ABSTRACT

Title of Document: A COMPARISON OF DIFFERENT METHODS THAT DEAL WITH CONSTRUCT SHIFT IN VALUE ADDED MODELING: IS VERTICAL SCALING NECESSARY?

Yong Luo, Doctor of Philosophy, 2013

Directed By: Professor Hong Jiao
Department of Human Development and Quantitative Methodology

Construct shift is a term used to describe the change of tests in the construct they intend to measure. In tests across multiple grades where curriculum change occurs, construct shift is expected to exist. This presents a problem to many VAM models that assume scores across multiple grades are on a common developmental scale, since these scores cannot be placed on the same scale through vertical scaling. There are three methods currently available to deal with construct shift: the CU method ignores construct shift, carry out the vertical scaling process with a unidimensional IRT model, and directly use the vertically scaled scores in specific VAM models that require vertical scaling; the CB method models construct shift, carry out the vertical scaling process with a bifactor model and use the scores on the general factor in specific VAM models that require

vertical scaling; the SU method does not use vertical scaling but directly applies the scores at each grade in the generalized persistence (GP) model. A simulation study was conducted to compare the impacts of construct shift upon teacher rank ordering estimation with those three methods.

Results suggest that the performances of all three methods are subject to the influence of magnitude of construct shift and choice of teacher effect persistence pattern. The CB and the CU methods perform similarly, while the SU method is superior to them in most simulation conditions. Only with large magnitude of construct shift is the CB method slightly better than the SU method in terms of the last year's teacher effect estimation. The CB method performs better than the CU method with large magnitude of construct shift, while they perform similarly with small or medium magnitude of construct shift. It is concluded that the SU method performs the best among those three methods and is recommend for use in practice.

A COMPARISON OF DIFFERENT METHODS THAT DEAL WITH CONSTRUCT SHIFT IN VALUE ADDED MODELING: IS VERTICAL SCALING NECESSARY?

By

Yong Luo

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Hong Jiao, Chair
Professor Robert Lissitz
Professor Gregory Hancock
Professor Robert Croninger
Professor Laura Stapleton

Dedication

This dissertation is dedicated to my parents Guiming Luo and Runcai Liu.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my dissertation advisor, Dr. Hong Jiao, for her guidance, encouragement and support in the past a few years. I have been constantly inspired by her passion in research and seemingly limitless research ideas. I am also grateful to Dr. Robert Lissitz for being so generous with his time whenever I wanted to discuss some research ideas, as well as his providing me with a RA position at the Maryland Assessment Research Center for Education Success (MARCES) when I needed financial assistance the most. I want to thank Dr. Gregory Hancock for trusting me with the first RA position in his collaboration project with ETS and helping me find a summer internship at Certified Financial Planner (CFP). He is undoubtedly the best and funniest teacher I've ever met. My gratitude also goes to the other two committee member, Dr. Laura Stapleton and Dr. Robert Croninger, for their invaluable comments and suggestions. I also want to thank Dr. Jeffrey Haring for all the wonderful classes I took with him, in which I learnt a lot in both statistics and programming. Last but not least, I want to thank Dr. Ji Seung Yang for her help with the automation of IRTPRO, without which I might still be manually running IRTPRO.

I am also grateful to Dr. Louis Mariano at the Rand Corporation. I met him at the psychometric society conference in Nebraska and he pointed me to his paper about persistence model in VAM, which became a major component of this dissertation. My gratitude also goes to Dr. Ying Li at American Nursing Association, for all the inspiration I drew from her dissertation and numerous conversations I had with her.

Finally, I want to thank my parents for their unconditional support and love. Both of them only finished elementary school, but they started emphasizing the importance of education with wisdom and foresight when I was a little boy. I hope they will be proud of their son and themselves. I apologize with all my heart to them for those years of absence and the failure to be a filial son.

Table of Contents

CHAPTER 1: INTRODUCTION

1.1 Background

The 2002 No Child Left Behind (NCLB) Act, which "marks decisive shift away from evaluating districts and schools on the basis of inputs to judging them on the basis of outcomes" (Braun & Wainer, 2007), stipulates that all students in grades 3-8 and one high school grade be tested in reading and mathematics by 2006 and in science by 2008. As a result, testing of students in k-12 setting with standardized assessment has grown rapidly during the past decade. States and districts have been actively expanding their testing program and data system, which usually includes student achievement data across multiple years and linkages between students and teachers. In 2009, the Obama administration announced the Race to the Top Assessment Program to fund states' development of valid and informative assessments to ensure that students gain the knowledge and skills needed for college and workplace readiness. It is anticipated that statewide testing programs and the corresponding data system will continue to grow.

While such longitudinal data can be used to track students' growth, the linkage between the teacher and the students makes the evaluation of educational effectiveness of teachers and schools possible. Value-added modeling (VAM), a family of statistical models that use students' scores as the outcome variable and estimate the contribution of teachers or schools to learning, is a popular approach that is gaining momentum recently. It is believed that VAM provides a fair comparison of teachers or schools with proper control of the effects due to the demographic and other relevant covariates of their students, thus it is superior to other approaches that only consider the proportion of

students reaching adequate yearly progress (AYP) at one time point while ignoring the cohort differences. Consequently, VAM finds applications in teacher and school evaluation: it has been explored to determine teacher pay and bonuses (Lissitz, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Sanders, Saxton, & Horn, 1997), support school decisions (Fitz-Gibbon, 1997), and even to certify and promote teachers (Gordon, Kane, & Staiger, 2006).

1.2 Current State of VAM Research

A VAM can be written as a mathematical equation where the left side is students' test score and the right side describes how the test score can be discomposed into different parts such as teacher effect, random error, and relevant covariates such as background and prior scores. Despite a large number of studies on the applications of VAM in school and teacher evaluation, most VAM studies focus on the right side of the equation while the left side receives little attention. As nicely summarized by Briggs and Weeks (2009), those studies focusing on the right side of the equation fall into four main categories: whether teacher effect parameters persist undiminished (McCaffrey et al. 2004); whether covariates at student, teacher, or school level should be included (Ballou, Sanders, and Wright 2004); whether teacher effect estimates should be treated as fixed or random (Harris 2008); whether causal inference can be made in regards to teacher effect estimates (Rubin, Stuart, and Zanutto 2004; Raudenbush 2004).

No doubt the right side of the VAM equation is of vital importance if VAM is expected to produce valid estimates of teacher or school effects. However, no matter how statistically sound the right side is, VAM can only be as good as the test scores used as

the outcome variables in VAM. Koretz (2008) stated that the utility of the teacher effect estimates out of VAMs depends on the left side - the test score quality.

The test scores currently used in the VAM studies are usually scale scores, which are transformations of the scores based on certain measurement models (CTT or IRT). There are potentially two main issues with the test scores. One is the measurement error, which, regardless of the measurement models, is relevant to all VAMs. The other issue is related to the common scale built based on vertical scaling, a psychometric procedure which puts the scores across grades on the same developmental scale. There are two concerns associated with vertical scaling in the VAM context. The first one is that when the assumptions of vertical scaling are satisfied, the teacher effect estimates may still vary depending on the vertical scaling method. McCaffrey et al. (2003) suspected that VAM estimates might be sensitive to different vertical scaling methods. Briggs et al. (2008) showed empirically that such a concern was not unwarranted. The other concern is when vertical scaling is used across multiple grades, the assumptions of unidimensionality and construct invariance are likely to be violated. The violation of construct invariance is especially troubling since the basic idea of VAM is to compute teacher effect based on student growth across grades on a common scale, which should remain constant in order to have a meaningful discussion about growth and change. Numerous researchers have expressed concerns over the potential threats that a changing measure can pose to validity (Bereiter, 1963; Lord, 1963; Angoff, 1971; Bergman, Eklund, & Magnusson, 1991; Williamson, Appelbaum, & Epanchin, 1991; Willett, 1997; Bryk, Thum, Easton, & Luppescu, 1998; Linn, 2001; Thum, 2012). In the context of VAM, tests at different grades are the measures used to quantify student growth, which are used to evaluate

teacher effect. If the tests themselves across grades do not measure the same latent construct and construct equivalence is not achieved, it seems reasonable to question the validity and accuracy of VAM estimates.

Construct shift, a term synonymous with the violation of construct invariance, is often used to describe the change of tests in the construct they intend to measure. In tests commonly used in the K-12 setting, construct shift may be a common phenomenon. Some states, such as Pennsylvania, use criterion-referenced tests that are not designed to construct a common developmental scale across grades. Essentially, the tests are not linked and the tests of higher grades may have very limited overlap with those of lower grades in terms of content coverage. When content coverage differs across adjacent grades, it seems reasonable to suspect that such tests are affected by construct shift.

Even for tests that are designed to be linked to construct a common developmental scale, it is still hard to believe that construct remains equivalent across multiple grades (Hamilton, McCaffrey, & Koretz, 2006; Reckase, 2004; Schmidt, Houang, & McKnight, 2005). Schmidt et al. (2005) demonstrated the high likelihood of construct shift along a vertical scale using the following figure:

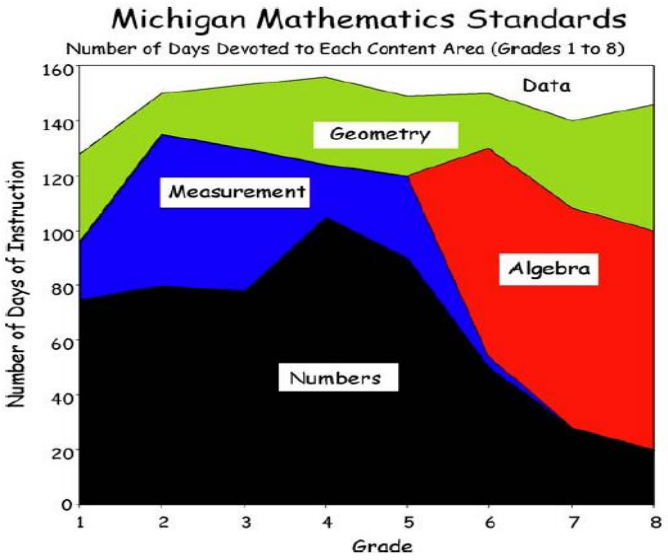

Figure 1.1 Illustration of Math Curriculum Change across Multiple Grades

The first noticeable observation about the above figure is that although math might be the single umbrella term to describe what the tests intend to measure, it does not necessarily suggest unidimensionality and actually consists of different subcontent domains: numbers, measurement, geometry, algebra, and data across different grades. In other words, math becomes a composite of different content strands, which may represent different constructs that require different cognitive processes. For this reason, math might be perceived as multidimensional. Another observation is that from grades 2 through 8, the proportion of different content strands change and there is a major shift in terms of content coverage. Schmidt et al. (2005) drew the conclusion that "math is not math". In other words, math taught at different grades does not contain the same content areas and the content coverage differs as well. This can be well illustrated with the comparison of grade 2 and grade 8 in the above figure: the grade 2 math is composed of data, geometry, measurement, and numbers, while the grade 8 math replaces measurement with algebra.

Such a shift in content coverage is likely to cause construct shift, This example might be considered as a typical scenario of construct shift due to changes in content coverage where some old content (measurement) is dropped out and new content (algebra) is added in the math curriculum, although it should be noted that construct shift may also occur only due to changes in the coverage proportion of different content areas across different grades though the content areas remain the same.

Depending on the subject area, the magnitude of construct shift seems to vary. Skaggs and Lisstz (1988) suggested that reading tests seemed to be more unidimendional across grades and the assumption of construct invariance may reasonably hold. Wang and Jiao (2009) found evidence for the construct invariance across grades of a reading test in a K-12 setting. Reckase and Martineau (2004) suggested that for science tests the assumption of construct invariance across grades is more likely to be violated due to the drastic shift of content. It should be noted that in real settings, the assumption of construct invariance might be violated to varying degrees, depending on the subject area and the number of grade levels involved.

If multidimensionality occurs without construct shift, which means the content mix of the tested subject remains the same, unidimensional vertical scaling is not an option. Wang (1986) showed that when unidimensional IRT models were applied to test items that were multidimensional, the result was the linear composite of those dimensions in the data. Reckase (2004) warned that "projecting the complex data onto a line results in the loss of information and when that linear scale is extended over many grade levels, the loss of information might be extensive." Literature provides some guidance on how to conduct multidimensional vertical scaling when the assumption of construct invariance

holds (Beguin & Hanson, 2001; Beguin, Hanson, & Glas, 2000; Patz & Yao, 2007; Simon, 2008).

When multidimensionality is combined with construct shift, those multidimensional vertical scaling methods are not applicable because they assume construct invariance. This presents a challenge to the use of VAM since many VAM models require that test scores are vertically scaled, and lack of available vertical scaling methods in the case of construct shift leaves researchers with two options: one is to ignore construct shift and proceed assuming that construct shift should not be a problem; the other is to develop some statistical procedures to address construct shift accordingly. Martineau (2006) found that when construct shift is ignored in VAM, teacher estimates may be severely biased. Therefore, the first option is not acceptable.

In terms of the second option, there have been some methodological advances more recently. Mariano et al. (2010) realized that in previous persistence VAM models, the assumption that a teacher's effect in the current year should correlate perfectly with the effect in the later years is overly restrictive in the case of construct shift. To address this problem, they developed a "generalized persistence" (GP) model that relaxes this overly restrictive assumption of perfect correlation and stated that the GP model does not assume vertical scaling and deals with the issue of construct shift. If the GP model is shown to be superior or similar to other VAM models that require vertical scaling in terms of teacher effect estimation, it is certainly attractive since all the pitfalls inherent with vertical scaling upon which the practitioners and researchers might stumble can be circumvented without sacrificing the model performance. To show this, however, a comparison between the GP model and other VAM models that require vertical scaling

has to be conducted. Mariano et al. (2010) directly generated scale scores and imperfectly correlated teacher values to simulation the scenario of construct shift. Without the item level data it is simply impossible to carry out vertical scaling and make such a comparison.

While Mariano et al.'s study focused on the modeling issue in the VAM framework, Li (2012) used the bifactor model to simulate a scenario of construct shift at the item level and proposed a full-information bifactor model used in the vertical scaling process to investigate the recovery of the bifactor model parameters. Comparing to the conventional MIRT models, the bifactor model seems to be a convenient and innovative framework to model growth in the context of construct shift. With the conventional MIRT models, there are a series of difficult questions remaining to be answered when conceptualizing growth with latent dimensions dropping out or entering the latent constructs across grades: If a latent dimension drops out, should it be modeled as part of the growth? If yes, what should its value be in the next grade? If a latent dimension enters due to the introduction of new content area, how should it be modeled? Should its value at previous years be set at zero or negative infinity? Li circumvented those difficult questions by proposing a bifactor model framework in which growth only occurs on the general dimension and all the grade-specific dimensions are irrelevant in regards to growth. Although it may seem unrealistic to assume that growth only occurs on the general dimension but not on the grade-specific dimensions, the bifactor model at least offers a convenient framework to model growth in the context of multidimensionality and construct shift. Li showed that compared to the conventional method of fitting a unidimensional IRT model and thus ignoring construct shift, the bifactor model method

recovers the item parameters and person parameters more accurately. However, she limited her study in the IRT framework and did not apply her proposed method in the VAM context.

1.3 Statement of the Problem

Many testing programs use IRT as the measurement model and consequently test scores used in VAMs are predominantly IRT scale scores. However, there seems to be a disconnection in the literature between studies focusing on IRT and those on VAM. On the IRT side, despite the advances in multidimensional vertical scaling methodology, seldom are those methods applied in the VAM context to investigate their impacts upon teacher effect estimation. On the VAM side, despite the abundance of IRT models and relevant literature, VAM researchers tend to directly generate the scale score at the test level, ignoring that in practice the scale score comes from the item level response data, which are usually assumed to follow a specific IRT model. This practice presents a challenge to investigate the effect of the psychometric properties of test scores upon the parameter estimates in VAM since many psychometric issues are investigated at the item level by utilizing the item level data.

With the bifactor model as the true measurement model driving the item level responses, this study combines the IRT framework and the VAM in a simulation to investigate how different methods impact the teacher effect estimates in VAM when faced with construct shift. Specifically, it compares the performance of the relatively new GP model with the other VAM models that require vertical scaling in scenarios where

construct shift occurs. In addition, the GP model is applied in an empirical data set to evaluate its performance.

CHAPTER 2: LITERATURE REVIEW

This chapter is organized as follows: the first section reviews the definition of teacher effect in VAM and the current VAM models with a focus on the persistence models and ends with a discussion of challenges and issues facing VAM researchers. The second section reviews the current IRT vertical scaling methodologies with an emphasis on the bifactor model method. The last section summarizes the gap in current literature.

2.1 VAM and Teacher Effect

2.1.1 Definition of Teacher Effect

In VAM, the estimate of teacher effect is a measure of a teacher's contribution to student growth and learning. The contribution is often referred as a causal effect in the sense that the contribution made by the teacher causes the growth of the students. McCaffrey et al. (2004) defined teacher effect in this way: "Conceptually, the teacher effect on a student is defined as the difference between the student's achievement after being in the teacher's class compared with his/her achievement in another plausible setting, such as with a teacher of average effectiveness." They also suggested that in order to make inferences about such a causal effect, the following issues have to be clarified.

The first issue is the definition of a plausible alternative with which a student's growth with the current teacher is compared. If other teachers are the plausible alternative, are they in the same school, in the same district, in the same state? Or should they include all teachers teaching similar students? Do we consider a particular teacher or the average

of them? In current practices teacher effects are often estimated at the district level, which means the plausible alternative is the average teacher in the district level.

Which students should be considered? This is the second relevant question when defining teacher effect. Due to the probable existence of teacher student interaction, teacher effects are likely to be non-constant across students. If so, which effects should be considered? If an average of effects is to be considered, which population of students should be averaged? Answers to those questions affect the definition of teacher effect.

Another issue meriting consideration is the confounding of teacher effect and the indirect effect of school or school district that affect students through teachers. Meyer (1997) suggested that it is impossible to isolate the teacher effect from such indirect effects. Whether agreeing with him or not, when one defines teacher effect, an explicit definition is needed in this regard.

The last issue is the possibility that teacher effects vary across time. There is empirical evidence that teacher effects change due to increment of teaching experience, change of class sizes, and other external factors (Shkolniket al., 2002; Rivkin, Hanushek, and Kain, 2000; Kane & Staiger, 2001). Without a constant single teacher effect, it should be explicitly stated in regards to the teacher effect, may it be the current year effect, the average effect of recent years, or the trend in the effect.

2.1.2 Main VAM Models

McCaffrey et al. (2003) classified the main VAM models into the following categories: covariate adjustment models, gain score models, and multivariate models.

2.1.2.1 Covariate Adjustment Models

One unique characteristic of the covariate adjustment models (Diggle, Liang, & Zeger, 1996; Meyer, 1997; Rowan, Correnti, & Miller, 2002) is that test scores appear on the right side of equation. Specifically, prior scores are used as predictors, along with other covariates, to predict the current score. The mathematical equation of covariate adjustment models, as used by Rowan et al. (2002), is as follows:

$$y_{ig} = \mu_g + \beta_g \boldsymbol{X}_i + \gamma y_{ig-1} + \gamma'_{ig} \boldsymbol{Z}_{ig} + \theta_g + \varepsilon_{ig} \qquad (1)$$

In this equation, y_{ig} and y_{ig-1} are student i's test scores at grade g and grade g-1; μ_g is the grade-specific mean; $\boldsymbol{X}_i$ and $\boldsymbol{Z}_{ig}$ are the time invariant and time varying covariates from student I, with β_g and γ'_{ig} being the corresponding vector of coefficients of those covariates; θ_g is the teacher's effect upon student *i*'s score at grade g, which is above and beyond μ_g, and it can be either fixed or random in a normal distribution; ε_{ig} is the residual error term. Both θ_g (when considered as a random effect) and ε_{ig} are assumed to be *i.i.d* normal random variables with a mean of 0. While this model is often used with two years of data, when there are more than two years of data there is an important assumption: the residual errors across years are independent. McCaffrey et al. (2003) stated that "If the model is extended to allow for correlation among the residual errors across years, then standard mixed model estimation would yield biased estimates of fixed effects because of the correlation between the covariate and the residual error term."

One of the main advantages of the covariate adjustment model is its intuitiveness and easy applicability. Another main advantage is that it does not require vertical scaling. Green (2010) concludes that "this is particularly beneficial for school systems using a mixture of norm-referenced and/or criterion-referenced tests, where reported student

scores from the two types of instruments reflect different measurements: either relative academic performance or proficiency on predetermined criteria, respectively." Its main disadvantage is that this model, by only including the test score of the previous year in the equation, fails to take into consideration students test scores in prior years. Another disadvantage is its stringent requirement of data completeness: only students with complete record can be used for model estimation.

2.1.2.2 Gain Score Models

Gain score models can be seen as a special case of covariate adjustment models when the scores have been vertically scaled and thus put on the same development scale (Rowan et al., 2002; Shkolnik, Hikawa, Suttorp, Lockwood, Stecher, & Bohrnstedt, 2002). Its mathematical equation is as follows:

$$d_{ig} = y_{ig} - y_{ig-1} = \mu_g + \beta_g \boldsymbol{X}_i + \gamma'_{ig} \boldsymbol{Z}_{ig} + \theta_g + \varepsilon_{ig} \qquad (2)$$

It is quite straightforward why this model is a special case of the previous one: By setting γ, the coefficient of y_{ig-1} in the covariate adjustment model, to be 1 and moving y_{ig-1} from the right side to the left side of the equation, the gain score model is derived. All the model assumptions remain the same.

It should be worth reiterating that while gain score models seem to be the product of a simple mathematical manipulation of the covariate adjustment model, it has a much more stringent requirement in terms of the psychometric properties of the test scores. Since only the difference of two scores on the same scale makes sense, vertical scaling is a necessity in this model, which, as will be reviewed in the second part of the literature, may bring unintended consequences.

2.1.2.3 Multivariate Models

A main feature that distinguishes the multivariate models from the previous two models is the simultaneous modeling of all student scores. While computationally intensive, multivariate models provide much flexibility such as omission of covariates, ability to utilize records with missing data, consideration of the persistence of teacher effect. There are mainly three kinds of multivariate models: cross-classified models, layered model, and persistence models.

2.1.2.3.1 Cross-Classified Models

Cross-classified models are developed by Raudenbush and Bryk (2002) to model cross-grade correlations and the impact of persisting teacher effects on test scores. The mathematic equation is as follows:

$$y_{i0} = \mu + \mu_i + \emptyset'_{i0}\theta_0 + \varepsilon_{i0}$$

$$y_{i1} = \mu + \gamma + \mu_i + \gamma_i + \emptyset'_{i0}\theta_0 + \emptyset'_{i1}\theta_1 + \varepsilon_{i1}$$

$$y_{i2} = \mu + 2\gamma + \mu_i + 2\gamma_i + \emptyset'_{i0}\theta_0 + \emptyset'_{i1}\theta_1 + \emptyset'_{i2}\theta_2 + \varepsilon_{i2}$$

$$y_{i3} = \mu + 3\gamma + \mu_i + 3\gamma_i + \emptyset'_{i0}\theta_0 + \emptyset'_{i1}\theta_1 + \emptyset'_{i2}\theta_2 + \emptyset'_{i3}\theta_3 + \varepsilon_{i3}$$

In those equations, y_{i0}, y_{i1}, y_{i2}, and y_{i3} are scores for student *i* in grades 0 to 3. Grade 0 here represents the base time point for the growth model, where g is equal to 0. The εs are the residual errors that are assumed to be *i.i.d.* normally distributed with mean of 0. The θs are the teacher effect that is also assumed to be normally distributed with a constant variance across grades. A linear trend $y_{ig} = \mu + g\gamma + \mu_i + g\gamma_i$ is used to model student growth, and similar to hierarchical models, the random intercepts μ_i and slopes γ_i are assumed to be normally distributed with mean of 0 and variance and

covariance terms. It should be noted that in cross-classified models, student scores are decomposed into two parts: the linear growth of students and the teacher effects, and the teacher effects persist undiminished into the future years. Take grade 2 as an example,

$$y_{i2} = (\mu + 2 * \gamma + \mu_i + 2 * \gamma_i) + (\emptyset'_{i0}\theta_0 + \emptyset'_{i1}\theta_1 + \emptyset'_{i2}\theta_2) + \varepsilon_{i2}$$

where the terms in the parenthesis model student *i*'s linear growth, and those in the second parenthesis is the accumulated teacher effect from grade 0, grade 1, and grade 2.

