

# The Robustness of Vertical Scaling Methods to Violation of Unidimensionality

Liquan Yin

Dr. Suzanne lane

Research of Methodology

[Liy15@pitt.edu](mailto:Liy15@pitt.edu)

(412)366-0902

## 1) Statement of problem

In recent years, many states adopted Item Response Theory (IRT) based vertically scaled tests in response to the federal education policy NCLB since IRT-based scales have several compelling features under accountability context. Selecting a practical and effective method among different vertical scaling methods is critical to educators and other related researchers. Comparisons among various IRT-based scaling methods then become necessary. One fundamental assumption of IRT-based scaling is that item responses are based on the same skill or same composite of multiple skills. This assumption is very likely to be violated in vertical scaling context since the tests to be scaled usually cross a wide range (*e.g.*, grade 3 to 8). It remains an open question whether or not a unidimensional IRT model can be successfully applied to vertical scaling that may not be strictly unidimensional in nature (Boughton, Lorie, & Yao, 2005). Therefore, investigating the behavior of these various scaling methods under different test conditions is the main motivation of this study. More specifically, this study focuses on the robustness of unidimensional scaling methods to the inevitable violation of unidimensional assumptions while developing vertical scales in practice.

The following are the two main research questions:

- 1) When unidimensional assumption holds, which IRT scaling methods (concurrent, semi-concurrent, and separate calibrations by using ICC, TCC, and mean/sigma linking methods) yield less biased ability estimates in vertical scaling context?
- 2) When unidimensional assumption does not hold, which IRT calibration /scaling methods (same as those in question 1) yield less biased ability estimates in vertical scaling context?

## 2) Theoretical Framework

Vertical scaling refers to a conversion of raw scores onto a scale that is common to all assessments which have similar constructs but with different difficulties across grades. With a vertical scale, the students' or schools' year-to-year growths and changes then are comparable (Kolen & Brennan, 2004). Early scaling studies mostly focused on comparisons among traditional Thurstone scaling methods. With well-developed Item Response Theory, IRT-based scaling methods have received particularly considerations in early 1980s because of its independency of item characteristics (Yen & Burket, 1997).

Common-items non-equivalent groups design is the most widely used vertical scaling design. In the Common-items design, tests appropriate for each grade level (referred to as level tests) are constructed with a set of common item blocks. Common items are used as a "bridge" to link adjacent grades. This study will focus on the common-item design.

The process of estimating the item parameters is referred to as "calibration" in IRT. There are two general IRT-based calibrations in vertical scaling: concurrent and separate calibrations. In concurrent calibration, a vertical scale is established by calibrating data from

all grade levels in a single computer run. After concurrent calibration, the item parameter estimates and ability estimates for all grades are already on the same scale.

When separate calibration is used, the IRT parameters are estimated separately for each test. These estimations are usually based on different scales due to IRT indeterminacy problem. Scaling methods (such as mean/sigma, TCC, and ICC) are then used to put item parameters on a common metric (Kolen & Brennan, 2004; Kim & Cohen, 1998).

Theoretically, concurrent calibration is expected to produce more stable results because it uses all information simultaneously. When vertical scale is developed across several grades, separate calibration might be safer since the violation of unidimensionality is less distorted between adjacent grades (Kolen & Brennan, 2004). On the other hand, separate calibration might introduce more measurement errors since much more linking steps are involved. Furthermore, different linking methods also likely behave differently. Comparisons among these scaling methods are necessary and critical in vertical scaling context (Tong & Kolen, 2007). Very limited research has been done to compare these calibration/scaling methods in vertical scaling context. Especially, very few researchers explored the robustness of various scaling procedures to the inevitable violation of unidimensional assumptions in developing vertical scales.

Therefore, the purpose of this simulation study is to compare different scaling methods under different conditions, both unidimensional condition and non-unidimensional conditions.

### 3) Methods

*Vertical scaling design and IRT model.* In this common-items design, the scaling processes span 6 grades, grade 3 through grade 8. Grade 5 is treated as a base grade (Figure 1):

Grades	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Grade 3	G3_on					
Grade 4	G3_c	G4_on				
<b>Grade 5</b>		G4_c	G5_on			
Grade 6			G5_c	G6_on		
Grade 7				G6_c	G7_on	
Grade 8					G7_c	G8_on

(Figure 1: common-items non-equivalent groups design)

IRT model on this study is an unidimensional 3P logistic IRT model (Embretson & Reise, 2000):  $p(u = 1/\theta) = c + (1 - c) \frac{e^{Da(\theta_j - b)}}{1 + e^{Da(\theta_j - b)}}$ , where  $a, b,$  and  $c$  are item parameters;  $\theta$  is the ability level, and  $D$  is a scaling parameter.

