Investigating Linear and Nonlinear Item Parameter Drift with Explanatory IRT Models

Luke Stanke
University of Minnesota

Okan Bulut
University of Alberta

Michael C. Rodriguez & José Palma
University of Minnesota

Minnesota Youth Development Research Group

April, 2016

Paper presented at the annual meeting of the
American Educational Research Association
Washington DC.

Citation:

Stanke, L., Bulut, O., Rodriguez, M.C., & Palma, J. (2016, April). *Investigating linear and nonlinear item parameter drift with explanatory IRT models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington DC.

Investigating Linear and Nonlinear Item Parameter Drift with Explanatory IRT Models

INTRODUCTION

Test scores can be distorted by shifts in item performance over time because of cognitive or noncognitive examinee characteristics (Bulut, Palma, Rodriguez, & Stanke, 2015), examinees' opportunities to learn (Albano & Rodriguez, 2013), and changes in curriculum and teaching methods (DeMars, 2004; Miller & Linn, 1988). In the context of item response theory (IRT), these distortions in item parameters over multiple administrations of a test are called item parameter drift (IPD; Goldstein, 1983).

IPD is typically considered as a result of construct-irrelevant variability in test items over time. However, drift can also occur due to the difference in the perception of a construct across grade spans and developmental levels in a single occasion. This type of drift is construct relevant. Martineau (2006) described the presence construct-relevant variability in item parameters as construct shift.

Given the constant physical, social, and emotional development of students through grades at school, developmental measures are more likely to be exposed to construct shift. In standard measurement literature, IPD is considered construct irrelevant, however this drift might really be construct relevant drift. Like IPD, the presence of construct shift may still lead to systematic errors in equating, scaling, and consequently scoring (Kolen & Brennan, 2004). By using incorrect measurement models, our assumptions about students' scores on these developmental measures are false, which might lead to inappropriate conclusions about students' developmental performance.

METHODS

Research Questions

This proposal has one research question, however more are answered for the presentation: How often does IPD model or construct shift misspecification occur under controlled conditions? For this study, we study construct shift as measure of IPD.

Simulation Conditions

Table 1 shows the simulation conditions of the study. The simulation conditions of this study included drift type (linear, quadratic, and offset quadratic), drift magnitude (0, 0.1, 0.2), and sample size (500, 1000), resulting in 18 crossed conditions. Test length was fixed to 10 items and there were seven hypothetical grade levels for all crossed conditions.

Data Generation

Item difficulty parameters and abilities were drawn from a normal distribution. Two items were considered as anchor items with no drift across grades. The remaining items were modified to drift linearly (i.e., the same magnitude of drift across grades) and nonlinearly (i.e., the magnitude of drift increases quadratically across grades). 50 response data sets were generated for each crossed condition.

Data Analysis

Four explanatory IRT models were fit to the data sets using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) in R: 1) Rasch model assuming model invariance; 2) linear IPD model with grade as a continuous predictor; 3) Quadratic IPD model with grade as a continuous predictor; and 4) nonlinear IPD model with grade as a categorical predictor. For brevity, the Quadratic IPD model is specified as

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \theta_p - (\beta_{i0} + \beta_{i1}(Grade) + \beta_{i2}(Grade)^2)$$

where the log odds of obtaining a correct response, Pij, is equal the trait level of an individual, minus the item difficulty, β_{i0}, the linear drift constant, β_{i1}, and the quadratic drift constant, β_{i2}, of item *i*.

RESULTS

The simulation results summarizing model fit are presented in Tables 2 and 3. The models with the lowest AIC and BIC values were selected as the best-fitting model in each of the 50 replications across 18 crossed conditions. The findings suggest that the factor IPD model provided the best model fit under nonlinear and offset nonlinear conditions. However, the quadratic IPD model outperformed the factor IPD model when sample size was small. When linear IPD was present, the linear IPD model was the best fitting model due to its greater parsimony in explaining linear drift.

CONCLUSION

This study shows that either factor or quadratic IPD models can be highly useful in detecting drift when the amount of drift varies across testing occasions or grade levels. If, however, linear IPD is present, quadratic and factor IPD models may be highly redundant and laborious. Considering the computational demands of the three IPD models, one may want to begin with a linear IPD model first, and then move to the quadratic or factor IPD models if the magnitude of drift is not fixed across grade levels. In addition to model fit results, the impact of drift on estimated item parameters and ability estimated will be discussed in the final form of the proposal.