2.1.2.3.2 Layered Model

Sanders et al. (1997) developed the Tennessee Value-Added Assessment System (TVAAS) model to estimate teacher effects. It is also known as layered model "because the model for later years adds layers to the model for earlier years" (McCaffrey, et al., 2004). Its mathematic equations for grade 0 to 3 are as follows:

$$y_{i0} = \mu_0 + \emptyset'_{i0}\theta_0 + \varepsilon_{i0}$$

$$y_{i1} = \mu_1 + \emptyset'_{i0}\theta_0 + \emptyset'_{i1}\theta_1 + \varepsilon_{i1}$$

$$y_{i2} = \mu_2 + \emptyset'_{i0}\theta_0 + \emptyset'_{i1}\theta_1 + \emptyset'_{i2}\theta_2 + \varepsilon_{i2}$$

$$y_{i3} = \mu_3 + \emptyset'_{i0}\theta_0 + \emptyset'_{i1}\theta_1 + \emptyset'_{i2}\theta_2 + \emptyset'_{i3}\theta_3 + \varepsilon_{i3}$$

In those equations, μ_g is the grade-specific mean, while y_{ig}, θ_g, and ε_{ig} remain the same as in the cross-classified models. Similar to the cross-classified models, the εs are assumed to be normally distributed and independent across students, and the θs are assumed to be normally distributed and independent. The variance-covariance matrix of the εs is unrestricted, while across students the variance-covariance parameters are constrained to be the same. In addition, the variances of teacher effects in the layered model are allowed to vary across grades. The main difference between the layered model

and the cross-classified model is that instead of using a linear growth model to represent grade difference, the layered models use the grade specific mean μ_g; therefore, the student *i*'s score can be decomposed into the grade specific mean and the accumulated teacher effects, which is also assumed to persist undiminished into future years.

It should be noted that the above equations are limited to the scenario of a single subject and cohort of students from one system, and they can be extended to model multiple subject and cohorts of students from multiple systems.

2.1.2.3.3 Persistence Models

McCaffrey et al. (2004) summarized the relations among the aforementioned models and the persistence models. Specifically, they showed that if restrictions are applied to the overall time trend and/or the distribution of residual errors, the gain score models and the cross-classified models become special cases of the layered model. With restrictions imposed upon the persistence parameter and without covariates, the layered model becomes a special case of the persistence model. The covariate adjustment models and the gain score models can both be viewed as special cases of the persistence model with restrictions on the distribution of residual errors and the persistence parameters. This section review starts with the "Generalized Persistence" (GP) model (Mariano et al., 2010), the most general persistence model, then reviews other persistence models and shows that they are all special cases of the GP model.

2.1.2.3.3.1 GP Model

In GP models, student scores are calculated based on the sum of teacher effects across years. Since different students may change teachers every year and have different membership in multiple group units, GP models are also referred to as "multiple membership" models (Browne, Goldstein, &Rasbash, 2001; Rasbash & Browne, 2001). Assuming a single cohort of students and a single subject, the mathematical equation of the GP model is as follows:

$$Y_{it} = \mu_t + \sum_{t^* \le t} a_{tt^*} \boldsymbol{l}_{\boldsymbol{t}^*} + \varepsilon_{it}.$$

In this equation, y_{it} is the test score of student *i* in year t, μ_t is the year-specific mean, and $\varepsilon_{it}$is the residual error. $\boldsymbol{l}_{\boldsymbol{t}^*}$ is a vector of teacher effects at year t, and a_{tt^*} is the persistence parameter which is equal to 1 when $t = t^*$, and between the range of 0 to 1 when $t^* \le t$. Here it is assumed that each student has only one teacher each year for the sake of simplicity.

The residual error terms $\boldsymbol{\varepsilon}_i{}' = (\boldsymbol{\varepsilon}_{i1}, \ldots, \boldsymbol{\varepsilon}_{it})$ are assumed to be normally distributed random variables, independent across students. They have a mean of 0 and an unstructured covariance matrix $\sum$:

$$\boldsymbol{\varepsilon}_i \sim MVN(\mathbf{0}, \sum).$$

For each grade t, the current and future effects of the teachers teaching grade t with a K_t - dimensional multivariate normal distribution with mean vector 0 and unstructured covariance matrix $\boldsymbol{\Gamma}_t$:

$$\boldsymbol{l}_{\boldsymbol{t}^*} \sim MVN(\mathbf{0}, \boldsymbol{\Gamma}_t).$$

It is assumed that the vectors of teachers' effects are independent across both teachers at the same grade and teachers at different grades. Moreover, they are independent of the residual errors.

The primary innovation of the GP model is its relaxation of the assumption that the current and future effects of a teacher are perfectly correlated. All the previous persistence models, including the "complete persistence" (CP) model (Raudenbush & Bryk, 2002; Sanders et al., 1997; Harris & Sass, 2006) and the "variable persistence" (VP) model (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Lockwood, McCaffrey, Mariano, & Setodji, 2007), assume a perfect correlation between the current and future effects of a teacher. When construct shift occurs across grades, which means that the tests of different grades are measuring different constructs, this assumption seems overly restrictive and unrealistic, since the current effect of a teacher on a construct in grade g cannot be perfectly correlated to the future effect of him or her on a different construct. Realizing the restrictiveness of the assumption of perfect correlation, Mariano et al. allowed in the GP model the current and future effects of teachers to have an arbitrary covariance structure that is estimated from the data, and they claim that the GP model "is flexible enough to accommodate both teacher effect decay and scale changes, including content shift, across tests from different grades". However, they did not use simulations to empirically compare the GP model with other persistence models when construct shift occurs.

The GP model can be extended to include time invariant and time varying student background variables $\boldsymbol{x_{it}}$:

$$Y_{it} = \mu_t + \boldsymbol{\beta'_t x_{it}} + \sum_{t^* \le t} a_{tt^*} \boldsymbol{l_{t^*}} + \varepsilon_{it}$$

2.1.2.3.3.2 ZP Model, CP Model, and VP Model

The ZP model, CP model, and VP model have the same mathematical equation as the GP model. The "zero persistence" (ZP) model is a special case of the GP model in the sense that $a_{tt^*} = 0, t^* < t$. In other words, the teacher effect does not persist into future years. The CP model is another special case of the GP model since it constrains $a_{tt^*} = 1, t^* < t.$, which means the teacher effect persists undiminished into future years. In the VP model, $0 < a_{tt^*} < 1, t^* < t$, which means that a teacher's future effects are simple rescaling of the proximal year effect. Therefore, the VP model is also a special case of the GP model.

The fact that the GP model is a generalized case of the ZP model, the CP model, and the VP model can also be shown through the different assumptions about the covariance matrix $\mathbf{\Gamma}_t$, which can be decomposed as

$$\mathbf{\Gamma}_t = \boldsymbol{S}_t^{1/2}\mathbf{C}_t\boldsymbol{S}_t^{1/2}.$$

In this equation, $\boldsymbol{S}_t^{1/2}$is a nonnegative diagonal matrix of the variances of grade t teacher effects in each outcome year $t \geq t^*$ and C_t is the nonnegative definite correlation matrix of those effects.

The GP model places no constraints on both the $\mathbf{S}_t$ and $\mathbf{C}_t$**,** which means that the variance of the teacher effects are allowed to vary and the correlation set to be arbitrary. For the ZP model, the CP model, and the VP model, $\mathbf{C}_t$ is constrained to be **J**, a matrix of all 1s indicating the perfect correlation. These three models are different from each other in the sense that the ZP model constrains $\mathbf{S}_t$to be only one parameter – the variance of the proximal teacher effect due to no teacher effect persistence, the CP model constrains the

diagonal parameter of $\mathbf{S}_t$ to be the same, and in the VP model the diagonal parameters of $\mathbf{S}_t$ are just the product of the square of the persistence parameter and the teacher effect in the preceding year.

2.1.2.3.3.3 Estimation of Persistence Models

Following Lockwood, McCaffrey, Mariano, and Setodji (2007), Mariano et al. (2010) also adopted the Bayesain framework (Carlin & Louis, 2000; Gelman, Carlin, Stern, & Rubin, 1995; Gilks, Richardson, & Spiegelhalter, 1996) for the estimation of teacher effects in the GP model and implemented it using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). Specifically, they used independent, minimally informative, natural semiconjugate (Gelman et al., 1995) priors of low precision normal distributions for each μ_t, the grade specific mean parameter and a Wishart distribution for $\sum^{-1}$ with T+1 degrees of freedom. For each g, other than assuming that $\mathbf{\Gamma}_t{}^{-1}$ is distributed Wishart, they showed that using Wishart as the prior for $\mathbf{\Gamma}_t$ with t+1 degrees of freedom is superior since when the variances of the proximal teacher effect and the future teacher effects differ considerably, the correlation parameters are quite sensitive to even the minimally informative Wishart priors for $\mathbf{\Gamma}_t{}^{-1}$. With the choice of the above priors for the unknown parameters, they successfully estimated the GP model using Markov chain Monte Carlo (MCMC) methods in WinBUGS.

Karl, Yang, and Lohr (2012) tackled the estimation issue of multiple membership linear mixed models such as the GP model using the frequentist approach. Specifically, they developed a method to compute maximum likelihood estimates with an EM algorithm. This method takes advantage of matrix sparcity and only inverts a matrix with

dimensions depending on the number of random effects other than on the total number of observations. Comparing to the Bayesian estimation framework, this estimation method produces standard errors that will not be influenced by the choice of priors. They implement this estimation method in the R (R Development Core Team, 2012) package GPvam (Karl et al., 2012).

2.1.2.4 Challenges Facing VAM Researchers

2.1.2.4.1 Non-random Assignment of Students

Under experimental conditions, teacher effect estimation is a much simpler problem since with random assignment of students, the confounding variables are randomly distributed across class and therefore teachers are fairly treated regardless of the classes they teach.

In reality, students are not randomly assigned to different classes. For example, Goldhaber and Anthony (2004) found that successful teachers tend to be able to select their students. If students are sorted based on their prior academic performance, students with similar abilities tend to be in the same class. This non-random assignment of students causes the student performance to be correlated with classrooms, and it is not easy for the current VAM models to separate compositional effects due to students clustering from teacher effects (Murphy, 2012). Rothstein (2009) shows that when students are systematically sorted into classes, the higher test scores cannot be easily attributed to either the teacher effects or the student characteristics. There is belief among the VAM critics that this issue of non-random assignment might not be compensated by the current statistical models, even with prior scores included as the covariates.

2.1.2.4.2 Correct Model Specification

While it is agreed that no model is correct, the bulk of VAM literature investigate the issue of correct model specification. As nicely summarized by Briggs and Weeks (2009), most VAM studies fall into the following categories:

> 1) Should teacher effect parameters be specified such that they persist over time, or should they be allowed to decay (McCaffrey et al. 2004)?
>
> 2) Should student, teacher, or school covariates be included (Ballou, Sanders, and Wright 2004)?
>
> 3) Should value-added effects be modeled as fixed or random (Harris 2008)?
>
> 4) Can value-added estimates be given a causal interpretation (Rubin, Stuart, and Zanutto 2004; Raudenbush 2004)?"

While the fourth point seems to be related to the issue of non-random assignment, the other three points are all relevant to the issues of correct model specification. Several studies (Briggs & Domingue, 2011; McCaffrey et al., 2004; Tekwe et al., 2004) show that different model specifications lead to different teacher effect estimates. Another even thornier issue related to model specification is the inadvertent omission of causative variables, which may cause biased estimates (Hibpshman, 2004). Although McCaffrey et al. (2003) note that if the VAM models are reasonably robust to the omission of variables, as long as they are randomly distributed. This precondition, however, is not realistic considering the non-random assignment of students. This is an even bigger problem because "determining whether causative variables are omitted in practice is impossible" (Murphy, 2012).

2.1.2.5 Use of VAM in Policy and Practice

Due to the aforementioned challenges facing VAM, it is not recommended to be used for high – stake decisions, although there is no evidence that VAM would be more detrimental than the currently employed methods for accountability purposes (McCaffrey et al., 2003). For low-stake and diagnostic purposes, VAM are useful in the sense that it may help identify the most and the least effect teachers, which can be used by administrators as starting points for thorough review. When the stakes increased, VAM should not be used alone to make inferences; instead, it should be used as one of the multiple indicators to inform decision makers. Nevertheless, in stake-attached settings VAM should be used in an extremely cautious manner. McCaffrey et al. (2003) suggested that sensitivity analyses should be conducted to investigate the effect upon students, teachers, and schools.

2.2 IRT Vertical Scaling

Kolen and Brennan (2004) provide the following definition of vertical scaling:

> To measure student growth, performance on each of the test level is related to a single score scale. The process used for associating performance on each test level to a single score scale is vertical scaling and the resulting scale is a developmental score scale.

This definition can be further elaborated as follows. First, it explicitly states that the purpose of vertical scaling is to measure student growth, although they did not define the seemingly vague term “growth”. Second, the product of vertical scaling is a

developmental score scale called vertical scale, which is assumed to exist. Similarly, Lissitz and Huynh (2003) define vertical scale as "a single (unidimensional) scale that summarizes the achievement of students".

Despite the seemingly simple definition, vertical scaling is complicated. As stated by Kolen and Brennan (2004), "…vertical scaling is a very complex process that is affected by many factors. These factors likely interact with one another to produce characteristics of a particular scale. The research record provides little guidance as to what methods and procedures work best for vertical scaling." What is worse is that "research does not provide a definitive answer concerning the characteristics of growth on educational tests."

While there are traditional and IRT vertical scaling methods, this review mainly focuses on the latter. IRT is a commonly used framework in today's large-scale testing and also the measurement model used in this study. The main factors that affect IRT vertical scaling usually include vertical scaling designs, choice of IRT models, different calibration methods, and different scoring methods.

2.2.1 Vertical Scaling Designs

According to Kolen and Brennan (2004), there are three basic data collection designs in vertical scaling: common item design, equivalent groups design, and scaling test design. In the common item design, examinees of different grades are administered the test forms of a corresponding level. Since different groups of examinees are from different grades and considered non-equivalent, a common set of items to the two

adjacent grades, are placed on the same positions of different forms and presented to examinees of two adjacent grades. Those common items are later used to link those two forms based on the common-item linking procedure (Kolen & Brennan, 2004). In the equivalent groups design, test forms corresponding to the grade or one level below are randomly assigned to examinees. Since those groups of examinees are considered equivalent, different forms can be linked through the random equating design. In the scaling test design, a test covering the content across all of the grade levels is administered to students of all grades, who also take the test corresponding to their own level. The scaling test is used to link different test forms of each grade level.

2.2.2 Choice of an IRT Model

2.2.2.1 Unidimensional IRT Model

In a two-parameter logistic (2PL) UIRT model, the probability of answering a dichotomous item is

$$P(X_i = 1|\theta_j, a_i, b_i) = \frac{1}{1 + \exp[-(a_i(\theta_j - b_i))]}$$

where θ_j is examinee *j*'s latent ability, a_i is the discrimination parameter of item *i*, and b_i is the difficulty parameter of item *i*. If a_i is constrained to be equal across items, the 2PL UIRT model becomes a 1PL UIRT model. If a guessing parameter is incorporated into the above equation, the 2PL UIRT model becomes a 3PL UIRT model which has the following equation:

$$P(X_i = 1|\theta_j, a_i, b_i) = c_i + \frac{1 - c_i}{1 + \exp[-(a_i(\theta_j - b_i))]}$$

where c_i is the guessing parameter of item *i* and the other terms remain the same as in the 2PL UIRT model. These UIRT models are unidimensional in the sense that only one latent trait is measured.

2.2.2.2 Multidimensional IRT Model

By assuming the existence of a vector of latent abilities, Reckase (1985) extended the 2PL UIRT model to a 2PL multidimensional IRT (MIRT) model which models the probability of a correct item response as

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{i1}\theta_1 + a_{i2}\theta_2 + \cdots + a_{in}\theta_n + d_i)]}$$

where $\boldsymbol{\theta}_j$ is a vector of latent abilities of examinee *j*, $\boldsymbol{a}_i$ is a vector of discrimination parameters of item *i*, and d_i is a scalar of item *i*, which can be calculated using the following formula:

$$d_i = -b_i\sqrt{a_{i1}^2 + a_{i2}^2 + \cdots + a_{in}^2}$$

Here b_i is the difficulty of item *i*. It should be noted that regardless of the number of latent abilities in MIRT models, there is always one difficulty parameter with a corresponding number of discrimination parameters. Another point that should be noted is that MIRT models do not constraint the latent abilities to be independent, and the correlations among latent abilities are freely estimated.

2.2.2.3 Bifactor Model: A Special Case of MIRT Model

Gibbons and Hedeker (1992) derived a factor model for dichotomous item response data based on the work of Holzinger and Swineford (1937). In the bi-factor model the probability of answering a dichotomous item correctly can be modeled as

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}\theta_g + a_{is}\theta_s + d_i)]}$$

where θ_g is the general ability and a_{ig} is the discrimination parameter of item *i* for this general ability, θ_s is the grade specific ability and a_{is} is the discrimination parameter of item *i* for this grade specific ability, $d_i$is a scalar parameter related to the overall item difficulty parameter. Orthogonality between the general ability and any grade specific ability is assumed. The main computational advantage of the bifactor model is that regardless of the overall number of dimensions involved, each item only loads on two dimensions, which makes it similar to a two-dimension factor model in terms of the computational intensiveness.

Based on the above discussion, the bifactor model can be conceptualized as a MIRT model which place constraints on the number of latent dimensions each item loads on and the relation among the latent dimensions.

2.2.2.4 IRT Models in Vertical Scaling

The validity of IRT related vertical scaling methods is based on the satisfaction of respective IRT model assumptions, such as dimensionality, local independence, and model fit. The unidimensional IRT model seems to fit nicely in the assumption of vertical scaling of the existence of a one-dimension scale, which might explain the dominant

choice of unidimensional IRT model in vertical scaling. However, in reality tests are often multidimensional, and relatively much fewer studies chose multidimensional IRT models (Beguin & Hanson, 2001; Beguin, Hanson, & Glas, 2000; Patz & Yao, 2007; Simon, 2008). Li (2012) proposed to use a bifactor IRT model to deal with construct shift that often occurs with the change of content across multiple grades.

2.2.3 Different Calibration Methods

Concurrent calibration requires just one computer run to establish a common scale and simultaneously estimate parameters for all items at all the grades involved. Separate calibration, however, requires one computer run for each grade and links the scales produced within each grade through the common items between adjacent grades. Common IRT linking method include the Mean-Mean (MM) method (Loyd & Hoover, 1980), the Mean-Sigma (MS) method (Marco, 1977), the Stocking and Lord test characteristic curve method (TCC; Stocking & Lord, 1983). Another calibration method, which is a hybrid of the previous two, requires a concurrent calibration between the non-overlapping adjacent grades first and then uses one of the linking methods employed in separate calibration to create the common scale.

2.2.4 Different Scoring Methods

Scoring refers to estimation of examinee proficiency. A decision has to be made concerning whether scoring of examinees' latent ability is based on number correct scoring or pattern scoring. If pattern scoring is used, it has to be decided whether to use

maximum likelihood or Bayesian estimation (EAP, MAP). Based on Kolen & Brennan (2004), "the decision of how to estimate examinee proficiency can have a significant effect on the properties of the resulting scale scores".

2.2.5 The Precarious Nature of Vertical Scaling

Numerous studies using either real data sets or simulations (Harris & Hoover, 1987; Holmes, 1982; Loyd & Hoover, 1980; Marco, Petersen, & Stewart, 1983; Slinde & Linn, 1978, 1979; Kolen, 1981; Marco et al., 1983; Skaggs & Lissitz, 1986a; Sykes & Yen, 2000; Tong, 2005; Harris, 1991; Tong, 2005; Camilli, Yamamoto, & Wang, 1993; Pomplun, Omar, & Custer, 2004; Williams, Pommerich, & Thissen 1998; Chin, Kim, & Nering, 2006; Karkee et al., 2003; Karkee, Wang, & Green, 2006; Meng, Kolen, & Lohman, 2006; Tong & Kolen, 2006; Yen, 1985; Hendrickson, Cao, Chae, & Li, 2006; Kim, Lee, & Kim, 2008) show that different combinations of those aforementioned factors, as well as other factors such as choice of software, have impact upon the resulting vertical scale.. Kolen and Brennan (2004) list the most influential factors as the following: the design for data collection; the complexity (dimensionality) of the subject matter area; the curriculum dependence of the subject matter area; test characteristics, including average item difficulty and discriminations, and relationships of the item characteristics to group proficiency; item types, such as multiple-choice and constructed response items; grade levels; nonlinear scale transformation following implementation of a scaling method.

Briggs (2008) further investigated these issues in the value-added model framework. His study results show that with different combination of those aforementioned vertical scaling procedures, very different conclusions can be drawn concerning estimates of

teacher effects and school effects. He concluded his study by saying "…the variability of score along a vertical scale is very sensitive to the way the scale has been created. This can be problematic when change along the scale is given an absolute or criterion-based interpretation. Hence it would seem that state considering the application of growth to standard models should be especially cognizant of psychometric decisions being made in establishing their vertical scales. These seemingly esoteric decisions appear to have potentially substantial impact on students and schools." However, he only considered the unidimensional IRT models.

2.2.6 Vertical Scaling with Multidimensionality

Li (2012) noticed that there is an unbalance in the literature in terms of investigating the violation of the two assumptions of vertical scaling: unidimensionality and construct invariance. A bulk of the literature, including the cited studies in the previous paragraphs, focuses on the ideal situation when both assumptions are assumed to be met. Only a few studies (Beguin & Hanson, 2001; Beguin, Hanson, & Glas, 2000; Patz & Yao, 2007; Simon, 2008) investigate vertical scaling methods under the scenario of multidensionality and construct invariance, and even fewer studies to investigate vertical scaling methods under the scenario of multidmeneisonality and construct shift (Li, 2012). Specifically, Li uses a bifactor model (Gibbons & Hedeker, 1992) that models the general dimension across grades and treats the noise dimension as grade specific to model construct shift, and the magnitude of construct shift is represented by the variance of the grade specific dimensions. Combined with common item design and concurrent calibration, Li explored the use of a bifactor model and compared that to the use of

unidimensional IRT model in vertical scaling with different magnitudes of construct shift. Under the simulation conditions investigated in her study, she drew the following conclusions:

1. The item parameters of the bifactor model were well recovered, but those of the general dimension were recovered better than the grade specific dimensions.
2. With unidimenisonal IRT model, item discrimination parameters were overestimated.
3. The bifactor model recovered the person parameters better.
4. Group mean estimates for the bifactor model were better.

2.3 Gap in Current Literature and Research Questions

2.3.1 Gap in Current Literature

As mentioned in chapter 1, literature on VAM seems to be lacking in regards to the impact of psychometric properties of the test scores upon teacher effect estimates. One of the reasons for the lack of attention may be because the current simulation studies used in the VAM context usually directly generated the scale scores and applied them in different VAMs without taking into consideration that the test scores used to evaluate teachers in practice come from the item level response data. Direct generation of the scale scores causes the difficulty to create scenarios with different psychometric problems, and as a result, the impact of different psychometric issues of the scores upon teacher effect estimate in VAM remains untapped. This study intends to fill the gap, combining IRT

and VAM in data simulation by generating item level response data with IRT models while using VAM to estimate the teacher effect based on the IRT scale scores coming out of scoring the item level response data. The main advantage of this combination is since the data are generated at the item level, the psychometric properties of the test scores can be easily manipulated.

Before Li's study (2012), there were no studies that provided advice on how to generate item level response data with construct shift, not to mention how to create vertically scaled scores when faced with construct shift. Li's bifactor model vertical scaling method (2012) not only provides an innovative vertical scaling method but also offers a convenient framework to generate multidimensional data of construct shift. However, this method has never been used in the VAM context before and it remains unknown how the superiority of Li's bifactor model vertical scaling method to the traditional unidimensional model vertical method is transferred to VAMs and consequently impact the teacher effect estimates.

