \tag{10}$$

Similarly, the non-reference IPD effects for M2 were obtained using

$$\eta_{ij} = \gamma_{01} + \frac{7}{8}\gamma_{q1} - \frac{1}{8}\sum_{all\ q'}\gamma_{q'1}, \tag{11}$$

and the reference item IPD was obtained with

$$\eta_{ij} = \gamma_{01} - \frac{1}{8}\sum_{q=1}^{7}\gamma_{q1}. \quad (12)$$

Since Equations (9) and (10) exclude the cohort effects, they represent item effect estimates at $W = 0$, and they are thus reduced for M1 and M2 in comparison to M0. Statistical tests for the item and item-cohort effects were not considered, since the test statistics and *p*-values returned by *HLM6* are designed to test for differences from zero, which were not of interest. Instead, the main interest was to examine the magnitude of the main effect for cohort, γ_{01}, from M1 and M2, and the IPD effects from M2, which were obtained using Equations (11) and (12). As indicated by the main effects for cohort (0.17 for M1, and 0.15 for M2), students' average perceived safeness at school was estimated to increase by an average of 0.17 and 0.15 logits per survey administration. Were the cohort covariate *W* instead coded as year (e.g., 2004, 2007, 2010), this would result in a logit change per year of 0.05 for M2. These estimates indicate that students are, on average, rating their schools as safer in later years.

After controlling for average change in perceived school safeness, the M2 cohort effects (IPD) were all smaller than 0.10 in absolute value. Item easiness for items 2, 3, and 4 were negative and were thus estimated to decrease by 0.03, 0.05, and 0.09 logits respectively for each cohort. Easiness for the remaining items were estimated to increase by cohort, with the largest increase being 0.09 logits for item 8. Although model M2 seemed to fit the data best, according to the results in Table 3, the magnitudes of M2 IPD at the item level were small after controlling for average change in safety rating via γ_{01}.

Table 4

Estimates of Item and Cohort Effects for Models M0, M1, and M2

Item	Effect	M0	M1	M2
Mean	γ_{00}	2.10	1.77	1.79
	γ_{01}		0.17	0.15
1	γ_{10}	3.87	3.54	3.52
	γ_{11}			0.02
2	γ_{20}	3.44	3.11	3.19
	γ_{21}			-0.03
3	γ_{30}	2.20	1.87	1.99
	γ_{31}			-0.05
4	γ_{40}	2.34	2.01	2.22
	γ_{41}			-0.09
5	γ_{50}	1.84	1.51	1.45
	γ_{51}			0.04
6	γ_{60}	0.54	0.21	0.18
	γ_{61}			0.03
7	γ_{70}	1.57	1.24	1.25
	γ_{71}			0.01
8	γ_{80}	0.98	0.65	0.50
	γ_{81}			0.09

Note: Mean represents the intercept. Effects for the reference item 8 were not estimated, but were obtained using Equations (10) and (12).

Figure 1 contains a plot of the combined M2 item difficulty and cohort effects for each item across cohorts 1, 2, and 3. Effects are in the log-odds or logit metric on the *y*-axis. The slopes of each line indicate a combination of the average logit change by cohort, captured by γ_{01}, and the item IPD, captured by γ_{q1}. As indicated by the small IPD values in Table 4, all of the items appear to share the average positive slope across cohort γ_{01}, with only slight deviations from it in γ_{q1}.

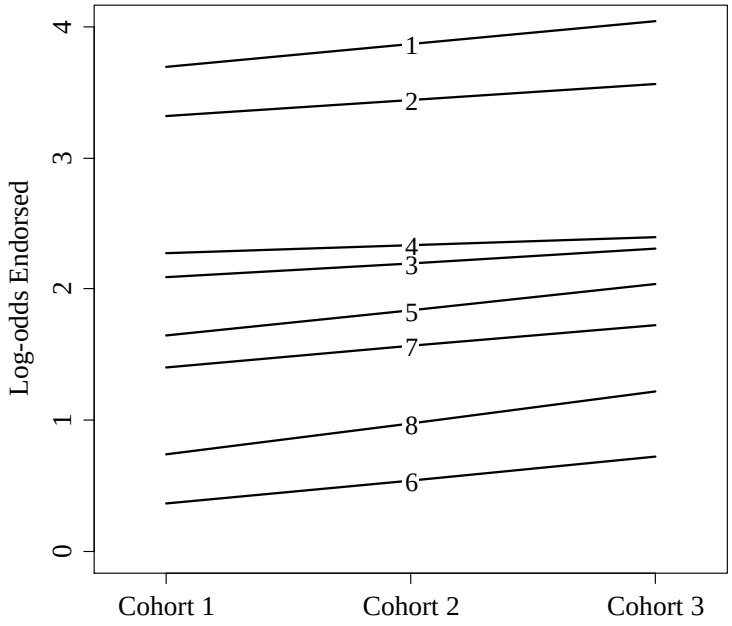

Figure 1. M2 item effects and cohort slopes. Each line represents an item and is labeled with the item number.

Discussion

This study demonstrates a unique application of the HGLM to a scenario where variability in item parameters may reduce the stability of the measurement scale. The IPD HGLM provides estimates of the effect of IPD on each item parameter simultaneously, while also controlling for person ability. Additional benefits include the potential for considering a third level of nesting, say, students within classrooms or districts, and other covariates and grouping variables, for example, to examine differential item functioning across gender or ethnicity and interactions between these covariates and the cohort.

Increases in observed proportion endorsed for these survey items would indicate, overall, an increase in student ratings of school safety, and vice versa for decreases. Within the HGLM, changes in proportion endorsed, or log-odds endorsed, over time are estimated while controlling

for overall differences in safety ratings by cohort. As a result, the slopes represent change in responses to the safety items for students having the same overall safety score. These changes can be interpreted as bias due to student cohort.

The model fit results support the inclusion of the interaction terms in the IPD model M2. However, investigation of individual effects reveal small IPD estimates. In terms of DIF, researchers have identified effects greater than or equal to 0.50 logits as problematic (e.g., Cheong, 2006). Still, small logit changes associated with cohort may deserve further investigation.

A confounding factor not addressed in this study is the presence of the same students across cohorts. Since three years had passed between administrations it seemed that, should students appear in the survey sample at more than one cohort, they would likely not recall their previous responses. Thus, in this study responses across time were treated as if they were unique at the person level. An additional limitation is the small item set, which is expected to produce less reliable estimates of u_{0j} than would a longer scale. The survey scale used in this study was useful for demonstration purposes. However, with educational tests, especially high-stakes ones, additional items should be used to provide a more comprehensive and reliable estimate of ability.

The majority of scales in measurement applications are maintained beyond an initial test administration, often across multiple years and many cohorts of examinees. The IPD HGLM has practical applications and implications for measurement in education, in that it can be used to improve the stability of a measurement scale, and thus improve the accuracy of student ability and growth estimates and decisions used in the placement and classification of students.

References

Babcock, B., & Albano, A. D. (2011). *Rasch scale drift over time: Examining when to reset the scale.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1998). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25,* 275-285.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing, 6*, 57-79.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20,* 369-377.

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

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 345-388). Charlotte, NC: Information Age Publishing.

Pastor, D. A., (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education, 16*, 223-243.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement, 30*, 100-120.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*, 63-84.

Wu, A. D., Li, Z., Ng.,S. L., & Zumbo, B. D. (2006). *Investigating and comparing the item parameter drift in the mathematics anchor/trend items in TIMSS between Singapore and*

the United States. Paper presented at the International Association for Educational Assessments, Singapore.