While this paper only examined one research question, our presentation would expand on model selection, bias, and person estimates. Practically, if we are obtaining responses of students and these response are not being scaled correctly, we may be producing incorrect conclusions about the developmental levels of students. As more and more schools are interested in understanding the developmental levels of students in areas other than academic performance, correct measurement models will become even more important, as will understanding construct variability.

Table 1

Simulation Design of the Study

Cell	Drift Type	Drift Magnitude	Sample Size
1	Linear	None	Small
2	Linear	None	Large
3	Linear	Small	Small
4	Linear	Small	Large
5	Linear	Large	Small
6	Linear	Large	Large
7	Quadratic	None	Small
8	Quadratic	None	Large
9	Quadratic	Small	Small
10	Quadratic	Small	Large
11	Quadratic	Large	Small
12	Quadratic	Large	Large
13	Offset Quadratic	None	Small
14	Offset Quadratic	None	Large
15	Offset Quadratic	Small	Small
16	Offset Quadratic	Small	Large
17	Offset Quadratic	Large	Small
18	Offset Quadratic	Large	Large

Table 2

Proportion of Models with the Lowest BIC Value by Cell

Simulation Conditions				IPD Models			
Cell	Drift Type	Drift Magnitude	Sample Size	Rasch	Linear IPD	Quadratic IPD	Factor Model
1	Linear	None	Small	1.00	.00	.00	.00
2	Linear	None	Large	1.00	.00	.00	.00
3	Linear	Small	Small	.00	1.00	.00	.00
4	Linear	Small	Large	.00	1.00	.00	.00
5	Linear	Large	Small	.00	1.00	.00	.00
6	Linear	Large	Large	.00	1.00	.00	.00
7	Nonlinear	None	Small	.00	.00	.80	.20
8	Nonlinear	None	Large	.00	.00	.00	1.00
9	Nonlinear	Small	Small	.00	.00	.80	.20
10	Nonlinear	Small	Large	.00	.00	.00	1.00
11	Nonlinear	Large	Small	.00	.00	.92	.08
12	Nonlinear	Large	Large	.00	.00	.00	1.00
13	Offset Nonlinear	None	Small	.00	.00	1.00	.00
14	Offset Nonlinear	None	Large	.00	.00	1.00	.00
15	Offset Nonlinear	Small	Small	.00	.00	1.00	.00
16	Offset Nonlinear	Small	Large	.00	.00	1.00	.00
17	Offset Nonlinear	Large	Small	.00	.00	1.00	.00
18	Offset Nonlinear	Large	Large	.00	.00	1.00	.00

Table 3

Proportion of Models with the Lowest AIC Value by Cell

Simulation Conditions				IPD Models			
Cell	Drift Type	Drift Magnitude	Sample Size	Rasch	Linear IPD	Quadratic IPD	Factor Model
1	Linear	None	Small	1.00	.00	.00	.00
2	Linear	None	Large	.96	.04	.00	.00
3	Linear	Small	Small	.00	1.00	.00	.00
4	Linear	Small	Large	.00	.98	.02	.00
5	Linear	Large	Small	.00	1.00	.00	.00
6	Linear	Large	Large	.00	.98	.02	.00
7	Nonlinear	None	Small	.00	.00	.00	1.00
8	Nonlinear	None	Large	.00	.00	.00	1.00
9	Nonlinear	Small	Small	.00	.00	.00	1.00
10	Nonlinear	Small	Large	.00	.00	.00	1.00
11	Nonlinear	Large	Small	.00	.00	.00	1.00
12	Nonlinear	Large	Large	.00	.00	.00	1.00
13	Offset Nonlinear	None	Small	.00	.00	.00	1.00
14	Offset Nonlinear	None	Large	.00	.00	.00	1.00
15	Offset Nonlinear	Small	Small	.00	.00	.00	1.00
16	Offset Nonlinear	Small	Large	.00	.00	.00	1.00
17	Offset Nonlinear	Large	Small	.00	.00	.00	1.00
18	Offset Nonlinear	Large	Large	.00	.00	.00	1.00

References

Albano, A. D., & Rodriguez, M. C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational and Psychological Measurement, 73*, 836–856.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4.* R package version 1.1-8. Retrieved from http://CRAN.R-project.org/package=lme4

Bulut, O., Palma, J., Rodriguez, M. C., & Stanke, L. (2015). Evaluating measurement invariance in the measurement of developmental assets in Latino English language groups across developmental stages. *Sage Open*, *5*(2), 1–18.

DeMars, C.E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*, 265–300.

Kolen, M. & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Miller, A. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, *25*, 205–219.