The GP model (Mariano et al., 2010) is claimed to be able to circumvent the issue of vertical scaling when faced with construct shift by relaxing the assumption of perfect correlation of the current and persisting future effects of teachers and hence requires no vertical scaling. They simulated the scenario of construct shift by generating a set of imperfectly correlated values that represent the proximal and future teacher effects. If the GP model can estimate teacher effects accurately, it should be superior to the other persistence models requiring vertical scaling due to its the precarious nature. However, they did not use any IRT models to generate data at the item level; therefore, it is

impossible to compare the performance of the GP model with other persistence models that require vertically scaled scores and hence item level data.

In addition, they used an empirical data set to empirically show that the relaxation of the assumption of perfect correlation is reasonable and the GP model provides the best model fit among all persistence models. They also found that the teacher effect estimates between the GP model and the VP model were extremely highly correlated. As a result, they concluded that for that specific data set they used, choosing the VP model over the theoretically superior GP model might not be problematic in terms of the proximal teacher effect estimates. However, they stated that "the results here are based on a single data set of assessment scores purported to be developmentally scaled. It is important for future work to carry out similar investigations with other data sets, particularly those with tests that are not on a vertical scale, to understand how generalizable our findings may be."

2.3.2 Research Questions

Based on the above summary, this study focuses on evaluation of the GP model and intends to answer the following research questions:

1. Does the GP model provide accurate teacher effect estimates without vertical scaling when data is generated from the bifactor IRT model to simulate the scenario of construct shift?
2. Which method performs better: the bifactor model vertical scaling method combined with VAM models that assume vertical scaling or the GP model that does not assume vertical scaling?

3. With tests that are not vertically scaled, are Mariano et al.'s findings generalizable?
4. Does the bifactor model yield a more accurate teacher effect estimate than the unidimensional model in the case of construct shift?

These four questions are addressed in the following chapters. Chapter 3 describes the simulation designed to answer question 1, question 2, and question 4, and chapter 4 presents the results of the simulation study. Chapter 5 focuses on question 3, fitting the GP model in an empirical dataset where scores are not vertically scaled. This dissertation ends with a discussion of the main findings and the suggestions for future research.

CHAPTER 3: METHODOLOGY

Chapter II reviewed the background relevant to the current study, the objectives of which were to investigate the performance of the GP model with test scores of construct shift, the performance of the bifactor model vertical scaling method in the VAM context, and to compare both to the current method of ignoring the construct shift and using the unidimensional vertical scaling method, which served as the baseline model. To meet these research objectives, a simulation study was designed to answer three of the four research questions listed at the end of chapter II:

1) Does the GP model provide accurate teacher effect estimate without vertical scaling when data are generated from the bifactor IRT model to simulate the scenario of construct shift?

2) Which method performs better: the bifactor model vertical scaling method combined with VAM models that assume vertical scaling or the GP model that does not assume vertical scaling?

4) Does the bifactor model vertical scaling method yield more accurate teacher effect estimate than the unidimensional model vertical scaling method in the case of construct shift?

Part I of this chapter provides a description of the simulation study used to answer the above questions, and Part II describes the empirical data set that is used to answer the third research question listed in chapter 2:

3) With tests that are not vertically scaled, are Mariano et al.'s findings generalizable?

3.1 Fixed Factors of the Simulation Study

There are six fixed factors in this simulation study: sample size, generating IRT model, common item design, test length, number of common items and concurrent calibration.

Sample size is fixed at 1,000. In VAM simulation studies, the class size is often set at 25. Assuming there are 40 classes in a school district, the sample size is therefore 40*25=1,000. It is also assumed that there are 40 comparable teachers, each teaching one of the classes.

The bifactor model is the measurement model used to generate the item level data. As mentioned in the previous chapters, it is a convenient framework to simulate construct shift and is also the only multidimensional IRT model in which a vertical scaling method exists to deal with construct shift. The general factor that is common to different grades is assumed to capture teacher effects, and it is assumed that teacher effects are irrelevant to the grade specific factors.

Common item design is used in the vertical scaling process. Relatively easy to implement, common item design is the most commonly used data design approach in commercial and state testing programs.

The test length is fixed at 60 items, and the number of common items is 18.

Those 2 factors are fixed because the focus was not on vertical scaling technique per se but on the teacher effect estimation. Kolen and Brennan (2004) suggested that the minimum number of common items should be 20%, and based on this suggestion Li used 20%, 30% and 40%. 30% was chosen to be the percentage of common items in this study.

The concurrent calibration is preferred over the separate calibration due to reasons such as avoiding linking errors and the use of a larger sample size.(Kolen & Brennan, 2004; Simon, 2008). Li also used concurrent calibration and she justified her choice by stating that the main focus of her study is to investigate whether the bifactor model vertical scaling method improves parameter estimation when the bifactor model is the correct model. According to Kolen and Brennan (2004), concurrent calibration is better than separate calibration when the model is correctly specified. Although Li's rationale for her choice does not apply in the current study, the concurrent calibration is also chosen here to be consistent with her study to more easily translate into the teacher effect estimates in VAM in this study.

3.2 Manipulated Factors of the Simulation Study

Table 1 lists the details of the proposed simulation study. As can be seen from the table, there are a total of 27 manipulated conditions, each of which has 100 replications. Specifically, the 3 levels of teacher effect persistence patterns are ZP, CP, and VP with the persistence parameter between two adjacent years set to be 0.5. This value is chosen to represent a medium persistence effect, to be distinguished from 0 and 1, which are the values of the persistence parameter in ZP and CP. The magnitude of the construct shift is represented with the variances of the grade specific dimensions, and the 3 levels of variance are chosen to be 0.25, 0.5, and 1 to represent small, medium, and large construct shift. The 3 vertical scaling methods are the traditional unidimensional IRT vertical scaling method, the bifactor model vertical scaling method, and the use of GP model that assumes no vertical scaling. In Table 3.1 each of the simulation conditions is described in detail.

Table 3.1 Simulation Manipulated Conditions

Conditions	Number of Levels
Teacher Effect Persistence Patterns	3
Magnitude of Construct Shift	3
Vertical Scaling Method	3
Total	27

3.2.1 Teacher Effect Persistence Patterns

The mathematical equation of the VP model is as follows:

$$Y_{it} = \mu_t + \boldsymbol{\beta'_t x_{it}} + \sum_{t^* \le t} a_{tt^*} \boldsymbol{l_{t^*}} + \varepsilon_{it}$$

where Y_{it} is the test score for student *i* in year *t*, and μ_t is the year specific mean, $\boldsymbol{x_{it}}$ is a covariate vector containing both time variant and varying background variables, and $\boldsymbol{l_{t^*}}$ is a vector of teacher effects at year t^*. If we ignore the covariates, the (VP) model can be simplified to:

$$Y_{it} = \mu_t + \sum_{t^* \le t} a_{tt^*} \boldsymbol{l_{t^*}} + \varepsilon_{it}$$

The value of a_{tt^*} changes with the change of *t*: when $t^* = t$, $a_{tt^*} = 1$, indicating the current teacher effect; when $t^* < t$ *t*, $a_{tt^*} < 1$, indicating the diminishing effect of a prior teacher. For example, if a student's current grade mean score is *s*, the teacher who taught him 2 years ago had an effect of l_{-2} then and the persisting effect now is $0.25 * l_{-2}$ ($a_{t(t-2)} = 0.25$), the teacher who taught him 1 years ago had an effect of l_{-1} then and the persisting effect now is $0.5 * l_{-1}$ ($a_{t(t-1)} = 0.5$), and the current teacher has an effect

of l_0, then the student's score at the end of this year would be s+0.25*l_{-2}+0.5*l_{-1}+l_0 plus some random error.

The ZP model is a special case of VP because a_{tt^*} is constrained to be 0, and the current teacher effect will not persist into future years. For example, if a student's current grade mean score is *s*, the teacher who taught him 2 years ago had an effect of l_{-2} and the teacher who taught him 1 years ago had an effect of l_{-1}, and the current teacher has an effect of l_0, then the student's score at the end of this year would be s+l_0 plus some random error.

The CP model is another special case of VP because a_{tt^*} is constrained to be 1, which means the current teacher effect will persist undiminished into future years. For example, if a student's current grade mean score is *s*, the teacher who taught him 2 years ago had an effect of l_{-2} and the teacher who taught him 1 years ago had an effect of l_{-1}, and the current teacher has an effect of l_0, then the student's score at the end of this year would be s+l_{-2}+l_{-1}+l_0 plus some random error.

Therefore, the above three models can be seen as the VP model where the persistence parameter is constrained to be 3 values: 0.5, 0, and 1. These three models are adapted to model student growth patterns due to different teacher effect persistence in this study. It should be noted that in the original models the teacher effects are modeled using the scale score units, while in this study the teacher effects are directly placed on the latent ability scale of the general dimension in the bifactor model. In the remaining part of this dissertation, CP, VP, and ZP refer to the persistence pattern being manipulated in the data generation process, in order to be differentiated from the CP model, the VP model, and the ZP model.

Specifically, the general dimension θ_g represents student common ability across grades, the first grade-specific dimension θ_3 represents grade 3 specific ability at the end of grade 3, the second grade-specific dimension θ_4 represents grade 4 specific ability at the end of grade 4, and the last grade-specific dimension θ_5 represents grade 5 specific ability at the end of grade 5. The persistence parameter a_{21} is set equal to 0.5, a_{31} equal to 0.25, and a_{32} equal to 0.5 with VP. In other words, the teacher effect is set to persist at

a decrease rate of 0.5, which is considered a medium persistence effect. Table 3.2 lists the student ability at the end of each grade with these three persistence patterns, where l_3, l_4, and l_5 are teacher effect at grade 3, 4 and 5.

Table 3.2 Student Ability on the General Dimension at Each Grade

Models	Grade 3	Grade 4	Grade 5
VP	$\theta_g + l_3$	$\theta_g + 1.5 * l_3 + l_4$	$\theta_g + 1.75 * l_3 + 1.5 * l_4 + l_5$
ZP	$\theta_g + l_3$	$\theta_g + l_3 + l_4$	$\theta_g + l_3 + l_4 + l_5$
CP	$\theta_g + l_3$	$\theta_g + 2l_3 + l_4$	$\theta_g + 3 * l_3 + 2 * l_4 + l_5$

The grade specific dimensions were assumed not to be affected by teacher effect. The bifactor IRT model combined these values on the general dimension with their corresponding values on the grade specific dimensions to generate the item level data.

3.2.2 Magnitude of Construct Shift

To represent the magnitude of the testlet effect, Li, Bolt, and Fu (2006) and Rijmen (2010) manipulated the variances of the testlet factors in the testlet model, which is a constrained version of the bifactor model. Inspired by their studies, Li (2012) manipulated the variances of the grade specific dimensions to represent the magnitude of construct shift. Specifically, she used 0.25, 0.5, and 1 as the variance values to represent small, medium, and large magnitude of construct shift. Those values were also used in this simulation study.

3.2.3 Vertical Scaling Method

With the traditional unidimensional IRT vertical scaling method, multidimensionality and construct shift are ignored and this condition represents a scenario where the measurement model is misspecified: the generating model is a bifactor model while the estimating model is a UIRT model. This condition can be considered analogous to the common practice of applying UIRT vertical scaling across grades while it is believed the test scores at different grades represent different multidimensional constructs. Using a common item design, this method uses a 2PL IRT model for the concurrent calibration. The calibrated item parameters are used to score the students, and student scores are then used in the VAM model same as the generating VAM model for teacher effect estimates. For example, if the ZP model was used to generate the persistence pattern on the general dimension, the ZP model was used to estimate teacher effects. The estimation VAM model and generating VAM model were the same in order to avoid confounding effects of VAM model misspecification and ignorance of construct shift – measurement model misspecification. Since this method involves concurrent calibration with a unidimensional IRT model, it will be abbreviated as the CU (concurrent and unidimensional) method in the remaining part.

With the GP model method, since no vertical scaling is required, the item parameters for each grade were separately calibrated using a unidimensional 2PL IRT model. These separately calibrated item parameters were directly used to score students. Without any vertical scaling procedure, these student scores were used in the GP model to estimate teacher effects. This method has a less stringent assumption than the previous one in that it does not assume the constructs in different grades remain invariant,

although it does misspecify the measurement model by assuming unidimensionality of test scores. The most attractive feature of the GP model is its ability to recover the teacher effect even in the face of construct shift, which is the focus of this study. Since this method uses separate calibrations for each grade, it will be abbreviated as the SU (separate and unidimensional) method in the remaining part.

With the bifactor model vertical scaling method, the item parameters were concurrently calibrated using a multi-group bifactor model; it will be abbreviated as the CB (concurrent and bifactor) method in the remaining part. The calibrated item parameters were then used to score students, and student scores on the general dimension were used in the VAM model same as the generating VAM model for teacher effect estimates. It is expected that since the CB method does not involve model misspecification, it should be superior to the CU method. A more interesting comparison is between the SU method and the CB method: the CB method does not involve model misspecification but introduces error with vertical scaling, while the SU method does not introduce vertical-scaling-related error but misspecifies the measurement model.

3.3 Data Generation

Item response data are generated based on the bifactor model, with the growth on the general factor based on different persistence models. The generation of the ability parameters, the item parameters, and the item response data are discussed in the following section.

3.3.1 Ability Parameter Generation

It should be noted that two sets of students are generated. The first set is only used for the item parameter calibration and not used for the teacher effect estimate. Assuming there are 3 independent cohorts of students at grade 3, 4, and 5, with each cohort having 2,000 students. Their latent abilities are generated with four-dimensional multivariate normal distributions. Specifically, both the general and the grade-specific dimensions have a fixed standard deviation of 1 in all grades; grades 3, 4, and 5 have respective means of -1.1, -0.6, and 0 on the general dimension and a fixed mean of 0 on each grade specific dimension. Table 3.3 summarizes the generating scheme for calibration item response data.

Table 3.3 Latent Trait Parameter Generation Used for Calibration

Grade Level	General Factor	Grade Specific Factor
Grade 3	N(-1.1, 1)	N(0,1)
Grade 4	N(-0.6,1)	N(0,1)
Grade 5	N(0,1)	N(0,1)

The reason why the first set of students is used for calibration is twofold. First, it is more common in practice to have readily available calibrated item parameters to score the students. Second, the second set of students is a cohort of students who progress from grade 3 to grade 5, and their ability values at each grade are correlated, it will violate

local independence assumption of IRT if they are used in the CU method and CB method when concurrent calibration is involved.

For the second set of students, this study assumed that there were 40 classes taught by 40 different teachers with each class size equal to 25 students at the beginning of the grade 3, at the beginning of grade 4 those 1,000 students were regrouped into 40 different classes of equal size taught by a different set of 40 different teachers, and at the beginning of grade 5 those students were regrouped again and taught by another 40 different teachers. To sum up, 120 teachers and 1,000 students were generated. It was also assumed that these students did not transfer to other schools so missing data was not an issue. It is acknowledged that this is an ideal scenario which might not be easily achieved in practice, but it will provide us with baseline results that can guide further research.

At the beginning of each year, these students changed classes but the class size remained the same. Consequently, students may have different teachers and teachers teach different classes for different years. Only grades 3, 4, and 5 were considered.

The general ability at the end of grade 2 was generated to be a normal random variable with mean of -1.6 and standard deviation of 0.2. Those values were chosen so that regardless of the persistence pattern, the general ability at grade 5 would still stay in (-3, 3), the common range of the latent variable scale of IRT. Depending on which persistence model used, different pattern of teacher effects are added to the grade 2 general ability to generate general abilities for grades 3, 4, and 5. When generating the teacher effects, two issues were taken into consideration. One is that the teacher effects

should not be negative on the latent variable scale of IRT. The other is the same concern that the general ability at grade 5 should remain in a reasonable range. Based on these considerations, the teacher effects for each grade are generated to be normal random variables with mean of 0.4 and standard deviation of 0.2. With the generated ability values at the beginning of grade 3 and the generated teacher effects, the general ability values at the end of grade 3, grade 4, and grade 5 can be calculated based on the formulas summarized in Table 3.2.

Generation of the grade specific ability values is straightforward since the teacher effects are assumed not to affect those dimensions. They are all generated with a standard normal distribution. The generation scheme is summarized in the following table. It should be noted again that the general ability values for grade 3, grade 4, and grade 5 are not generated from independent distributions. Instead, they were calculated based on the general factor in grade 2 and the teacher effects across three years. Table 3.4 summarizes the true model parameters for generating item response data for simulating growth from grades 3 to 5.

Table 3.4 Latent Trait Parameter Generation

Grade Level	General Factor	Grade Specific Factor
Grade 2	N(-1.6,0.2)	
Grade 3		N(0,1)
Grade 4		N(0,1)
Grade 5		N(0,1)

3.3.2 Item Parameter Generation

The generating scheme for the discrimination parameter *a* is the same as in Li's study (2012) for consistency. Specifically, the *a*s for the general factor are set to be 1.2, 1.4, 1.6, 1.8, 2.0 and 2.2 to represent moderate to well discriminating items. The *a*s for the grade specific factors are fixed to be 1.7, which is the mean of 1.2, 1.4, 1.6, 1.8, 2.0 and 2.2.

The difficulty parameter *b* is generated to be a normal random variable. Specifically, for the non-common items in grade 3, *b* is generated from a normal distribution with mean of -1.1 and standard deviation of 0.4. For the non-common items in grade 4, *b* is generated from a normal distribution with mean of -0.6 and standard deviation of 0.4. For the non-common items in grade 5, *b* is generated from a normal distribution with mean of 0 and standard deviation of 0.4. Those values are chosen because they match with the average values of the means and standard deviations of the general factors across the three grades.

For the common items, their difficulty parameter *b*s should be suitable to the two adjacent grades. In other words, those items should not be too easy to the higher grade or too difficult to the lower grade. For the common items between grade 3 and grade 4, the *b*s are generated from a uniform distribution ranging from -1.6 to -0.1. For the common items between grade 4 and grade 5, the *b*s are generated from a uniform distribution ranging from -1.1 to 0.5.

With the *a*s and the *b*s generated, the scalar parameter d can be computed using the following formula:

$$d_i = -b_i\sqrt{a_{i0}^2 + a_{ij}^2}$$

where a_{i0} is the discrimination parameter for the general factor, and a_{i0} is the discrimination parameter for the grade specific factor of grade j. Table 3.5 summarizes the generation of the difficulty parameters.

Table 3.5 Item Difficulty Parameter Generation

Items	Generating Distributions		
	Grade 3	Grade 4	Grade 5
Non-common Items	N(-1.1,0.4)	N(-0.6,0.4)	N(0,0.4)
Common Items	U(-1.6,-0.1)		
Common Items		U(-1.1,0.5)	

3.3.3 Item Response Data Generation

A student's probability of a correct response to an item from grades 3 to 5 can be specified as the following:

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}\theta_{g3} + a_{i3}\theta_3 + d_i)]}$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}\theta_{g4} + a_{i4}\theta_4 + d_i)]}$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}\theta_{g5} + a_{i5}\theta_5 + d_i)]}$$

where θ_{g3}, θ_{g4}, and θ_{g5} represent the general ability at grade 3,4 and 5 and θ_3, θ_4, and θ_5 represent the grade specific ability for grade 3, 4, and 5 respectively. Depending on which teacher effect persistence model, the above equations are different in terms of θ_g for grade 4 and grade 5. In other words, student general ability is simulated based on the equations presents in Table 3.2 for each true model.

Specifically, assuming the general ability at the end of grade 2 is θ_g, the students' probability of a correct response to the math tests from grade 3 to 5 for VP persistence pattern, can be specified as the following:

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}(\theta_g + l_3) + a_{i3}\theta_3 + d_i)]}$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}(\theta_g + 1.5 * l_3 + l_4) + a_{i4}\theta_4 + d_i)]}$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}(\theta_g + 1.75 * l_3 \ + 1.5 * l_4 + l_5) + a_{i3}\theta_5 + d_i)]}$$

For ZP persistence pattern:

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}(\theta_g + l_3) + a_{i3}\theta_3 + d_i)]}$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}(\theta_g + l_3 \ + l_4) + a_{i4}\theta_4 + d_i)]}$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}(\theta_g + l_3 \ + l_4 + l_5) + a_{i3}\theta_5 + d_i)]}$$

For CP persistence pattern:

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}(\theta_g + l_3) + a_{i3}\theta_3 + d_i)]}$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}(\theta_g + 2l_3 + l_4) + a_{i4}\theta_4 + d_i)]}$$

$$P(X_i = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_i) = \frac{1}{1 + \exp[-(a_{ig}(\theta_g + 3 * l_3 + 2 * l_4 + l_5) + a_{i3}\theta_5 + d_i)]}$$

It should be noted that for grade 3, the probability of a correct response remains the same regardless of the persistence models since it does not involve persisting teacher effect. With the generated student abilities, teacher effects, and item parameters, the item response data can be generated for each examine based on the above equations.

3.4 Identification of the Bifactor Model

In order to keep the bifactor model identified, either the variance or the discrimination parameter of the latent variable has to been fixed. Li (2012) fixed the variance of the general factor to be 1 and the discrimination parameters of the three grade specific factors to be 1.7 so that the variance of the grade specific dimensions could be freely estimated as a measure of the magnitude of construct shift. In this study, it should not matter which constraint is used since the interest is not the estimation of the magnitude of construct shift but its impact upon teacher effect estimation. To be consistent, however, the same constraints are used.

3.5 Calibration

The item response data based on the first set of generated ability values are used for calibration. As discussed previously, three calibration methods are explored, the CU method, the CB method, and the SU method.

With the CU method, multidimensionality and construct shift were ignored and all the items were assumed to load on the only factor in the data calibration process. With common item design, this method concurrently calibrated the item parameters based on the 2PL IRT model.

With the CB method, the item parameters were concurrently calibrated using a multi-group bifactor model. For the general factor, grade 5 students are treated as the reference group with a standard normal distribution. The SDs of grade 3 and 4 are also fixed at 1, while the means of these two grades are freely estimated.

The common item design was used in the simulation study. Under the common item design, common items are used for adjacent grades. In this simulation study there are three grades, so two sets of common items are needed. With the set of common items between grade 3 and 4 labeled as C34 and between grade 4 and 5 as C45, the following figure represents the model specification.

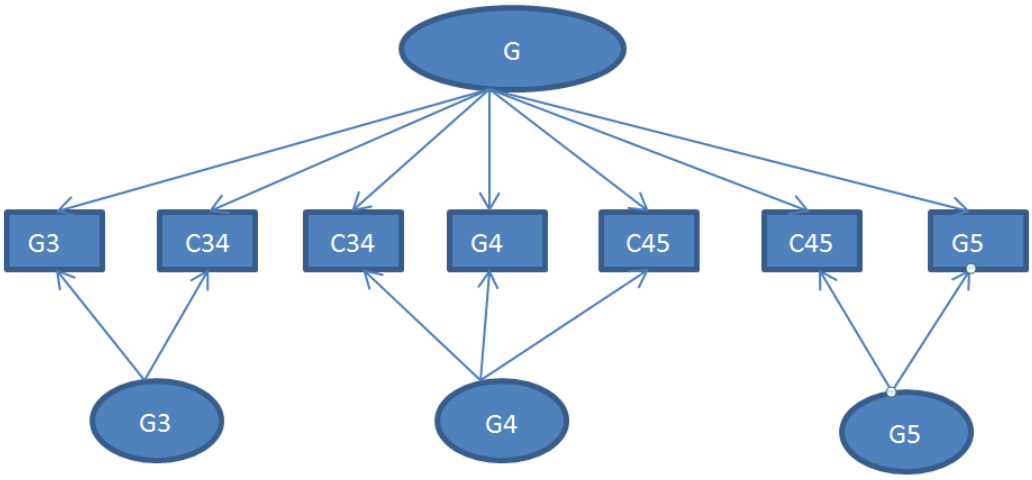

Figure 3.1 Nonequivalent Group Common Item Design with Bifactor Model

For the common items, students in each grade use both the general ability and their respective grade specific ability to answer those items. For example, although C34 is the set of items that is common to students in both grade 3 and 4, students in grade 3 use the general ability and the grade 3 ability to answer them, while students in grade 4 use the general ability and the grade 4 ability.

Based on the model specification, concurrent bi-factor calibration was conducted. Students only answered the grade specific items and those common items shared between their grade and the adjacent ones, and those items not answered by them were considered not reached and therefore treated as missing data in the calibration process.

The item parameters based on the bifactor model vertical scaling method were then used to score the students across grades. Only the scores on the general dimension were used to estimate the teacher effects.