*Simulation.* Item parameters and population ability distributions for simulation are based on the real results of Florida Comprehensive Assessment test (FCAT, 2006). Simulation

based on a real dataset is to get reasonable item parameters and avoid awkward combinations from computer random selection.

***Manipulated Factors.***

- I. Calibration and scaling methods (5 levels): concurrent calibration, separate calibrations by using different scaling methods: Mean-sigma, ICC, and TCC, and semi-concurrent calibration, which is a hybrid of concurrent and separate calibrations (Meng, 2007).
- II. Data structure and correlations between adjacent grade levels (3 levels): unidimensional, non-unidimensional with  $r = .85$  (high), and non-unidimensional, with  $r = .65$  (medium to high). *E.g.*, for grade 5 students, their performance on on-grade level items and on common items which are from grade 4 are likely different.  $r$  is the correlation between examinee's performances on these items.
- III. Number of common items(2 levels): 20% or 30%

***Fixed factors.*** Sample size,  $N = 1000$ . EAP and MAP are used to get ability estimates. Total number of items is fixed at 60 for each grade from grade 3 to grade 8. The study is replicated for 100 times to get stable results.

***Software and programs.*** The simulation is programmed in SAS 9.1 and BILOG-MG program is called in for calibrations and ability estimations. ST program is also called in to get ICC and TCC linking parameters.

***Evaluation criterion.*** The ability estimations are compared to "true" ability ( $\theta$ ) which is used as a base for previous data simulation for analyses. Both bias,  $\text{Bias} = \hat{\theta}_i - \theta_i$ , which gives the direction of bias and root-mean-square-deviation (RMSD) which gives the magnitude of bias are used for comparisons. Then the bias and RMSD are averaged across 100 replications and thus 180 means of RMSDs ( $5 \times 3 \times 2 \times 6$ ) and 180 means of biases are in final comparisons.

#### **4) Data analysis and anticipated conclusions**

The means of averaged RMSD across 100 replications and means of averaged biases are compared under 30 conditions for 6 grades. The descriptive statistics and graphs are the main source for comparisons. In addition, Analysis of variance (ANOVA) is also used to check if there is significant difference among the biases by using different calibration methods.

In general, concurrent calibration procedure is expected to produce better results than separate calibration when tests are assumed to be unidimensional. When tests are non-unidimensional, separate calibration methods are expected to yield better estimates. The performances among different linking methods are not predictable at present. When the correlation between grades is higher, the biases are expected to be smaller by using either concurrent calibration or separate calibration. The separate calibrations are expected to be robust against mild violation of unidimensionality. The scaling results are likely to be distorted by severe violation of unidimensionality such as  $r = .65$ .

## 5) Educational Significance

Selecting a practical and effective method among various vertical scaling methods is a real challenge to educators and other related researchers, especially under growth-based accountability system (Thum, 2003). This study evaluates the scaling methods under the most popular vertical scaling design, common-items design. Because there is very limited research under vertical scaling context, and almost no investigation of the robustness to the violation of unidimensional assumptions which is very typical in developing vertical scales in practice, the results reported in this study have significant implications for applied researchers and testing professionals at state assessment programs. Since assessment programs often vertically scale results across grade levels, it is critical that assessment professionals select an effective and practical scaling method. It is also important that these professionals understand that the different calibration methods affect scaling results differently under different test conditions.

## 6) References

- Boughton, K. A., Lorie, W., & Yao, L. (2005). A Multidimensional Multi-Group IRT Model for Vertical Scales with Complex Test Structure: An Empirical Evaluation of Student Growth using Real Data. *Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.*
- Embretson, S. E & Reise, S. P (2000) *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Inc.
- Kim, S.-H. & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22 (2), 131-143.
- Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking (2nd)*. NY: Springer Science + Business media, Inc.
- Meng, H. (2007). A comparison study of IRT calibration methods for mixed-format tests in vertical scaling. Un-published dissertation in University of Iowa.
- Thum, Y.M. (2003). *No Child Left Behind: Methodological Challenges & Recommendations for Measuring Adequate Yearly Progress* (No. 590). Los Angeles, CA: Center for the Study of Evaluation, University of California, Los Angeles (CSE).
- Tong, Y. & Kolen, M. J. (2007) Comparisons of Methodologies and Results in Vertical Scaling for Educational Achievement Tests. *Applied Measurement in Education*, 20 (2), 227-253.
- Yen, W. M. & Burket, G. R. (1997). Comparison of Item Response Theory and Thurstone Methods of Vertical Scaling. *Journal of Educational Measurement*. 34 (4), 293-313.

Budget of this study

Purchase of software BILOG-MG: \$300.00

Total: \$300.00

Item parameter and ability estimates are critical to this simulation study. BILOG-MG is used to get item parameter and ability estimates. It provides Expected a Posterior (EAP) estimation which is one popular estimation method in many state assessment programs. This software has a long-term use. It can be beneficial to many students in school of education, especially for students in Psychology in Education Department.