With the SU method, since no vertical scaling is required, the item parameters for each grade were separately calibrated based on the unidimensional 2PL IRT model.

The calibrated item parameters based on the above three methods are used to score the item response data based on the second set of generated ability values. It should be noted again that while each student has only one score with the CU and the SU methods, each student receives two scores on the general factor and the grade specific factor respectively with the CB method. Only the scores on the general factor are used for the teacher effect estimate. The computer program IRTPRO (Cai, du Toit, & Thissen, 2012) using marginal maximum likelihood estimation with an EM algorithm is used for both calibration and scoring.

3.6 Teacher Effect Estimation

Depending on the vertical scaling methods, student scores are used as the dependent variable in different persistence models for teacher effect estimation. Scores coming out of the CU method and the CB method are used in the ZP model, the CP model, or the VP model, depending on the corresponding generating model. Scores coming out of the SU method are used in the GP model. While both the Bayesian approach and the maximum likelihood (ML) approach can be used to estimate persistence models, the former one requires the specification of an informative prior distribution for the covariance parameters, the choice of which may affect parameter estimates. Therefore, the ML approach is chosen as the estimation algorithm. Specifically, the R package GPvam (Karl, Yang, & Lohr, 2012), which employs the maximum likelihood estimation

method for the multiple membership mixed models used in VAM, is used for teacher effect estimation.

3.7 Evaluation

One of the commonly used evaluation criteria is the classification accuracy of teachers or schools in terms of their added value (McCaffrey et al., 2004; Lockwood et al., 2007b; Briggs & Weeks, 2009). In this study, the classification accuracy is computed using three approaches: standard error based, tercile grouping, and quintile grouping. With the standard error based approach, a 95% confidence interval around individual teacher effect estimate is established by adding and subtracting 1.96 * standard error to the estimate. If the lower bound of the confidence interval is above 0 which is assumed as the cut score, this particular teacher is classified as effective; if the higher bound of the confidence interval is below 0, this teacher is classified as ineffective.

One issue with the previous classification scheme is that if the standard errors of teacher effect estimates are large, most of the teachers will be classified into the average category. Different from the standard error based approach that takes into consideration the estimate uncertainty due to sampling error, quintile and tercile grouping approaches divide the entire teacher effect distribution into several equal parts and classify teachers based on their individual percentiles. With tercile groupings, teachers are classified into three categories of equal size, with the bottom third classified as ineffective, the middle third as average, and the top third as effective. With quintile groupings, teachers are classified into five performance categories of equal size, which often are labeled as

ineffective, less ineffective, average, less effective, and effective. In order to make the three approaches comparable, in this study the top two categories are combined into one category as effective and the bottom two categories are combined into one category as ineffective.

To compute the classification accuracy, the numbers of correct classifications and incorrect classifications are calculated. First, teachers are classified as ineffective or effective based on whether their generating teacher effect values are below or above the population mean, which is their true status. Depending on which approach is used, an individual teacher's computed category is compared with his or her true status: if they match, it is defined as a correct classification; if they do not match, it is defined as an incorrect classifications.

Another commonly used evaluation criteria is the Pearson correlation of the estimated values of teacher effects and the generating values of teacher effects (Briggs & Weeks, 2009; Hong, 2010). Different from those studies with focuses on parameter recovery, the current study focuses on how construct shift impacts teacher classification and the rank order of teachers is considered more relevant. Therefore, the Spearman correlation is more appropriate and used as an evaluation criterion in this simulation study.

3.8 Analysis

Using the percentage of correctly classified teachers and the Pearson correlation coefficient, three-way analyses of variance (ANOVA) are conducted to determine the

existence of statistically significant effects, may it be the main effects or the interaction effects. In addition to statistical significance, eta-squared is computed and reported as an effect size index to address the impact of the manipulated factors.

CHAPTER 4: RESULTS

In this chapter, results of the simulation study are presented. Specifically, section 4.1 answers the first research question on the accuracy of teacher effect estimation with the SU method when data are generated with the bifactor model; section 4.2 answers the second and the last research questions by comparing the teacher effect estimation accuracy with the CU method, the CB method, and the SU method; in section 4.3, the results of three-way analysis of variance (ANOVA) are examined for the statistical effects of model choice, magnitude of construct shift, and different vertical scaling methods; in section 4.4, the main findings are summarized.

4.1 Accuracy of the SU method

Convergence was reached with each of the estimation runs. The accuracy of the SU method is described with two indices: the spearman correlation of the teacher effect estimates with the generating teacher effect values and the teacher classification accuracy. In terms of classification accuracy, both the numbers of incorrect and correct classifications under the three different classification schemes (standard error based, quintile grouping, and tercile grouping) were presented. Section 4.1.1 discusses the accuracy in terms of spearman correlation; section 4.1.2 addresses the classification accuracy, with 4.1.2.1 focusing on the standard error based approach, 4.1.2.2 on the quintile grouping approach, and 4.1.2.3 on the tercile grouping approach.

4.1.1 Spearman Correlation

The Spearman correlation coefficient is used to assess how well the rank orders of two variables agree. In Table 4.1, the values of the Spearman correlation coefficient between the teacher effect estimates of the SU method and the generating teacher effect values for each of the 9 simulation conditions are presented.

The correlation coefficient value ranges from 0.47 to 0.91. With a certain persistence pattern in a certain year, the pattern is that when the magnitude of construct shift decreases, the Spearman correlation coefficient increases, with the only exception of the last row of Table 4.1, which is ZP in year 3. This is somewhat unexpected considering that the SU method is created to address the issue of construct shift, while the results indicate its accuracy deteriorates with the increase of the magnitude of such shift. Another interesting finding is that CP has the highest correlation value with a mean of 0.80, while ZP has the lowest correlation value with a mean of 0.67. It seems that the correlation value decreases with the decrease of the value of the persistence parameter. Last, the impact of persistence pattern decreases from year 1 to year 3: in year 1 there are obvious differences among CP, VP and ZP; in year 2 the differences shrink; in year 3 the correlation values become virtually identical for the three persistence patterns.

Graphs showing the mean correlation between teacher effect estimates and the generating values in different years are presented in Figures 4.1a through 4.1c.

Table 4.1 Spearman Correlation between Teacher Effect Estimates and True Values for the SU method

		Variance (Magnitude of Construct Shift)		
Year	Persistence Pattern	1	0.50	0.25
Year 1	CP	0.85	0.89	0.91
	VP	0.72	0.75	0.78
	ZP	0.60	0.62	0.66
Year 2	CP	0.72	0.86	0.91
	VP	0.64	0.80	0.88
	ZP	0.50	0.74	0.85
Year 3	CP	0.47	0.80	0.80
	VP	0.49	0.82	0.82
	ZP	0.50	0.84	0.83

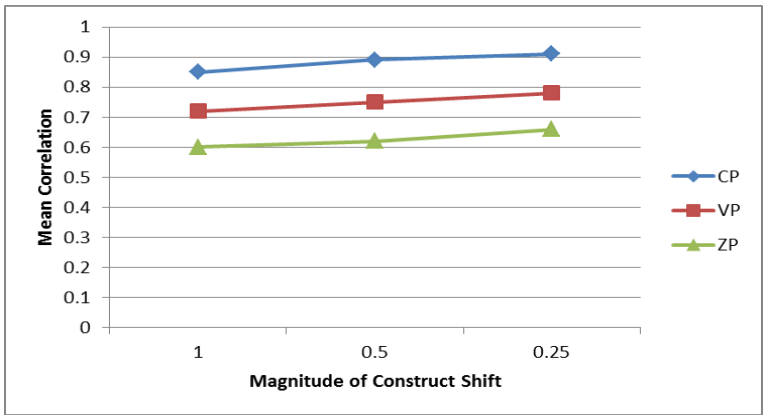

Figure 4.1a Mean Correlation between the Estimates and Generating Values in Year 1

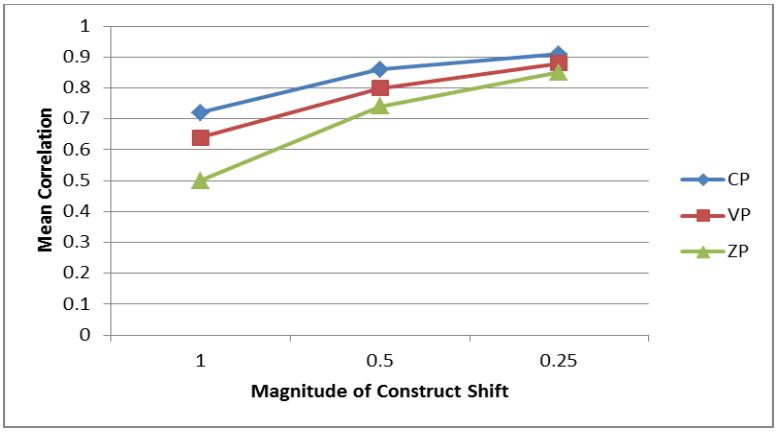

Figure 4.1b Mean Correlation between the Estimates and Generating Values in Year 2

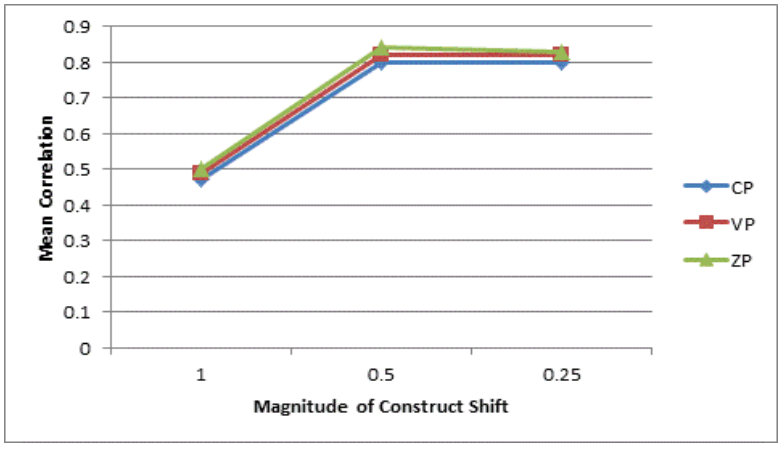

Figure 4.1c Mean Correlation between the Estimates and Generating Values in Year 3

4.1.2 Classification Accuracy

This section discusses classification accuracy of the SU method in terms of incorrect and correct classifications. Incorrect classification is defined as a teacher being classified as effective when the generated teacher effect value is below the population mean or a teacher being defined as ineffective when the generated teacher effect value is above the population mean; correct classification is defined as a teacher being classified as effective when the generated teacher effect value is above the population mean or a teacher being defined as ineffective when the generated teacher effect value is below the population mean. The average number of incorrect and correct classifications across replications is presented in Table 4.2, Table 4.3, and Table 4.4.

4.1.2.1 Standard Error Based Approach

The standard error based approach establishes the confidence interval of the teacher effect estimate by adding and subtracting 1.96 * standard error to the estimate. If the lower bound of the confidence interval is above 0, this particular teacher is classified as effective; if the higher bound of the confidence interval is below 0, this teacher is classified as ineffective.

The left portion of Table 4.2 presents the number of incorrect classifications of the SU method under different simulation conditions in different years. Except for VP and ZP with variance equal to 1 in year 1, it seems that the SU method performs extremely well in terms of making no incorrect classifications. Even for those two simulation conditions under which incorrect classification occurs, the number of incorrect classification is so small that on average far less than 1 teacher is misclassified.

The right portion of Table 4.2 presents the number of correct classification of the SU method, which ranges from 0 to 13.56. One consistent pattern across different persistence patterns in different years is that with the decrease of magnitude of construct shift, the number of correct classifications increases. Another pattern is that in year 1 and year 2, the number of correct classifications seems to decrease with the decrease of the value of the persistence parameter (CP > VP > ZP); while in year 3, this pattern seems to be reversed with ZP having the highest number of correct classifications.

Graphs showing the percentage of correct classifications based on standard errors in different years are presented in Figures 4.2a through 4.2c.

Table 4.2 Correct and Incorrect Classifications Based on Standard Error

Year	Persistence Pattern	Incorrect Classification			Correct Classification		
		Variance (Magnitude of Construct shift)			Variance (Magnitude of Construct shift)		
		1	0.5	0.25	1	0.5	0.25
Year 1	CP	0	0	0	13.56	10.74	11
	VP	0.09	0	0	7.41	3.55	4.96
	ZP	0.19	0	0	1.02	1.18	1.99
Year 2	CP	0	0	0	1.40	5.97	11.96
	VP	0	0	0	1.45	2.17	7.89
	ZP	0	0	0	0.67	0.63	4.22
Year 3	CP	0	0	0	0	0.46	5.95
	VP	0	0	0	0	1.05	7.11
	ZP	0	0	0	0	1.62	7.92

* The numbers in the above table are the average number of correctly or incorrectly classified teachers across 100 replications.

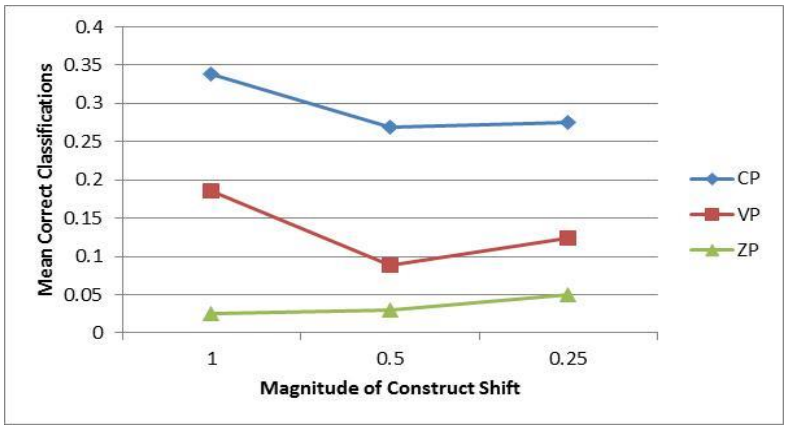

Figure 4.2a Mean Correct Classifications Based on SE in Year 1

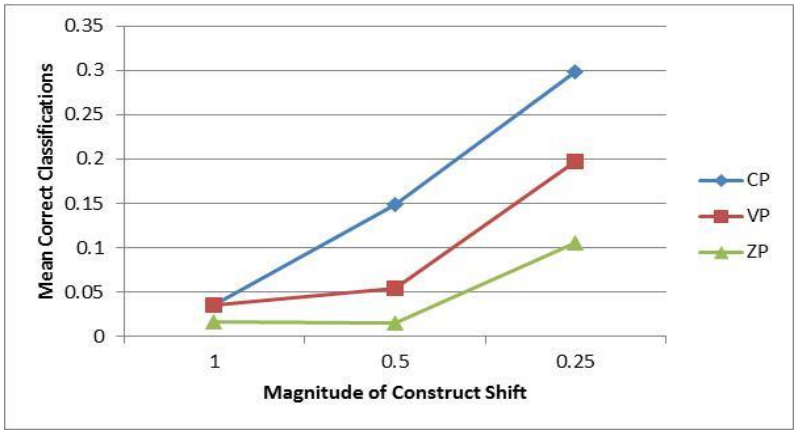

Figure 4.2b Mean Correct Classifications Based on SE in Year 2

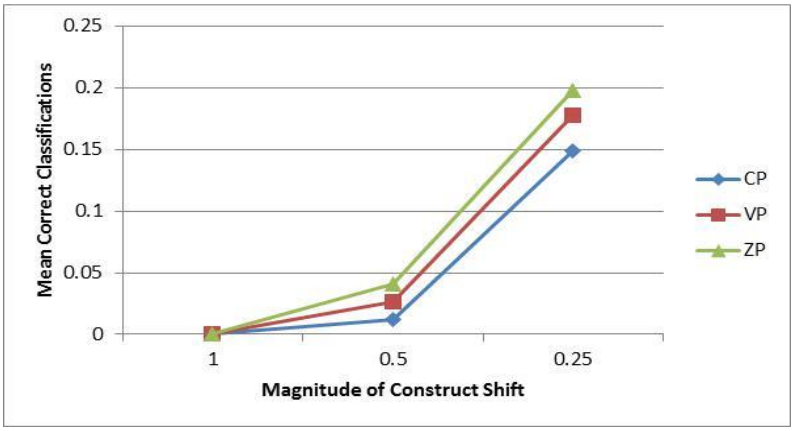

Figure 4.2c Mean Correct Classifications Based on SE in Year 3

4.1.2.2 Quintile Grouping Approach

The quintile grouping approach and the tercile grouping approach in the next section differs from the preceding standard error based approach in the sense that the uncertainty of statistical estimate is not taken into consideration. With the quintile grouping approaching, the whole distribution is divided into five parts using four quintile points, and the upper 40% is considered effective and the lower 40% is considered ineffective.

Table 4.3 presents the number of incorrect and correct classifications of the SU method under different simulation conditions in different years. The number of incorrect classifications in the left half ranges from 1.63 to 10.33, and the general pattern is that the number of incorrect classifications decreases with the decrease of the magnitude of construct shift regardless of the persistence pattern and year, with the exception of CP in year 1. Another pattern is that in year 1 and year 2, the number of incorrect classifications seems to increase with the decrease of the value of the persistence parameter (CP > VP > ZP); while in year 3, this pattern does not seem to hold with ZP having the lowest number of incorrect classifications except when the variance is equal to 0.25. One counterintuitive finding is that in year 3, when the magnitude of construct shift decreases from 0.5 to 0.25, the number of incorrect classifications increases marginally regardless of the persistence patterns. The persistence model is known to have the property of increasing score variance at higher grades due to the accumulation of teacher effect. It is believed that this surprising increase of incorrect classifications with the decrease of magnitude of construct shift happens because in year 3, when the score has the largest

variance across three years, the 0.25 increase of magnitude of construct shift is less influential than in the previous two years and is offset by sampling error.

The number of correct classifications in the right half of Table 4.3 ranges from 21.94 to 30.37, and the general pattern is that the number of correct classifications increases with the decrease of the magnitude of construct shift regardless of the persistence pattern and year. Another pattern is that in year 1 and year 2, the number of correct classifications seems to decrease with the decrease of the value of the persistence parameter (CP > VP > ZP); while in year 3, this pattern does not seem to hold with ZP having the highest number of correct classifications except when the variance is equal to 0.25.

Graphs showing the percentage of incorrect and correct classifications based on quintile grouping in different years are presented in Figures 4.3a through 4.3f.

Table 4.3 Correct and Incorrect Classifications Based on Quintile Grouping

Year	Model	Incorrect Classification			Correct Classification		
		Variance			Variance		
		1	0.5	0.25	1	0.5	0.25
Year 1	CP	2.94	3.16	1.65	29.06	28.84	30.35
	VP	6.14	4.97	4.74	25.86	27.03	27.26
	ZP	7.04	6.69	5.76	24.96	25.31	26.24
Year 2	CP	6.26	3.70	1.63	25.74	28.30	30.37
	VP	7.66	4.06	1.87	24.34	27.94	30.13
	ZP	9.74	4.32	1.73	22.26	27.68	30.27
Year 3	CP	10.33	2.30	3.73	21.67	29.70	28.27
	VP	10.06	2.11	3.44	21.94	29.89	28.56
	ZP	9.66	1.87	3.54	22.34	30.13	28.46

* The numbers in the above table are the average number of correctly or incorrectly classified teachers across 100 replications.

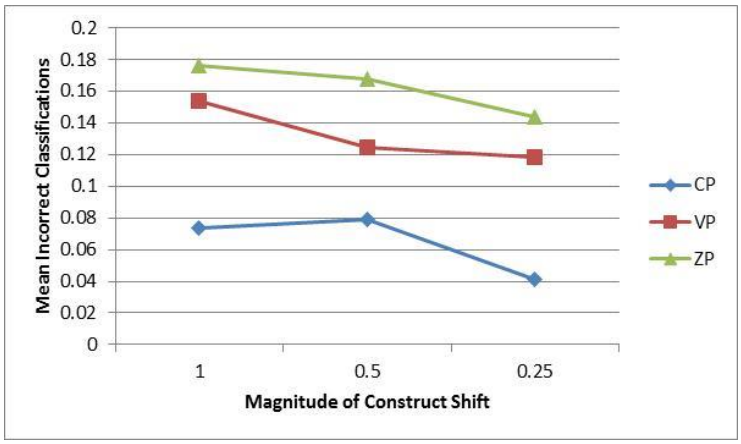

Figure 4.3a Mean Incorrect Classifications Based on Quintile Grouping in Year 1

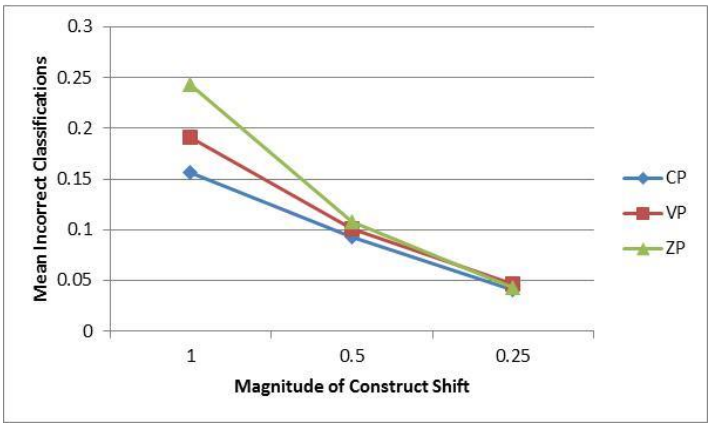

Figure 4.3b Mean Incorrect Classifications Based on Quintile Grouping in Year 2

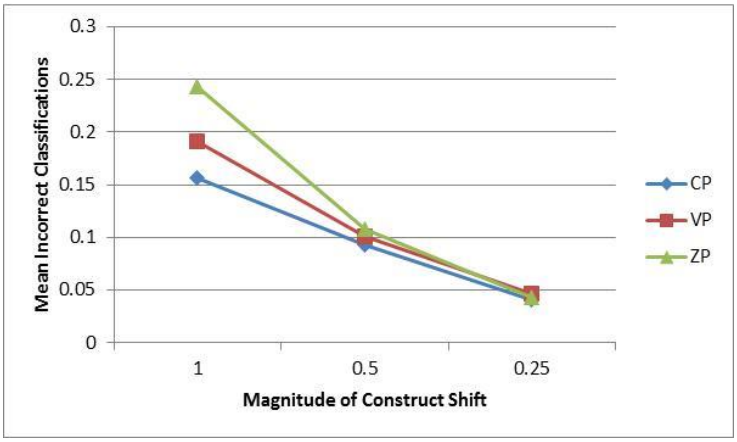

Figure 4.3c Mean Incorrect Classifications Based on Quintile Grouping in Year 3

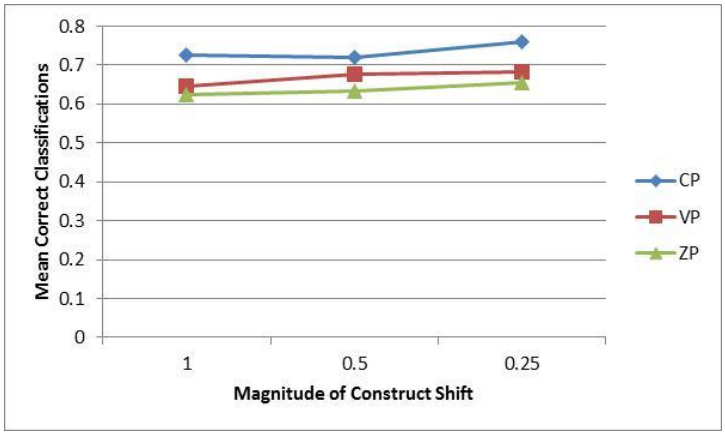

Figure 4.3d Mean Correct Classifications Based on Quintile Grouping in Year 1

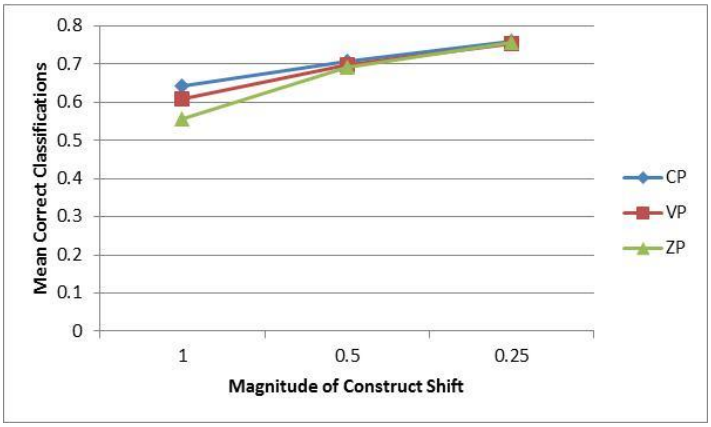

Figure 4.3e Mean Correct Classifications Based on Quintile Grouping in Year 2

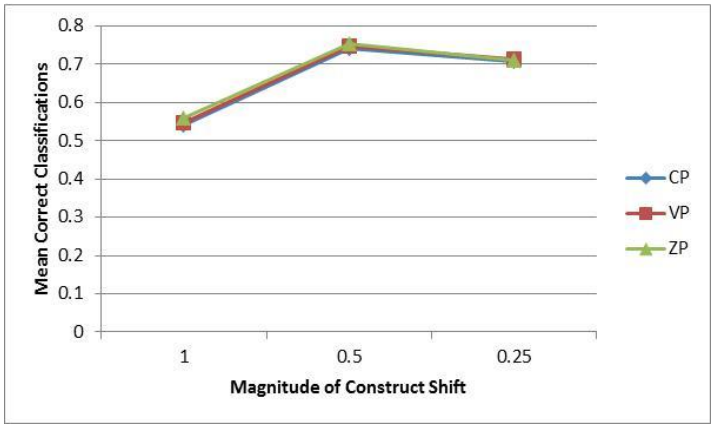

Figure 4.3f Mean Correct Classifications Based on Quintile Grouping in Year 3

4.1.2.3 Tercile Grouping Approach

With the tercile grouping approaching, the whole distribution is divided into three parts using two tercile points, and the top third is considered effective and the bottom third is considered ineffective.

Table 4.4 presents the number of correct and incorrect classifications of the SU method under different simulation conditions in different years. The number of incorrect classifications in the left portion of the table ranges from 0.31 to 7.24, and the general pattern is that the number of incorrect classifications decreases with the decrease of the magnitude of construct shift regardless of the persistence pattern and year, with the exception CP in year 1. Another pattern is that regardless of the year, the number of incorrect classifications seems to increase with the decrease of the value of the persistence parameter (CP > VP > ZP).

The number of correct classifications in the right portion of the table ranges from 18.76 to 25.69, and the general pattern is that the number of correct classifications increases with the decrease of the magnitude of construct shift regardless of the persistence pattern and year, with the exception of CP in year 1. Another pattern is that in year 1 and year 2, the number of correct classifications seems to decrease with the decrease of the value of the persistence parameter (CP > VP > ZP); while in year 3, this pattern does not seem to hold with ZP having the highest number of correct classifications except when the variance is equal to 0.25.

Graphs showing the percentage of incorrect and correct classifications based on tercile grouping in different years are presented in Figures 4.4a through 4.4f.

Table 4.4 Correct and Incorrect Classifications Based on Tercile Grouping

Year	Persistence Pattern	Incorrect Classification			Correct Classification		
		Variance (Magnitude of Construct Shift)			Variance (Magnitude of Construct Shift)		
		1	0.5	0.25	1	0.5	0.25
Year 1	CP	1.12	1.66	0.31	24.88	24.34	25.69
	VP	3.91	3.32	2.39	22.09	22.68	23.61
	ZP	5.48	3.95	4	20.52	22.05	22.00
Year 2	CP	3.10	2.08	0.81	22.9	23.92	25.19
	VP	4.23	2.62	0.9	21.77	23.38	25.10
	ZP	6.02	2.62	0.82	19.98	23.38	25.18
Year 3	CP	7.24	1.13	1.74	18.76	24.87	24.26
	VP	7.03	0.82	1.53	18.97	25.18	24.47
	ZP	6.89	0.79	1.18	19.11	25.21	24.2

* The numbers in the above table are the average number of correctly or incorrectly classified teachers across 100 replications.

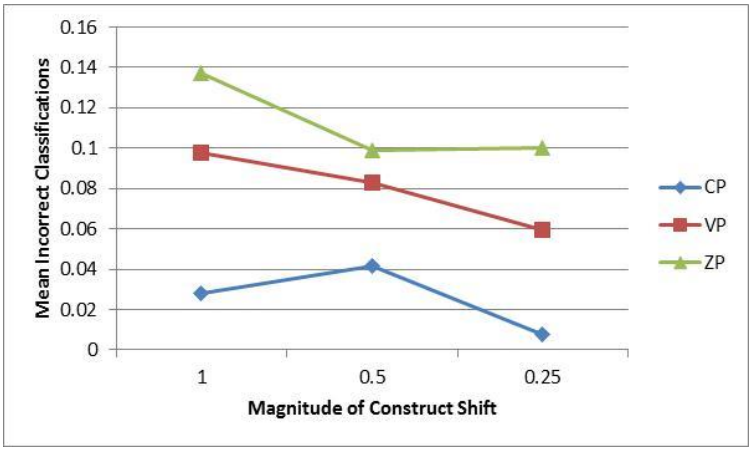

Figure 4.4a Mean Incorrect Classifications Based on Tercile Grouping in Year 1

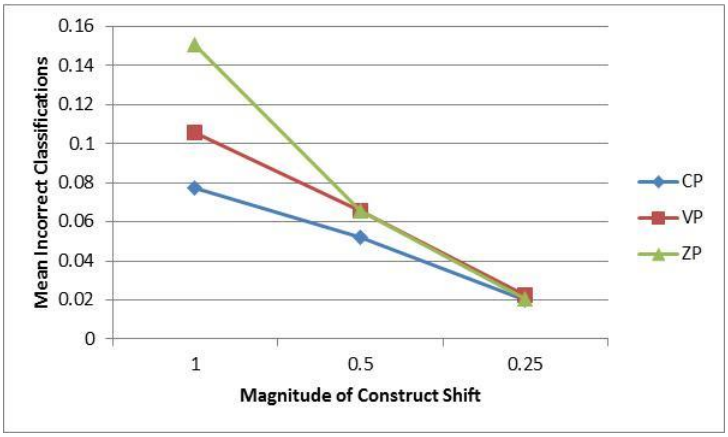

Figure 4.4b Mean Incorrect Classifications Based on Tercile Grouping in Year 2

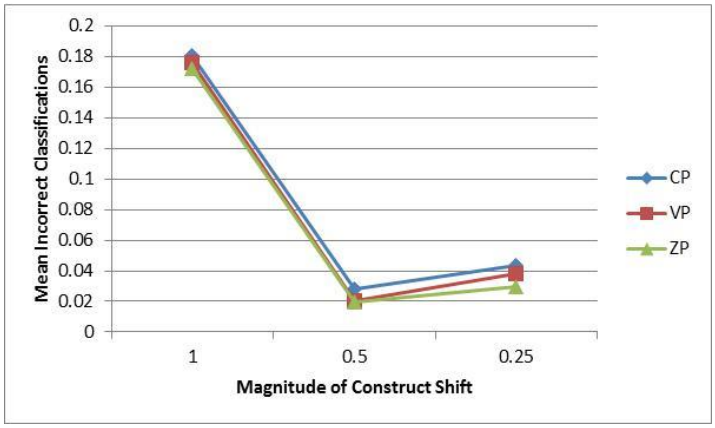

Figure 4.4c Mean Incorrect Classifications Based on Tercile Grouping in Year 3

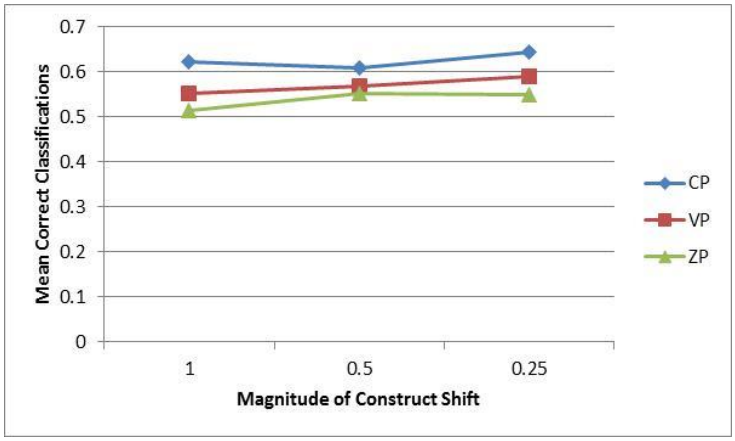

Figure 4.4d Mean Correct Classifications Based on Tercile Grouping in Year 1

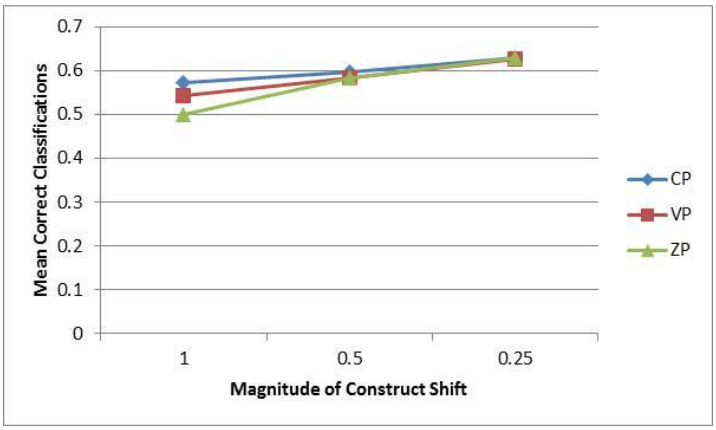

Figure 4.4e Mean Correct Classifications Based on Tercile Grouping in Year 2

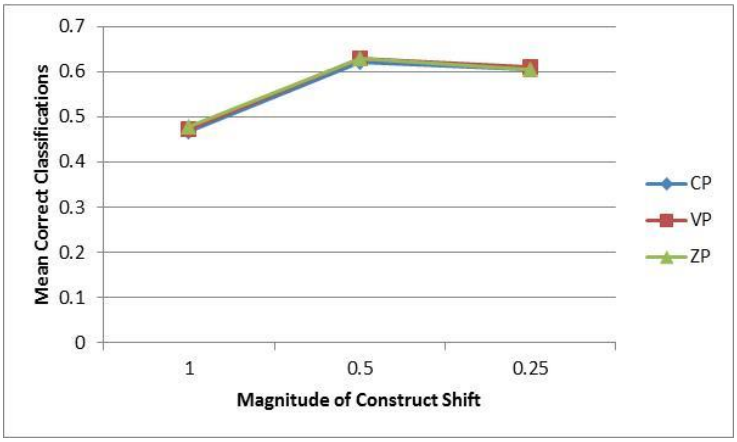

Figure 4.4f Mean Correct Classifications Based on Tercile Grouping in Year 3

4.2 Comparison of Teacher Estimation Accuracy of Different Methods

Similar to section 4.1, the comparison of teacher effect estimation across different methods was carried out using two sets of values as indices: the spearman correlation of the teacher effect estimates with the generating teacher effect values and the teacher classification accuracy under the three different classification schemes. Section 4.2.1 discusses the spearman correlation; section 4.1.2 addresses the classification accuracy, with 4.1.2.1 focusing on the standard error based approach, 4.1.2.2 on the quintile grouping approach, and 4.1.2.3 on the tercile grouping approach.

4.2.1 Comparison of the Spearman Correlation Values

Table 4.5 presents the mean Spearman correlation values under the 27 simulation conditions in different years and the corresponding standard deviations in the parenthesis. To address the second research question of comparing the CB and SU methods, comparison needs to be carried out between the first and the third column; to compare the CB and CU method, comparison needs to be carried out between the first and the second column. In the first column (CB), the Spearman correlation value ranges from 0.24 to 0.83 with a mean of 0.66, in the second column (CU) the Spearman correlation value ranges from 0.05 to 0.87 with a mean of 0.66, and in the third column (SU) the Spearman correlation value ranges from 0.47 to 0.91 with a mean of 0.74. Across different simulation conditions and different years it seems that overall the SU method does a better job than the CB method, which perform similarly to the CU method overall. It should be noted that the minimum value for the CU method is only 0.05.

To further investigate the difference between the CB and SU method, the comparison is broken down into different models in different years. For the CP persistence pattern in year 1, the SU method is consistently better than the CB method, although the difference is marginal (the difference is at the second decimal point); for VP in year 1, the SU method and the CB method seem to perform similarly; for ZP in year 1, the SU method seems to perform noticeably better than the CB method, with the mean difference of approximately 0.3. For CP and VP in year 2, the CB method seems to perform better than the SU method when the variance is equal to 1, and this pattern is reversed when the variance decreases; for ZP in year 2, the pattern is similar to year 1 in the sense that the SU method performs noticeably better than the CB method with a mean difference of approximately 0.25, regardless of the variance. For all the persistence patterns in year 3, the CB method seems to perform noticeably better than the SU method with a mean difference of approximately 0.1 when the variance is equal to 1, and this pattern is reversed except for ZP with variance equal to 0.25.

To further investigate the difference between the CB and CU method, the comparison is broken down into different persistence patterns in different years. For all the persistence patterns in year 1, the CB method and the CU method seem to perform similarly. For all the persistence patterns in year 2, the CB method seems to perform better than the CU method when the variance is equal to 1, and this pattern is reversed when the variance decreases. For all the persistence patterns in year 3, the same pattern seems to exist, and the CB method seems to perform noticeably better than the CU method with a mean difference of approximately 0.11 when the variance id equal to 1,

and this pattern is reversed when the variance decreases except for ZP with a variance equal to 0.25.

Graphs showing the mean correlation between teacher effect estimates and the generating values using different scoring methods in different years are presented in Figures 4.5a through 4.5i.

Table 4.5 Spearman Correlation between Teacher Effect Estimates and True Values

Year	Persistence Pattern	Variance	Vertical Scaling Method		
			CB	CU	SU
Year 1	CP	1	0.81(0.02)	0.80(0.01)	0.85(0.01)
		0.50	0.82(0.02)	0.81(0.01)	0.89(0.01)
		0.25	0.83(0.02)	0.86(0.01)	0.91(0.01)
	VP	1	0.73(0.03)	0.72(0.01)	0.72(0.01)
		0.50	0.76(0.03)	0.75(0.02)	0.75(0.02)
		0.25	0.77(0.04)	0.78(0.02)	0.78(0.02)
	ZP	1	0.33(0.09)	0.38(0.04)	0.60(0.01)
		0.50	0.25(0.12)	0.17(0.06)	0.62(0.02)
		0.25	0.31(0.12)	0.31(0.06)	0.66(0.02)
Year 2	CP	1	0.74(0.05)	0.68(0.03)	0.72(0.04)
		0.50	0.77(0.04)	0.81(0.03)	0.86(0.02)
		0.25	0.81(0.05)	0.87(0.02)	0.91(0.02)
	VP	1	0.74(0.05)	0.71(0.02)	0.64(0.05)
		0.50	0.77(0.05)	0.80(0.02)	0.80(0.02)
		0.25	0.81(0.04)	0.87(0.02)	0.88(0.02)
	ZP	1	0.24(0.11)	0.05(0.05)	0.50(0.04)
		0.50	0.57(0.10)	0.63(0.03)	0.74(0.02)
		0.25	0.51(0.12)	0.59(0.05)	0.85(0.02)
Year 3	CP	1	0.54(0.07)	0.41(0.03)	0.47(0.03)
		0.50	0.75(0.06)	0.79(0.02)	0.80(0.02)
		0.25	0.72(0.06)	0.80(0.03)	0.80(0.03)
	VP	1	0.61(0.07)	0.50(0.03)	0.49(0.04)
		0.50	0.80(0.05)	0.82(0.02)	0.82(0.02)
		0.25	0.77(0.06)	0.82(0.03)	0.82(0.03)
	ZP	1	0.60(0.08)	0.47(0.04)	0.50(0.04)
		0.50	0.80(0.05)	0.83(0.02)	0.84(0.02)
		0.25	0.77(0.05)	0.82(0.03)	0.83(0.03)

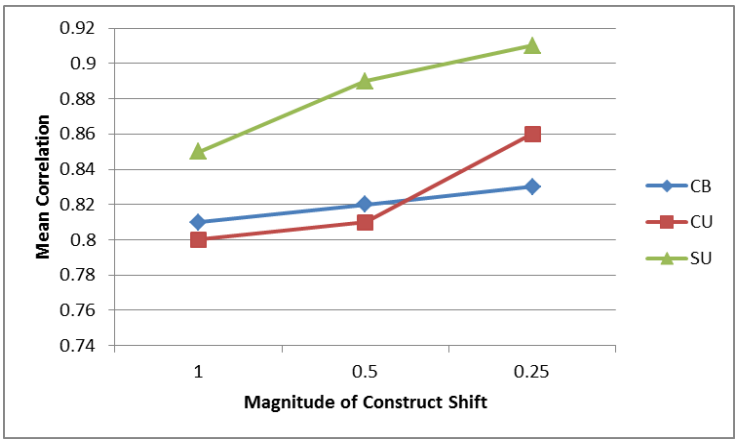

Table 4.5a Mean Correlation Comparison for CP in Year 1

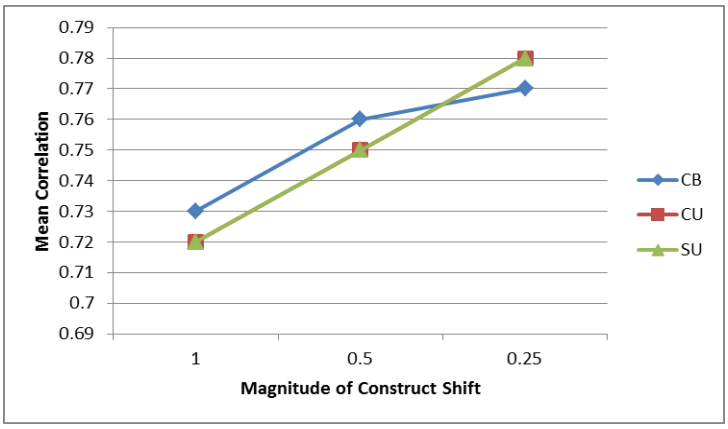

Table 4.5b Mean Correlation Comparison for VP in Year 1

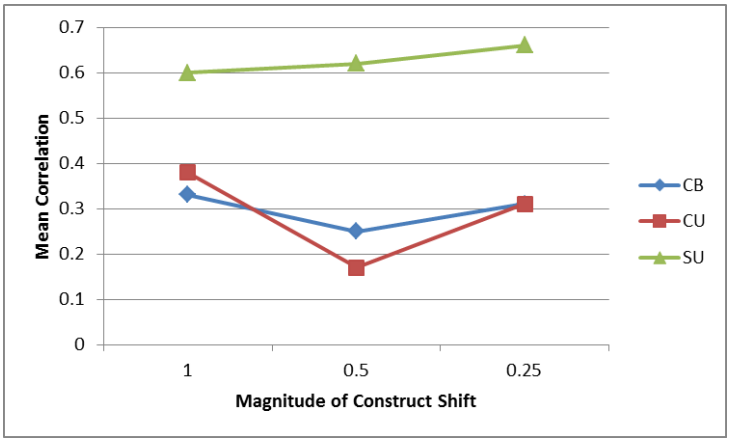

Table 4.5c Mean Correlation Comparison for ZP in Year 1

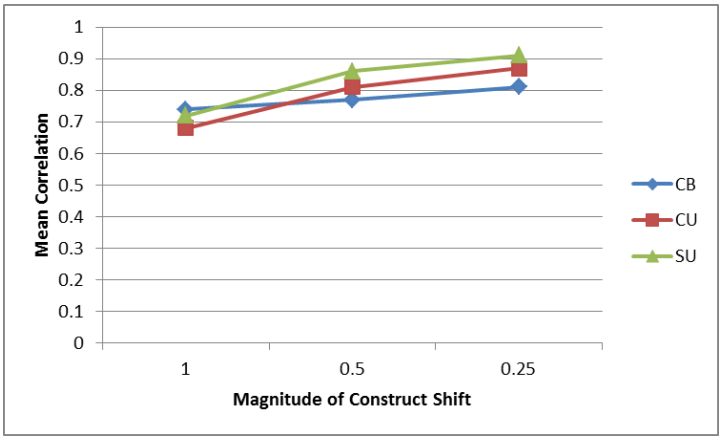

Table 4.5d Mean Correlation Comparison for CP in Year 2

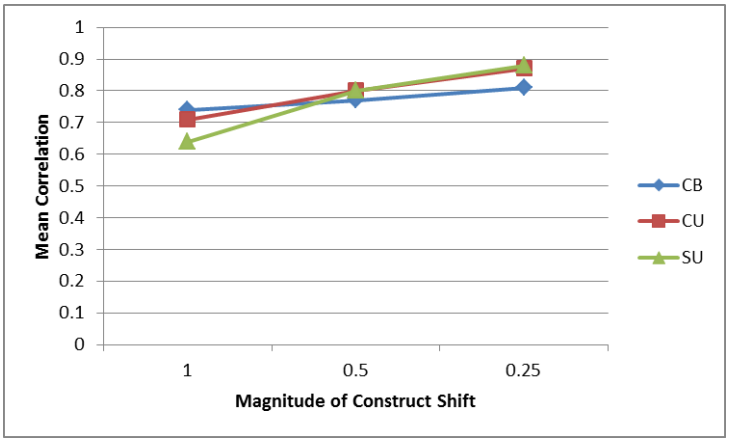

Table 4.5e Mean Correlation Comparison for VP in Year 2

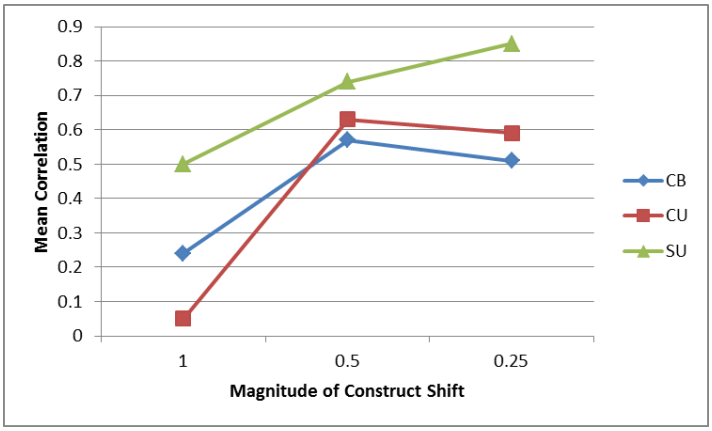

Table 4.5f Mean Correlation Comparison for ZP in Year 2

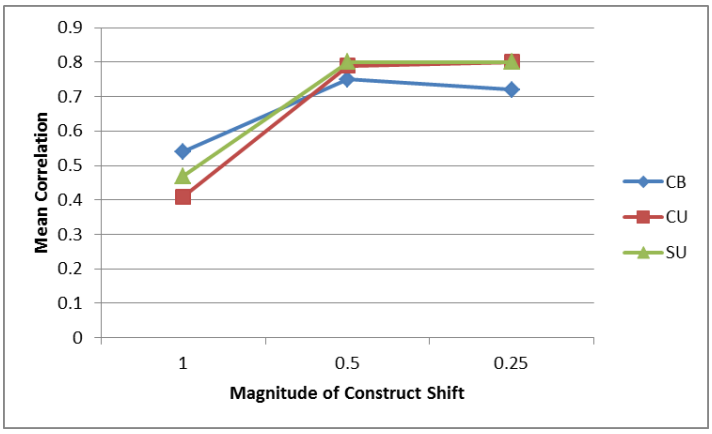

Table 4.5g Mean Correlation Comparison for CP in Year 3

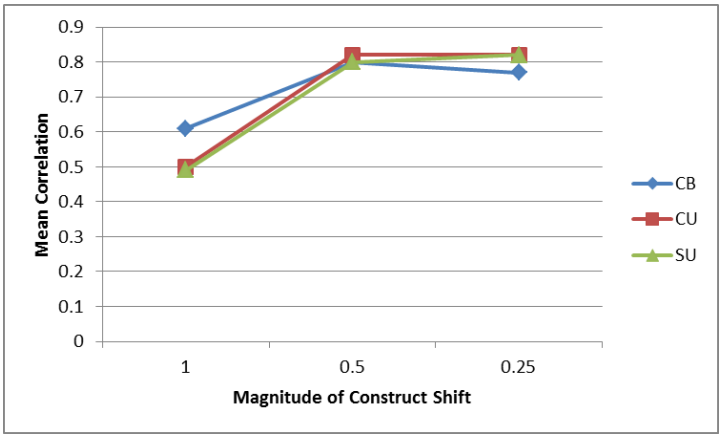

Table 4.5h Mean Correlation Comparison for VP in Year 3

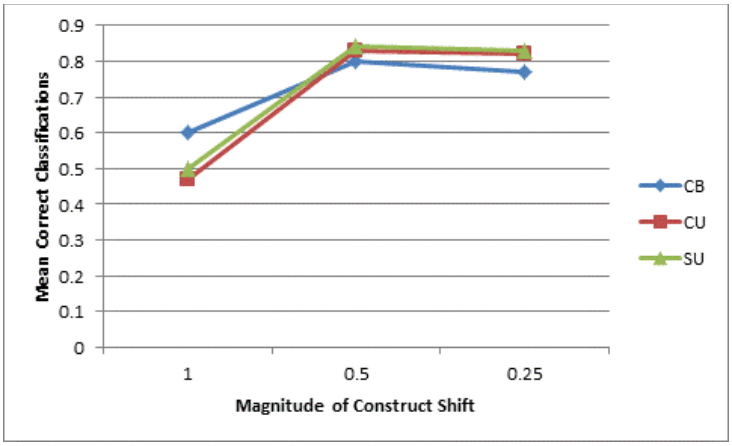

Table 4.5i Mean Correlation Comparison for ZP in Year 3

4.2.2 Comparison of the Classification Accuracy

4.2.2.1 Comparison Based on Standard Error

Table 4.6 presents the number of incorrect and correct classifications based on the standard errors under the 27 simulation conditions in different years and the corresponding standard deviations in the parenthesis. To compare the CB and SU method, comparison of the number of incorrect classifications needs to be carried out between the first and the third column; to compare the CB and CU method, the comparison needs to be carried out between the first and the second column. In the first column (CB), the number of incorrect classifications ranges from 0 to 0.31 with a mean of 0.05, in the second column (CU) the number of incorrect classifications ranges from 0 to 0.30 with a mean of 0.01, and in the third column (SU) the number of incorrect classifications ranges from 0 to 0.19 with a mean of 0.01. In terms of the number of correct classifications, comparison needs to be carried out between the fourth and the sixth column and the fourth and the fifth column. In the fourth column (CB), the number of correct classifications ranges from 0.01 to 11.81 with a mean of 3.80, in the fifth column (SU) the number of correct classifications ranges from 0 to 11.92 with a mean of 3.93, and in the sixth column (SU) the number of correct classifications ranges from 0 to 13.56 with a mean of 4.29. Across different simulation conditions and different years it seems that overall the SU method does a better job than both the CB and CU method, although it should be noted that all three methods have very low numbers of incorrect and correct classifications.

To further investigate the difference between the CB and SU method, the comparison is broken down into different persistence pattern in different years. In terms

of the number of incorrect classifications, it seems that the CB and the SU methods perform similarly for all the persistence patterns in year 1; for all the persistence patterns in year 2 and year 3, the number of incorrect classifications of the CB method is consistently no less than that of the SU method, although the differences are marginal. In terms of the number of correct classifications, the SU method seems to perform better than the CB method for all the persistence patterns in year 1; in year 2, while the same pattern exists for ZP, for CP and VP have a different pattern in the sense that only when variance is not equal to 1 does the SU method perform better than the CB method; in year 3, a different pattern was found: the SU method performs better than the CB method only when the variance is equal to 0.25.

To further investigate the difference between the CB and CU method, the comparison is broken down into different persistence patterns in different years. In terms of the number of incorrect classifications, the CB and the CU methods perform similarly for all the persistence patterns in year 1; for all the models in year 2 and year 3, the number of incorrect classifications of the CB method is consistently no less than that of the CU method with the exception of ZP with variance equal to 0.25 in year 2, although the differences are marginal. In terms of the number of correct classifications, the general pattern is that the CB method always performs better than the CU method unless the variance is equal to 0.25; another pattern is that in year 1 and year 2, the ZP has noticeably lower numbers of correct classifications than both CP and VP, while in year 3 the differences diminish considerably.

Graphs showing the percentage of correct classifications based on standard errors for different persistence models in different years are presented in Figures 4.6a through 4.6i.

Table 4.6 Correct and Incorrect Classifications Based on SE

Year	Persistence Pattern	Variance	**Incorrect Classifications** Vertical Scaling Method CB	CU	SU	**Correct Classifications** Vertical Scaling Method CB	CU	SU
Year 1	CP	1	0.00	0.00	0.00	11.81	11.92	13.56
			(0.00)	(0.00)	(0.00)	(1.33)	(1.03)	(1.38)
		0.50	0.01	0.00	0.00	9.53	8.36	10.74
			(0.10)	(0.00)	(0.00)	(1.81)	(1.11)	(2.35)
		0.25	0.00	0.00	0.00	10.11	13.70	11.00
			(0.00)	(0.00)	(0.00)	(1.68)	(1.65)	(2.22)
	VP	1	0.09	0.30	0.09	7.85	8.17	7.41
			(0.29)	(0.46)	(0.29)	(1.63)	(0.90)	(0.93)
		0.50	0.05	0.01	0.00	5.90	3.79	3.55
			(0.22)	(0.10)	(0.00)	(1.72)	(0.78)	(0.85)
		0.25	0.00	0.00	0.00	7.36	9.02	4.96
			(0.00)	(0.00)	(0.00)	(1.76)	(1.54)	(1.27)
	ZP	1	0.00	0.00	0.19	0.03	0.00	1.02
			(0.00)	(0.00)	(0.39)	(0.17)	(0.00)	(0.32)
		0.50	0.00	0.00	0.00	0.01	0.00	1.18
			(0.00)	(0.00)	(0.00)	(0.10)	(0.00)	(0.50)
		0.25	0.00	0.00	0.00	0.03	0.00	1.99
			(0.00)	(0.00)	(0.00)	(0.17)	(0.00)	(0.67)
Year 2	CP	1	0.01	0.00	0.00	3.70	2.60	1.40
			(0.10)	(0.00)	(0.00)	(1.41)	(0.49)	(0.75)
		0.50	0.07	0.00	0.00	4.50	3.05	5.97
			(0.26)	(0.00)	(0.00)	(2.08)	(1.16)	(1.24)
		0.25	0.17	0.00	0.00	6.33	9.94	11.96
			(0.40)	(0.00)	(0.00)	(2.02)	(1.63)	(2.08)
	VP	1	0.02	0.00	0.00	5.04	3.60	1.45
			(0.14)	(0.00)	(0.00)	(1.94)	(0.86)	(0.59)
		0.50	0.09	0.00	0.00	4.73	1.59	2.17
			(0.29)	(0.00)	(0.00)	(2.39)	(1.00)	(1.01)
		0.25	0.31	0.03	0.00	7.06	10.31	7.89
			(0.46)	(0.17)	(0.00)	(2.12)	(1.50)	(1.71)
	ZP	1	0.00	0.00	0.00	0.01	0.00	0.67
			(0.00)	(0.00)	(0.00)	(0.1)	(0.00)	(0.55)
		0.50	0.02	0.00	0.00	0.26	0.00	0.63
			(0.14)	(0.00)	(0.00)	(0.61)	(0.00)	(0.80)
		0.25	0.00	0.01	0.00	0.25	0.00	4.22
			(0.00)	(0.10)	(0.00)	(0.78)	(0.00)	(1.19)

		1	0.01	0.00	0.00	0.31	0.00	0.00
	CP		(0.10)	(0.00)	(0.00)	(0.66)	(0.00)	(0.00)
		0.50	0.01	0.00	0.00	0.80	0.00	0.46
			(0.10)	(0.00)	(0.00)	(1.15)	(0.00)	(0.66)
Year 3		0.25	0.02	0.00	0.00	1.82	3.62	5.95
			(0.14)	(0.00)	(0.00)	(1.60)	(1.48)	(1.87)
		1	0.03	0.00	0.00	0.82	0.00	0.00
	VP		(0.17)	(0.00)	(0.00)	(1.03)	(0.00)	(0.00)
		0.50	0.01	0.00	0.00	2.94	0.85	1.05
			(0.10)	(0.00)	(0.00)	(2.03)	(0.86)	(0.96)
		0.25	0.00	0.00	0.00	4.06	7.21	7.11
			(0.00)	(0.00)	(0.00)	(2.01)	(1.42)	(1.50)
		1	0.03	0.00	0.00	0.72	0.00	0.00
	ZP		(0.17)	(0.00)	(0.00)	(0.99)	(0.00)	(0.00)
		0.50	0.02	0.00	0.00	2.70	1.15	1.62
			(0.14)	(0.00)	(0.00)	(1.95)	(0.98)	(1.15)
		0.25	0.02	0.00	0.00	3.94	7.26	7.92
			(0.14)	(0.00)	(0.00)	(2.01)	(1.47)	(1.61)

* The numbers in the above table are the average number of correctly or incorrectly classified teachers across 100 replications.

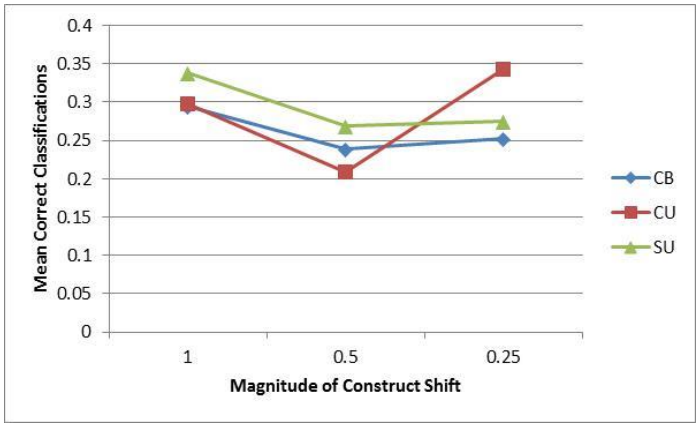

Figure 4.6a Mean Correct Classifications Based on SE for CP in Year 1

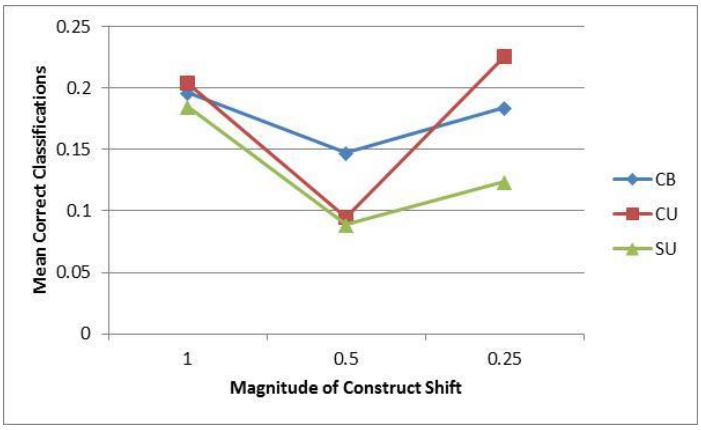

Figure 4.6b Mean Correct Classifications Based on SE for VP in Year 1

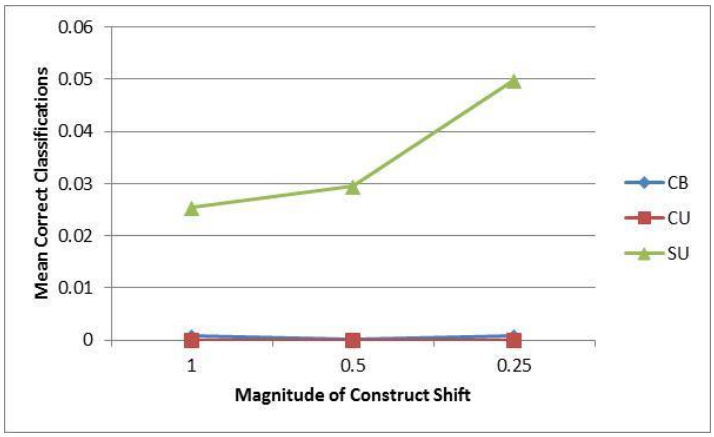

Figure 4.6c Mean Correct Classifications Based on SE for ZP in Year 1

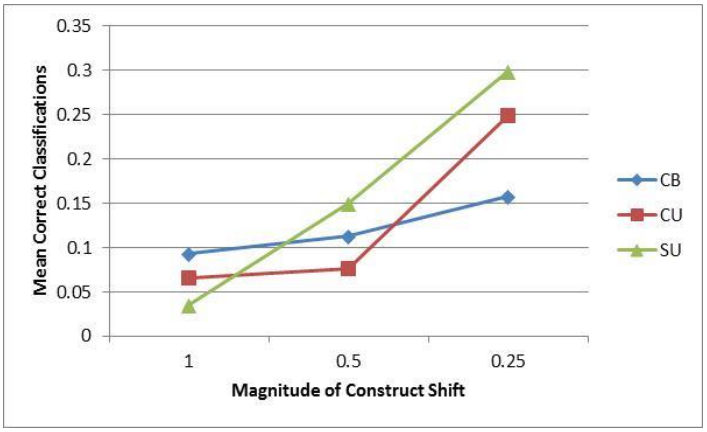

Figure 4.6d Mean Correct Classifications Based on SE for CP in Year 2

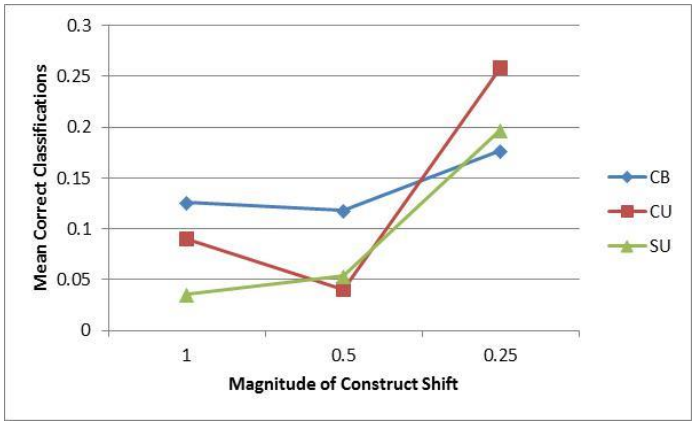

Figure 4.6e Mean Correct Classifications Based on SE for VP in Year 2

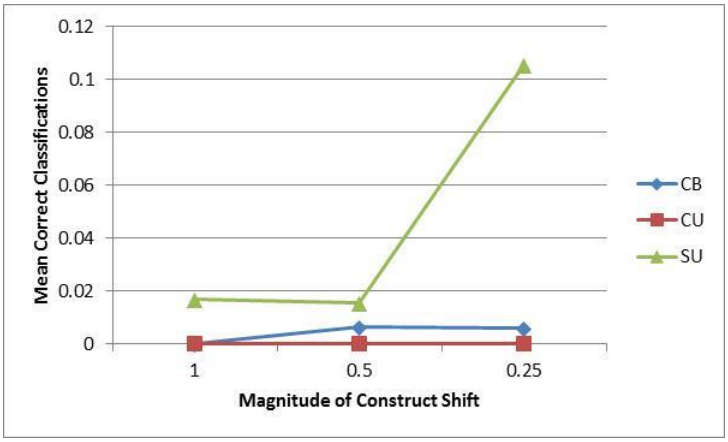

Figure 4.6f Mean Correct Classifications Based on SE for ZP in Year 2

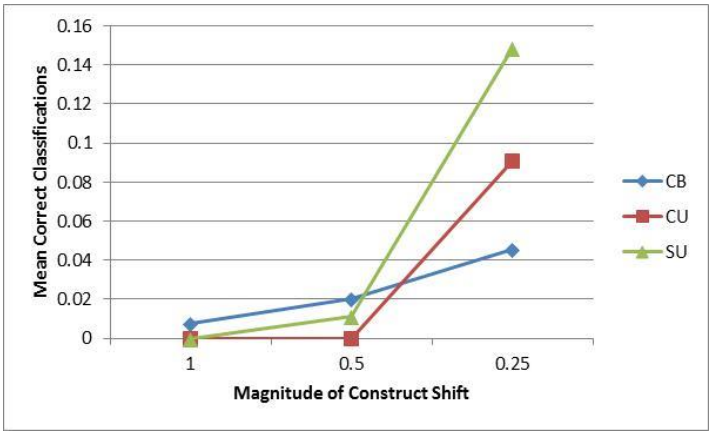

Figure 4.6g Mean Correct Classifications Based on SE for CP in Year 3

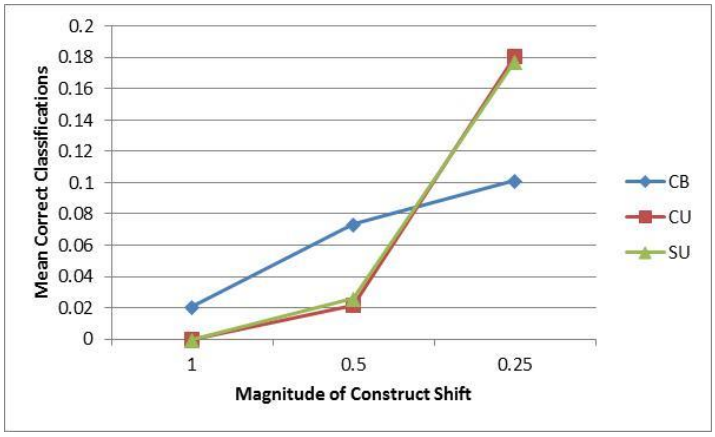

Figure 4.6h Mean Correct Classifications Based on SE for VP in Year 3

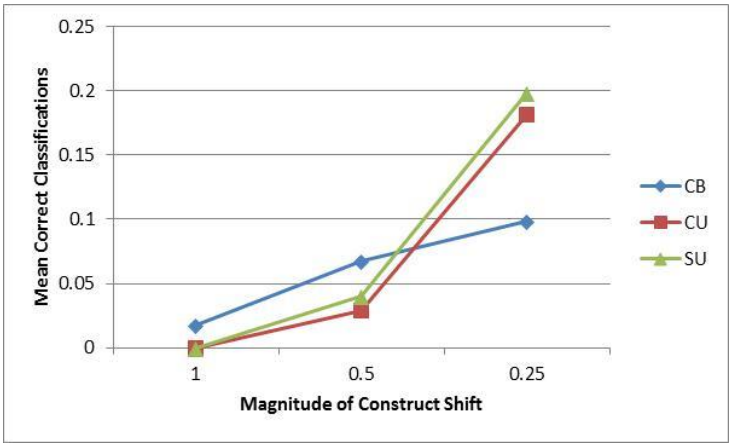

Figure 4.6i Mean Correct Classifications Based on SE for ZP in Year

4.2.2.2 Comparison Based on Quintile Grouping

Table 4.7 presents the number of incorrect and correct classifications based on quintile grouping under the 27 simulation conditions in different years and the corresponding standard deviations in the parenthesis. To compare the CB and SU method, comparison of the number of incorrect classifications needs to be carried out between the first and the third column; to compare the CB and CU method, comparison of the number of incorrect classifications needs to be carried out between the first and the second column. In the first column (CB), the number of incorrect classifications ranges from 3.49 to 12.62 with a mean of 6.39, in the second column (CU) the number of incorrect classifications ranges from 1.94 to 15.28 with a mean of 6.21, and in the third column (SU) the number of incorrect classifications ranges from 1.63 to 10.33 with a mean of 4.86. In terms of the number of correct classifications, comparison needs to be carried out between the fourth and the sixth column and between the fourth and the fifth column. In the fourth column (CB), the number of correct classifications ranges from 19.38 to 28.51 with a mean of 25.61, in the fifth column (CU) the number of correct classifications ranges from 16.72 to 30.06 with a mean of 25.78, and in the sixth column (SU) the number of correct classifications ranges from 21.67 to 30.37 with a mean of 27.14. Across different simulation conditions and different years, the SU method does a better job than the CB method in terms of classification accuracy based on quintile grouping: the mean number of incorrectly classified teachers with the SU method is 1.53 less than that with the CB method, and the mean number of correctly classified teachers with the SU method is 1.53 more than that with the CB method. The CU method does a slightly better job than the CB method in terms of classification accuracy based on quintile

grouping: the mean number of incorrectly classified teachers with the CU method is 0.18 less than that with the CB method, and the mean number of correctly classified teachers with the CU method is 0.17 more than that with the CB method.

To further investigate the difference between the CB and SU method, the comparison is broken down into different persistence patterns in different years. In terms of the number of incorrect classifications, the SU method is consistently lower than the CB method for all the persistence patterns in year 1, with the exception of the VP with variance equal to 1; In year 2 and year 3, the SU method is consistently lower than the CB method, with the exception of the CP and the VP with variance equal to 1. In terms of the number of correct classifications, the same pattern exists: the SU method is consistently higher than the CB method for all the persistence patterns in year 1, with the exception of the VP with variance equal to 1; In year 2 and year 3, the SU method is consistently higher than the CB method, with the exception of the CP and the VP with variance equal to 1.

To further investigate the difference between the CB and CU method, the comparison is broken down into different persistence patterns in different years. In terms of the number of incorrect classifications, it seems that the CB method has higher values than the CU method except for variance equal to 1 for the CP and VP in year 1, and for the ZP this pattern is reversed; for all the persistence patterns in year 2 and year 3, the number of incorrect classifications of the CB method is consistently higher than that of the CU method, except when the variance is equal to 1. In terms of the number of correct classifications, the patterns are similar to those of the number of incorrect classifications: the CB method has lower values than the CU method except when the variance is equal

to 1 for the CP and VP in year 1, and for the ZP this pattern is reversed; for all the models in year 2 and year 3, the number of correct classifications of the CB method is consistently lower than that of the CU method, except when the variance is equal to 1. The differences of both incorrect and correct classification rate between the CB and the CU methods are marginal.

Graphs showing the percentage of incorrect and correct classifications based on quintile grouping for different persistence models in different years are presented in Figures 4.7a through 4.7r.

Table 4.7 Correct and Incorrect Classifications Based on Quintile Grouping

Year	Persistence Pattern	Variance	**Incorrect Classifications** Vertical Scaling Method CB	CU	SU	**Correct Classifications** Vertical Scaling Method CB	CU	SU
Year 1	CP	1	4.47	4.59	2.94	27.53	27.41	29.06
			(0.77)	(0.51)	(0.60)	(0.77)	(0.51)	(0.60)
		0.50	3.70	3.10	3.16	28.30	28.90	28.84
			(0.73)	(0.30)	(0.51)	(0.73)	(0.30)	(0.51)
		0.25	3.87	3.48	1.65	28.13	28.52	30.35
			(0.90)	(0.67)	(0.89)	(0.90)	(0.67)	(0.89)
	VP	1	5.18	6.19	6.14	26.82	25.81	25.86
			(1.12)	(0.72)	(0.70)	(1.12)	(0.72)	(0.70)
		0.50	5.16	4.91	4.97	26.84	27.09	27.03
			(1.06)	(0.88)	(0.89)	(1.06)	(0.88)	(0.89)
		0.25	4.88	4.82	4.74	27.12	27.18	27.26
			(1.32)	(1.00)	(0.88)	(1.32)	(1.00)	(0.88)
	ZP	1	12.62	12.15	7.04	19.38	19.85	24.96
			(1.85)	(1.07)	(0.60)	(1.85)	(1.07)	(0.60)
		0.50	12.17	13.10	6.69	19.83	18.90	25.31
			(2.65)	(1.47)	(0.85)	(2.65)	(1.47)	(0.85)
		0.25	11.96	12.39	5.76	20.04	19.61	26.24
			(2.59)	(1.59)	(0.83)	(2.59)	(1.59)	(0.83)

		1	5.30	6.80	6.26	26.70	25.20	25.74
	CP		(1.59)	(0.90)	(1.02)	(1.59)	(0.90)	(1.02)
		0.50	5.18	4.09	3.70	26.82	27.91	28.30
			(1.28)	(0.59)	(0.76)	(1.28)	(0.59)	(0.76)
Year 2		0.25	3.65	1.94	1.63	28.35	30.06	30.37
			(1.40)	(0.66)	(0.71)	(1.40)	(0.66)	(0.71)
		1	5.41	7.02	7.66	26.59	24.98	24.34
	VP		(1.48)	(0.72)	(0.96)	(1.48)	(0.72)	(0.96)
		0.50	5.01	4.18	4.06	26.99	27.82	27.94
			(1.57)	(0.77)	(0.80)	(1.57)	(0.77)	(0.80)
		0.25	3.71	2.14	1.87	28.29	29.86	30.13
			(1.49)	(0.80)	(0.77)	(1.49)	(0.80)	(0.77)
		1	11.87	15.28	9.74	20.13	16.72	22.26
	ZP		(2.30)	(1.02)	(0.89)	(2.30)	(1.02)	(0.89)
		0.50	6.87	5.67	4.32	25.13	26.33	27.68
			(2.18)	(0.89)	(0.83)	(2.18)	(0.89)	(0.86)
		0.25	8.23	6.05	1.73	23.77	25.95	30.27
			(2.30)	(1.34)	(0.78)	(2.30)	(1.34)	(0.78)
		1	9.37	11.05	10.33	22.63	20.95	21.67
	CP		(1.53)	(0.81)	(1.10)	(1.53)	(0.81)	(1.10)
		0.50	4.70	2.83	2.30	27.30	29.17	29.70
			(1.62)	(0.90)	(0.89)	(1.62)	(0.90)	(0.89)
Year 3		0.25	5.90	4.24	3.73	26.10	27.76	28.27
			(1.93)	(1.34)	(1.06)	(1.93)	(1.34)	(1.06)
		1	7.92	9.87	10.06	24.08	22.13	21.94
	VP		(1.78)	(0.99)	(0.92)	(1.78)	(0.99)	(0.92)
		0.50	3.49	2.21	2.11	28.51	29.79	29.89
			(1.57)	(0.77)	(0.72)	(1.57)	(0.77)	(0.72)
		0.25	4.67	3.37	3.44	27.33	28.63	28.56
			(2.27)	(1.28)	(1.23)	(2.27)	(1.28)	(1.23)
		1	8.56	10.25	9.66	23.44	21.75	22.34
	ZP		(1.77)	(0.94)	(1.03)	(1.77)	(0.94)	(1.03)
		0.50	3.79	2.17	1.87	28.21	29.83	30.13
			(1.55)	(0.77)	(0.79)	(1.55)	(0.77)	(0.79)
		0.25	5.00	3.95	3.54	27.00	28.05	28.46
			(2.04)	(1.18)	(1.22)	(2.04)	(1.18)	(1.22)

* The numbers in the above table are the average number of correctly or incorrectly classified teachers across 100 replications.

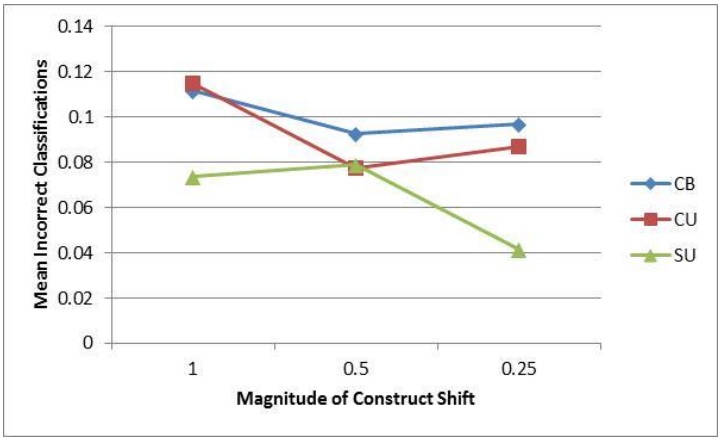

Figure 4.7a Mean Incorrect Classifications Based on Quintile Grouping for CP in Year 1

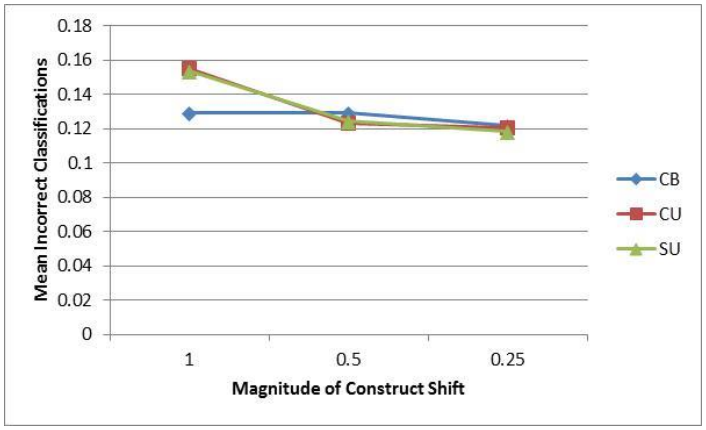

Figure 4.7b Mean Incorrect Classifications Based on Quintile Grouping for VP in Year 1

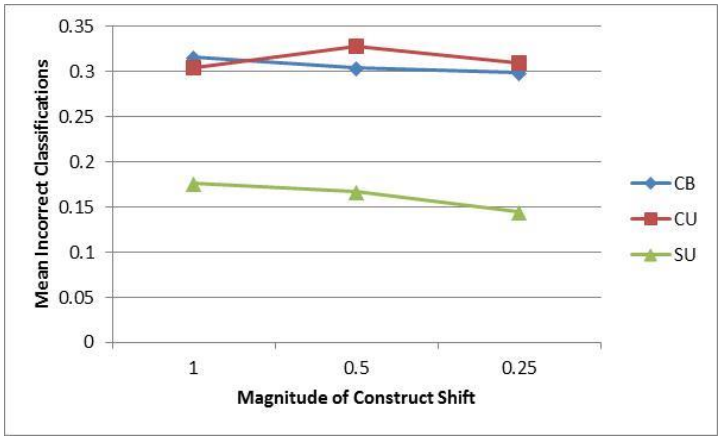

Figure 4.7c Mean Incorrect Classifications Based on Quintile Grouping for ZP in Year 1

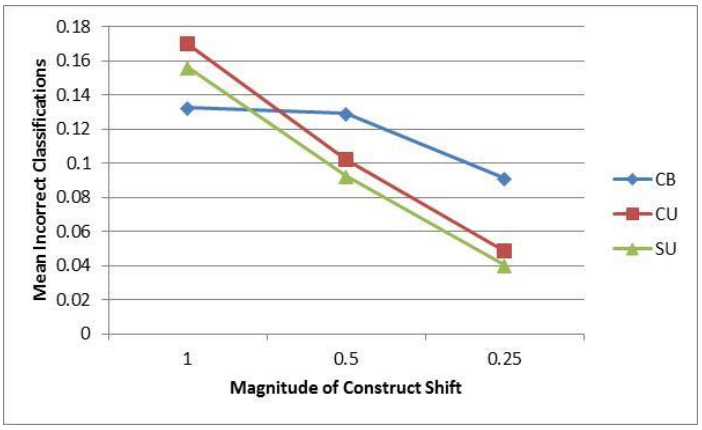

Figure 4.7d Mean Incorrect Classifications Based on Quintile Grouping for CP in Year 2

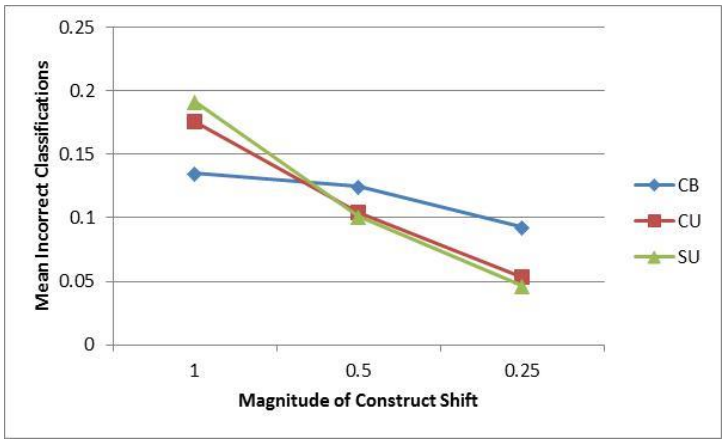

Figure 4.7e Mean Incorrect Classifications Based on Quintile Grouping for VP in Year 2

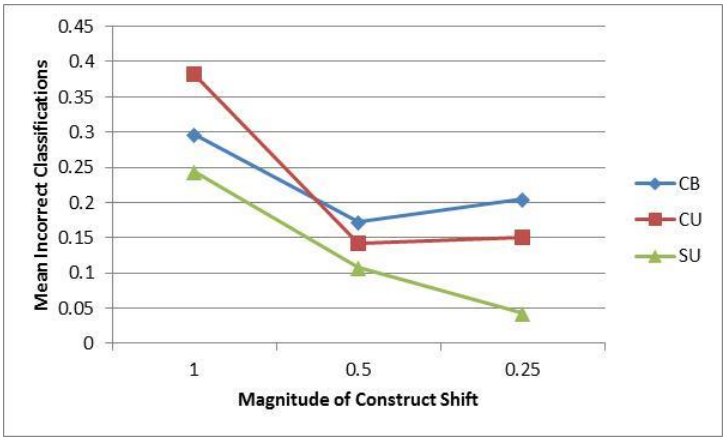

Figure 4.7f Mean Incorrect Classifications Based on Quintile Grouping for ZP in Year 2

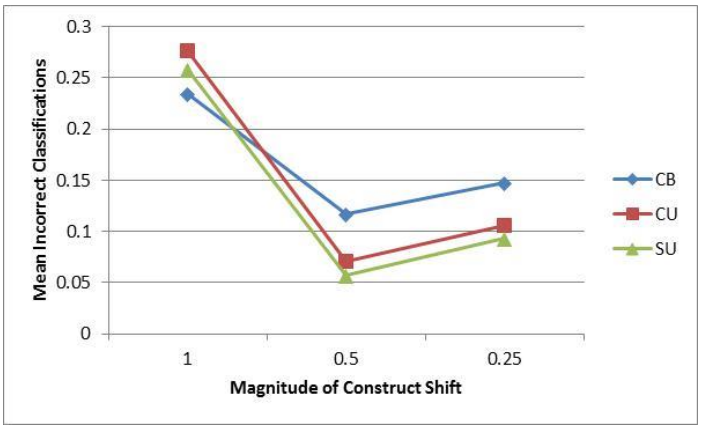

Figure 4.7g Mean Incorrect Classifications Based on Quintile Grouping for CP in Year 3

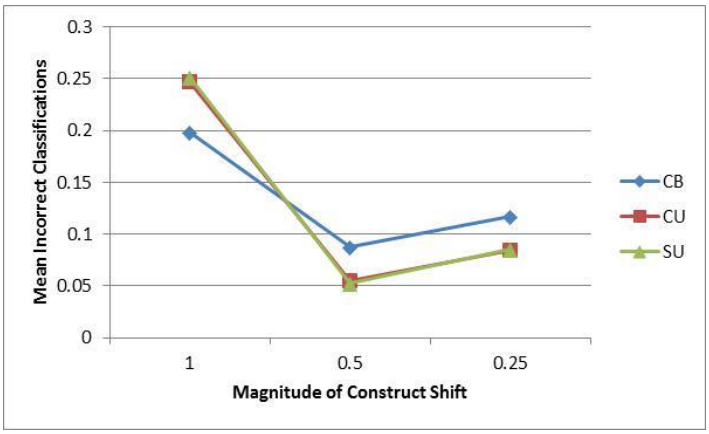

Figure 4.7h Mean Incorrect Classifications Based on Quintile Grouping for VP in Year 3

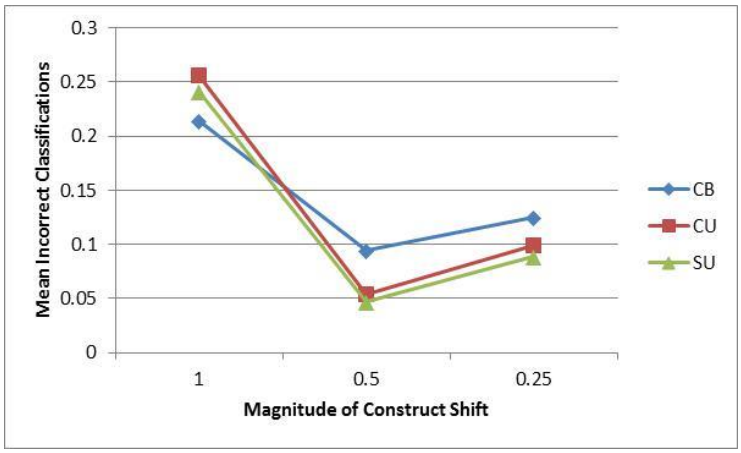

Figure 4.7i Mean Incorrect Classifications Based on Quintile Grouping for ZP in Year 3

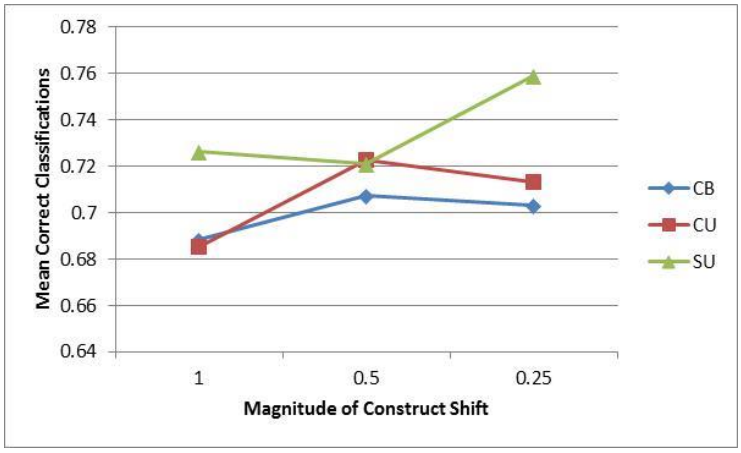

Figure 4.7j Mean Correct Classifications Based on Quintile Grouping for CP in Year 1

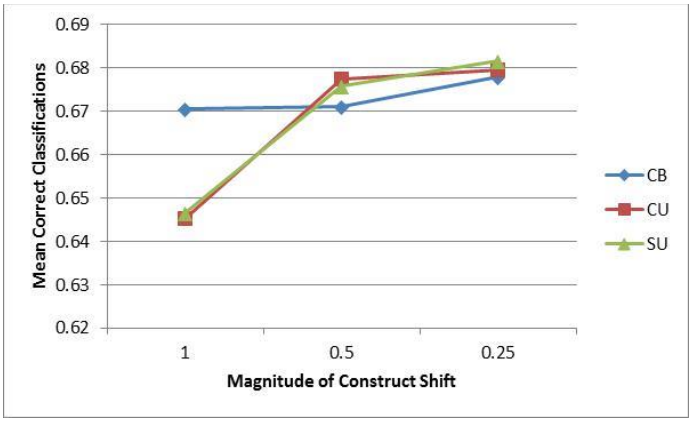

Figure 4.7k Mean Correct Classifications Based on Quintile Grouping for VP in Year 1

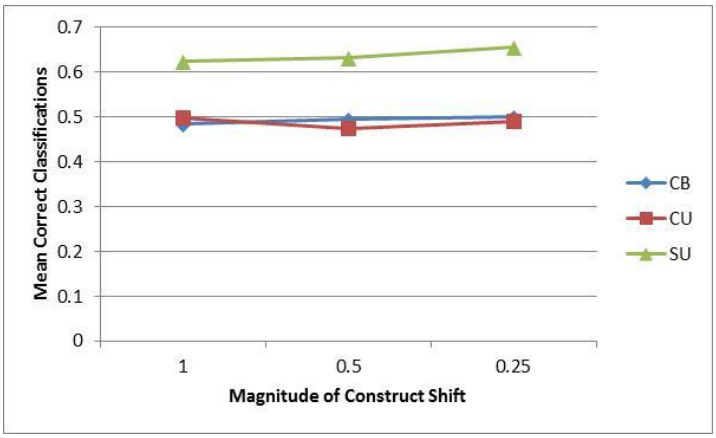

Figure 4.7l Mean Correct Classifications Based on Quintile Grouping for ZP in Year 1

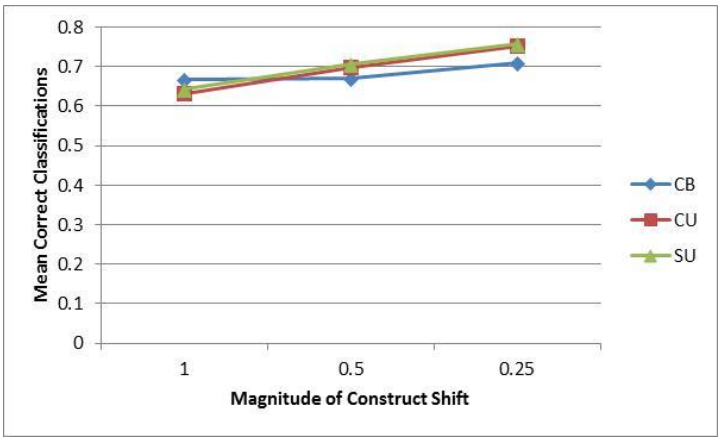

Figure 4.7m Mean Correct Classifications Based on Quintile Grouping for CP in Year 2

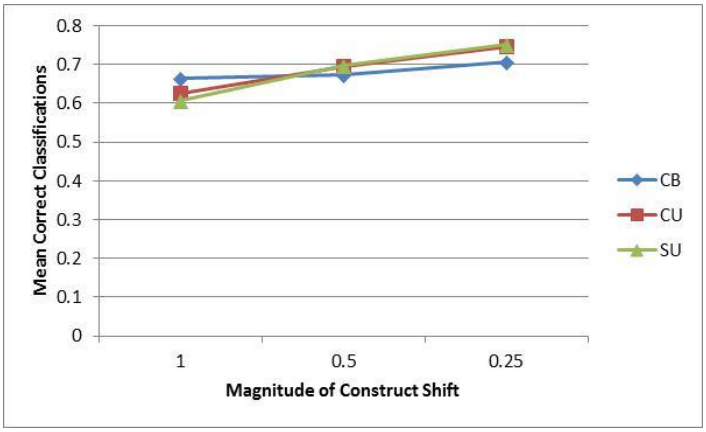

Figure 4.7n Mean Correct Classifications Based on Quintile Grouping for VP in Year 2

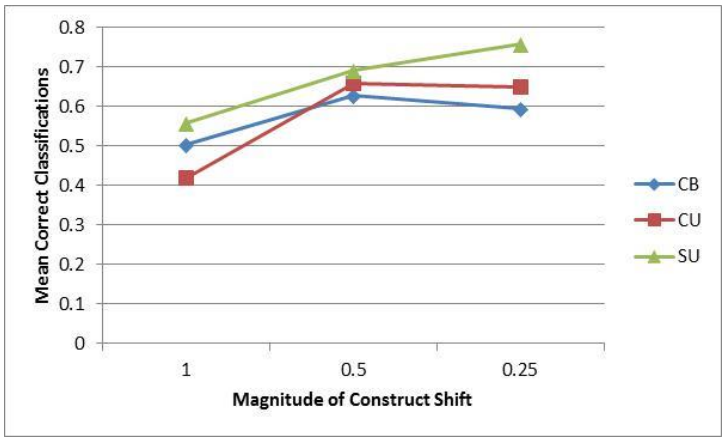

Figure 4.7o Mean Correct Classifications Based on Quintile Grouping for ZP in Year 2

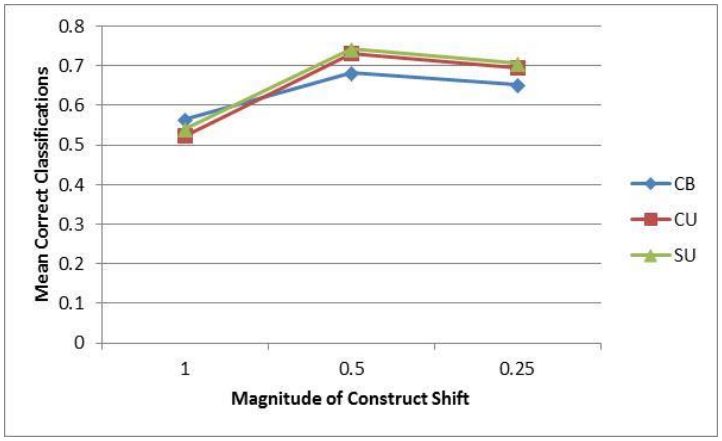

Figure 4.7p Mean Correct Classifications Based on Quintile Grouping for CP in Year 3

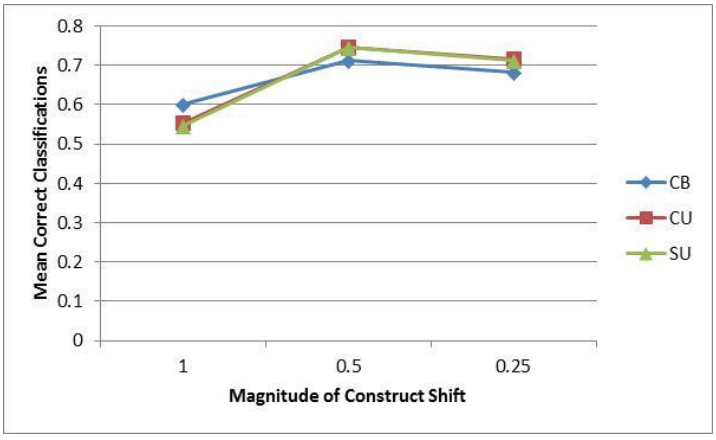

Figure 4.7q Mean Correct Classifications Based on Quintile Grouping for VP in Year 3

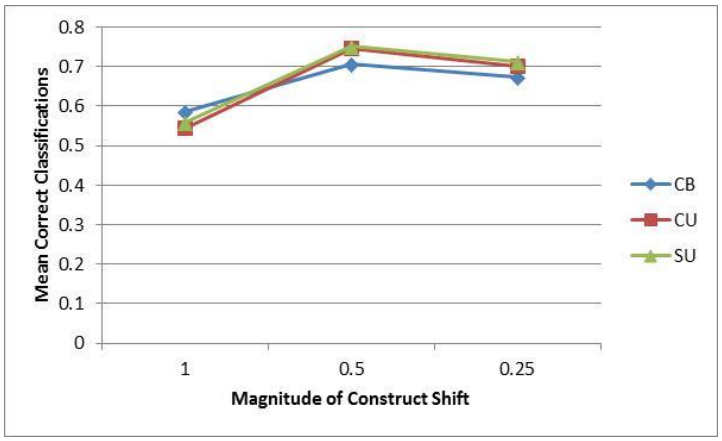

Figure 4.7r Mean Correct Classifications Based on Quintile Grouping for ZP in Year 3

4.2.2.3 Comparison Based on Tercile Grouping

Table 4.8 presents the number of incorrect and correct classifications based on tercile grouping under the 27 simulation conditions in different years and the corresponding standard deviations in the parenthesis. To compare the CB and SU method, comparison of the number of incorrect classifications needs to be carried out between the first and the third column; to compare the CB and CU method, comparison of the number of incorrect classifications needs to be carried out between the first and the second column. In the first column (CB), the number of incorrect classifications ranges from 1.9 to 9.75 with a mean of 4.32, in the second column (CU) the number of incorrect classifications ranges from 0.9 to 11.76 with a mean of 4.15, and in the third column (SU) the number of incorrect classifications ranges from 0.31 to 7.24 with a mean of 2.87. In terms of the number of correct classifications, comparison needs to be carried out between the fourth and the sixth column and between the fourth and the fifth column. In the fourth column (CB), the number of correct classifications ranges from 16.25 to 24.1 with a mean of 21.68, in the fifth column (CU) the number of correct classifications ranges from 14.04 to 25.1 with a mean of 21.85, and in the sixth column (SU) the number of correct classifications ranges from 18.76 to 25.69 with a mean of 23.10. Across different simulation conditions and different years it seems that the SU method does a better job than the CB method in terms of classification accuracy based on tercile grouping: the mean number of incorrectly classified teachers with the SU method is 1.45 less than that with the CB method, and the mean number of correctly classified teachers with the SU method is 1.42 more than that with the CB method. The CU method does a better job than the CB method in terms of classification accuracy based on tercile

grouping: the mean number of incorrectly classified teachers with the SU method is 0.17 less than that with the CB method, and the mean number of correctly classified teachers with the SU method is 0.17 more than that with the CB method.

To further investigate the difference between the CB and SU method, the comparison is broken down into different persistence patterns in different years. In terms of the number of incorrect classifications, the SU method is consistently lower than the CB method for all the persistence patterns in year 1, with the exception of VP with variance equal to 1 and 0.5; In year 2 and year 3, the SU method is consistently lower than the CB method, with the exception with variance equal to 1. In terms of the number of correct classifications, the same pattern exists: the SU method is consistently higher than the CB method for all the persistence patterns in year 1, with the exception of the VP model with variance equal to 1 and 0.5; In year 2 and year 3, the SU method is consistently lower than the CB method, with the exception with variance equal to 1.

To further investigate the difference between the CB and CU method, the comparison is broken down into different persistence patterns in different years. In terms of the number of incorrect classifications, it seems that in year 1 the CB method has higher values than the CU method except for variance equal to 0.5 for CP, and for VP the CB method has higher values than the CU method only when the variance is equal to 0.25, and for ZP the CB method has higher values than the CU method only when the variance is equal to 1; for all the models in year 2 and year 3, the number of incorrect classifications of the CB method is consistently higher than that of the CU method, except when the variance is equal to 1. In terms of the number of correct classifications, the patterns are similar to those of the number of incorrect classifications: in year 1 for

CP, the CB method has lower values than the CU method except for variance equal to 0.5, and for VP the CB method has lower values than the CU method only when the variance is equal to 0.25, and for ZP the CB method has lower values than the CU method only when the variance is equal to 1; for all the models in year 2 and year 3, the number of correct classifications of the CB method is consistently lower than that of the CU method, except when the variance is equal to 1. The differences of both incorrect and correct classifications between the CB and the CU methods, as can be seen in Table 4.8, are marginal.

Graphs showing the percentage of incorrect and correct classifications based on tercile grouping for different persistence models in different years are presented in Figures 4.8a through 4.8r.

Table 4.8 Correct and Incorrect Classifications Based on Tercile Grouping

Year	Persistence Pattern	Variance	**Incorrect Classifications** Scoring Method			**Correct Classifications** Scoring Method		
			CB	CU	SU	CB	CU	SU
Year 1	CP	1	3.08	2.96	1.12	22.92	23.04	24.88
			(0.93)	(0.57)	(0.36)	(0.93)	(0.57)	(0.36)
		0.50	2.75	2.99	1.66	23.25	23.01	24.34
			(0.72)	(0.10)	(0.81)	(0.72)	(0.10)	(0.81)
		0.25	2.11	1.66	0.31	23.89	24.34	25.69
			(0.98)	(0.65)	(0.49)	(0.98)	(0.65)	(0.49)
	VP	1	3.48	3.90	3.91	22.52	22.10	22.09
			(1.11)	(0.73)	(0.71)	(1.11)	(0.73)	(0.71)
		0.50	3.26	3.34	3.32	22.74	22.66	22.68
			(0.85)	(0.48)	(0.47)	(0.85)	(0.48)	(0.47)
		0.25	2.70	1.85	2.39	23.30	24.15	23.61
			(1.11)	(0.77)	(0.76)	(1.11)	(0.77)	(0.76)
	ZP	1	9.75	9.73	5.48	16.25	16.27	20.52
			(1.76)	(1.02)	(0.54)	(1.76)	(1.02)	(0.54)
		0.50	9.25	9.92	3.95	16.75	16.08	22.05
			(2.30)	(1.20)	(0.56)	(2.30)	(1.20)	(0.56)
		0.25	9.37	9.66	4.00	16.63	16.34	22.00
			(2.28)	(1.22)	(0.49)	(2.28)	(1.22)	(0.49)

		1	2.98	3.73	3.10	23.02	22.27	22.90
	CP		(1.11)	(0.85)	(1.10)	(1.11)	(0.85)	(1.10)
		0.50	3.11	2.49	2.08	22.89	23.51	23.92
			(0.98)	(0.86)	(0.91)	(0.98)	(0.86)	(0.91)
Year 2		0.25	2.04	0.98	0.81	23.96	25.02	25.19
			(0.97)	(0.70)	(0.49)	(0.97)	(0.70)	(0.49)
		1	3.01	3.36	4.23	22.99	22.64	21.77
	VP		(1.26)	(0.77)	(1.19)	(1.26)	(0.77)	(1.19)
		0.50	3.23	2.66	2.62	22.77	23.34	23.38
			(1.22)	(0.68)	(0.72)	(1.22)	(0.68)	(0.72)
		0.25	2.13	1.12	0.90	23.87	24.88	25.10
			(1.13)	(0.52)	(0.50)	(1.13)	(0.52)	(0.50)
		1	9.04	11.96	6.02	16.96	14.04	19.98
	ZP		(2.24)	(1.09)	(0.78)	(2.24)	(1.09)	(0.78)
		0.50	4.63	3.78	2.62	21.37	22.22	23.38
			(1.77)	(0.97)	(0.68)	(1.77)	(0.97)	(0.68)
		0.25	5.93	4.31	0.82	20.07	21.69	25.18
			(2.09)	(1.08)	(0.64)	(2.09)	(1.08)	(0.64)
		1	6.82	8.36	7.24	19.18	17.64	18.76
	CP		(1.37)	(0.81)	(0.93)	(1.37)	(0.81)	(0.93)
		0.50	2.75	1.57	1.13	23.25	24.43	24.87
			(1.47)	(0.81)	(0.73)	(1.47)	(0.81)	(0.73)
Year 3		0.25	3.96	2.05	1.74	22.04	23.95	24.26
			(1.67)	(0.95)	(1.00)	(1.67)	(0.95)	(1.00)
		1	5.64	7.11	7.03	20.36	18.89	18.97
	VP		(1.43)	(0.94)	(1.07)	(1.43)	(0.94)	(1.07)
		0.50	1.90	0.90	0.82	24.10	25.10	25.18
			(1.15)	(0.67)	(0.69)	(1.15)	(0.67)	(0.69)
		0.25	2.77	1.56	1.53	23.23	24.44	24.47
			(1.56)	(1.01)	(1.03)	(1.56)	(1.01)	(1.03)
		1	6.02	7.38	6.89	19.98	18.62	19.11
	ZP		(1.55)	(0.80)	(0.93)	(1.55)	(0.80)	(0.93)
		0.50	2.04	0.97	0.79	23.96	25.03	25.21
			(1.30)	(0.69)	(0.64)	(1.30)	(0.69)	(0.64)
		0.25	2.89	1.87	1.80	23.11	24.13	24.20
			(1.46)	(1.00)	(0.94)	(1.46)	(1.00)	(0.94)

* The numbers in the above table are the average number of correctly or incorrectly classified teachers across 100 replications.

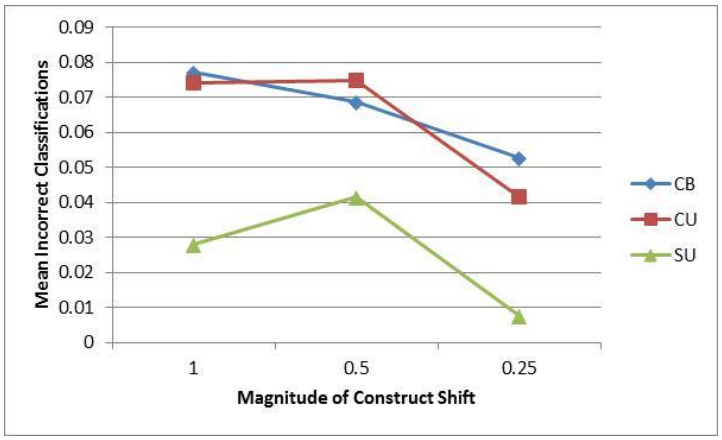

Figure 4.8a Mean Incorrect Classifications Based on Tercile Grouping for CP in Year 1

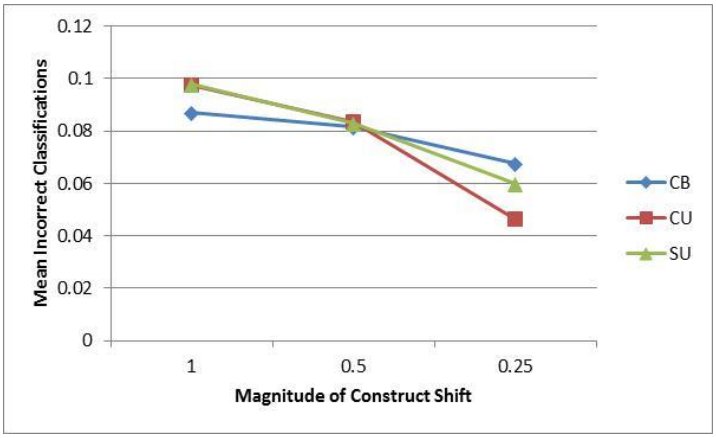

Figure 4.8b Mean Incorrect Classifications Based on Tercile Grouping for VP in Year 1

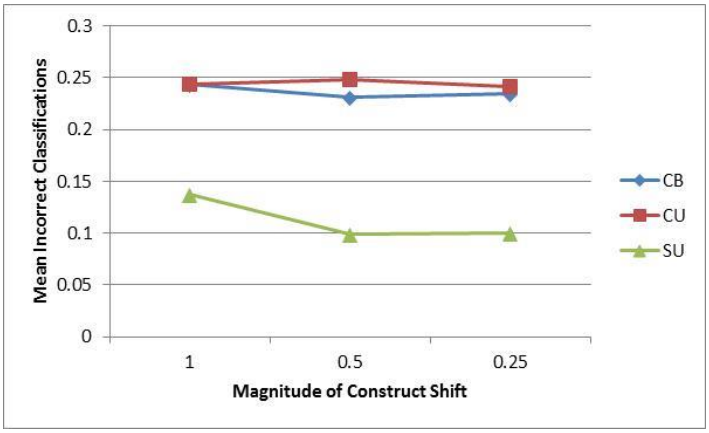

Figure 4.8c Mean Incorrect Classifications Based on Tercile Grouping for ZP in Year 1

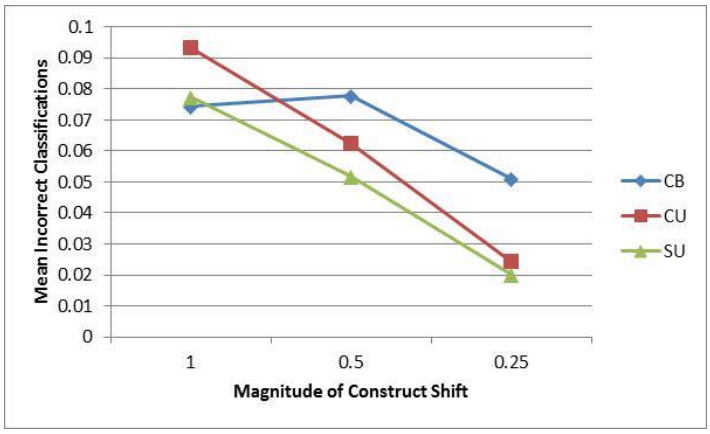

Figure 4.8d Mean Incorrect Classifications Based on Tercile Grouping for CP in Year 2

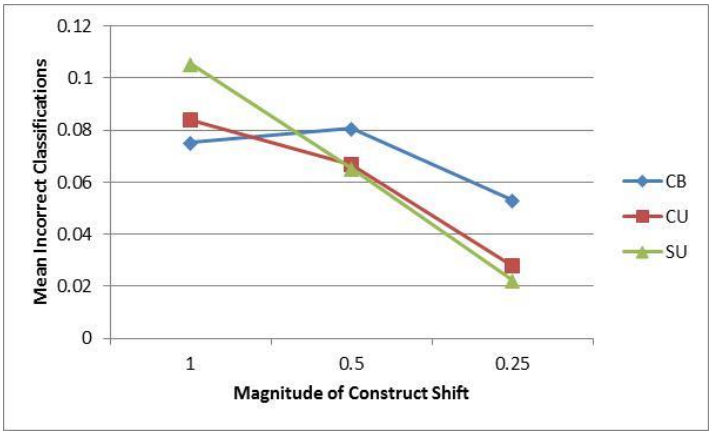

Figure 4.8e Mean Incorrect Classifications Based on Tercile Grouping for VP in Year 2

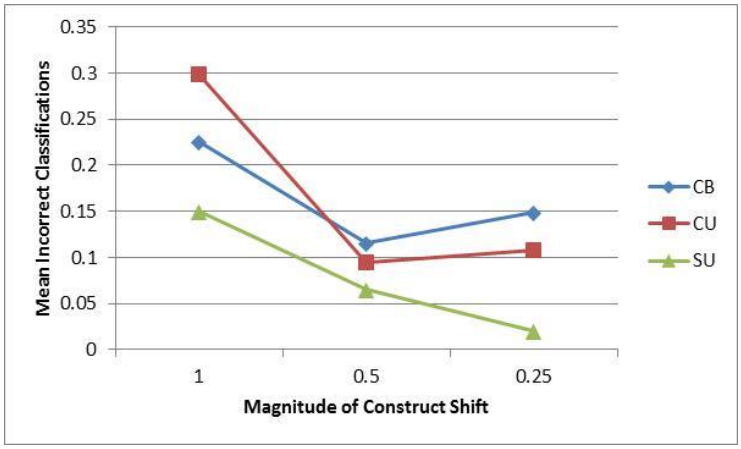

Figure 4.8f Mean Incorrect Classifications Based on Tercile Grouping for ZP in Year 2

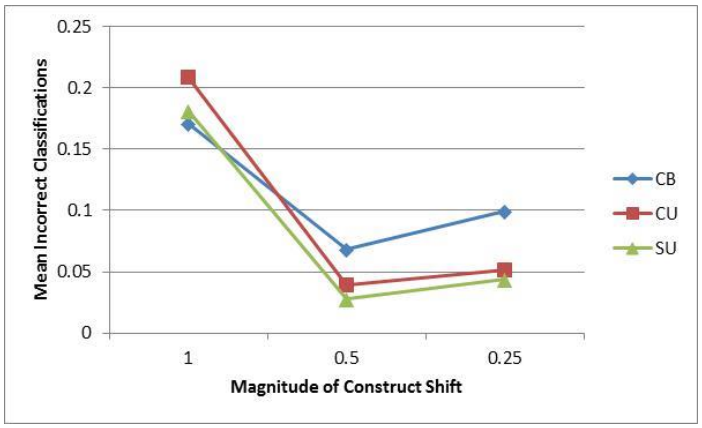

Figure 4.8g Mean Incorrect Classifications Based on Tercile Grouping for CP in Year 3

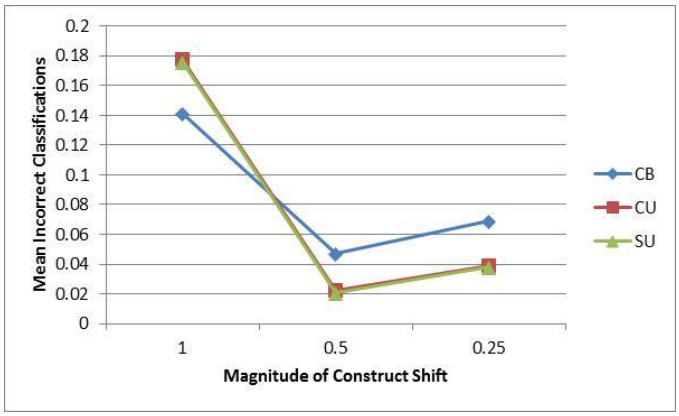

Figure 4.8h Mean Incorrect Classifications Based on Tercile Grouping for VP in Year 3

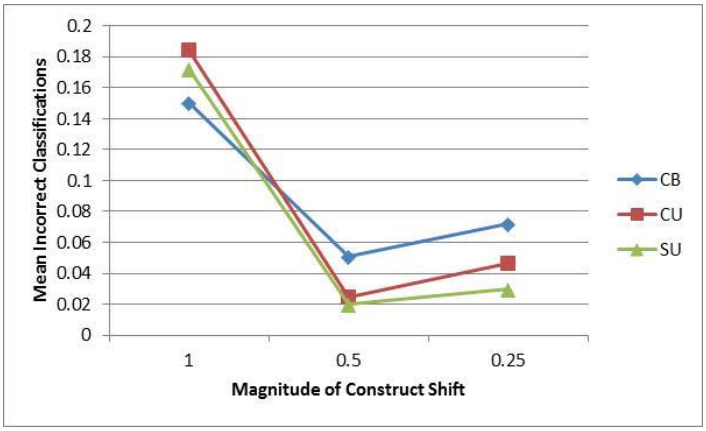

Figure 4.8i Mean Incorrect Classifications Based on Tercile Grouping for ZP in Year 3

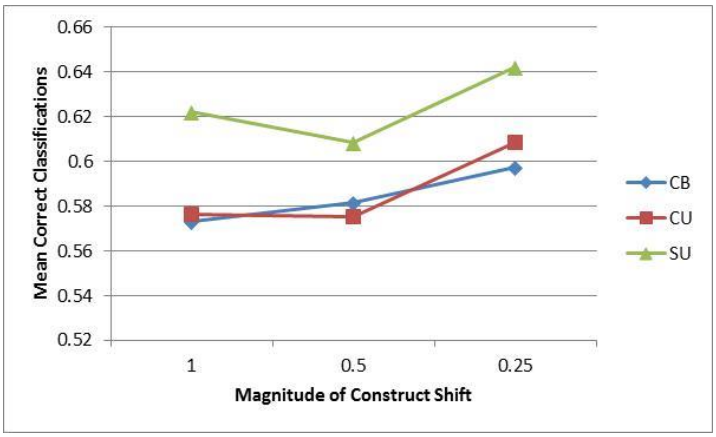

Figure 4.8j Mean Correct Classifications Based on Tercile Grouping for CP in Year 1

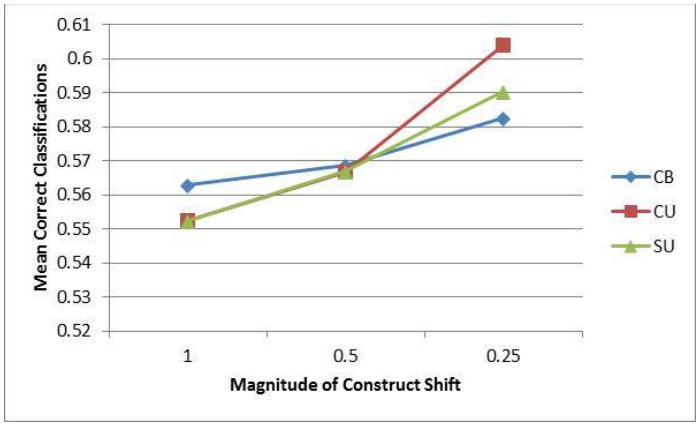

Figure 4.8k Mean Correct Classifications Based on Tercile Grouping for VP in Year 1

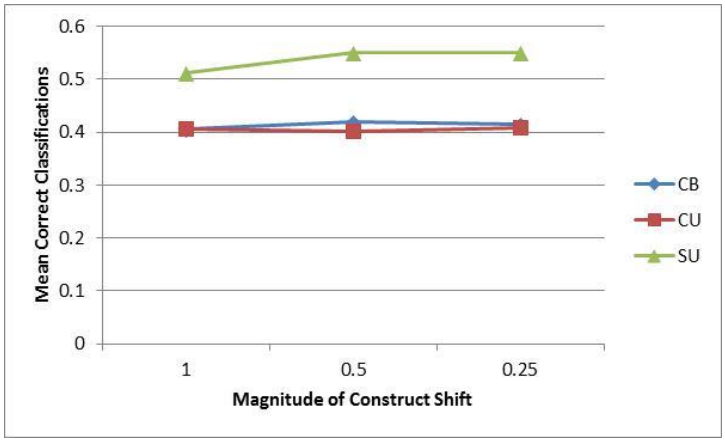

Figure 4.8l Mean Correct Classifications Based on Tercile Grouping for ZP in Year 1

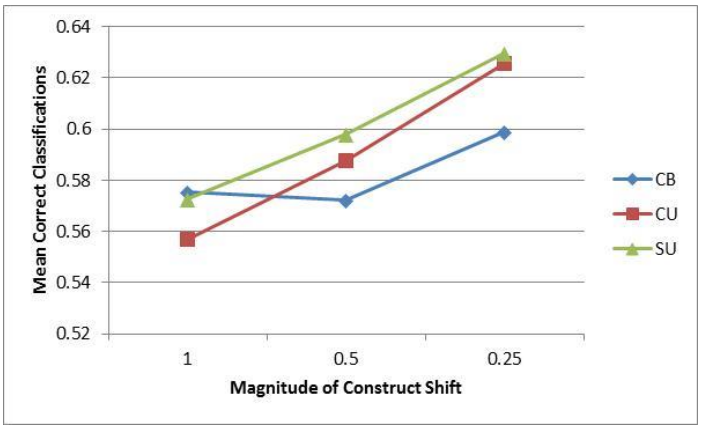

Figure 4.8m Mean Correct Classifications Based on Tercile Grouping for CP in Year 2

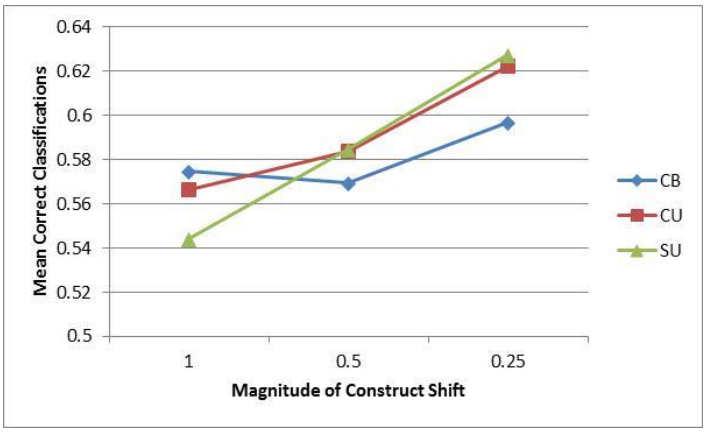

Figure 4.8n Mean Correct Classifications Based on Tercile Grouping for VP in Year 2

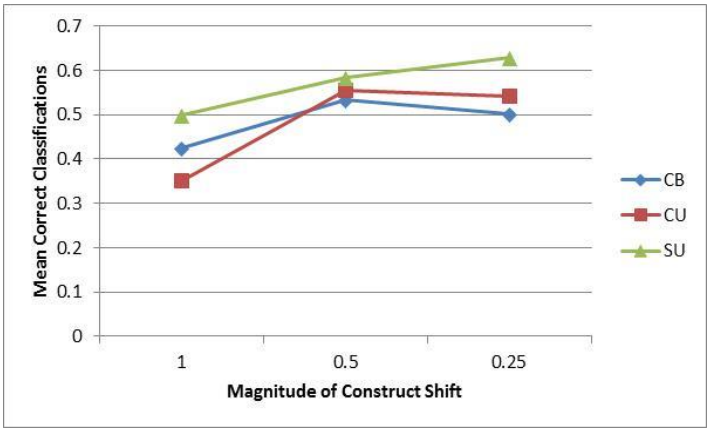

Figure 4.8o Mean Correct Classifications Based on Tercile Grouping for ZP in Year 2

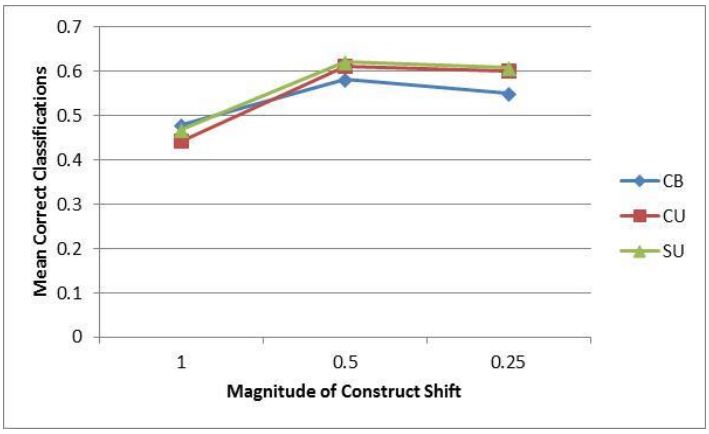

Figure 4.8p Mean Correct Classifications Based on Tercile Grouping for CP in Year 3

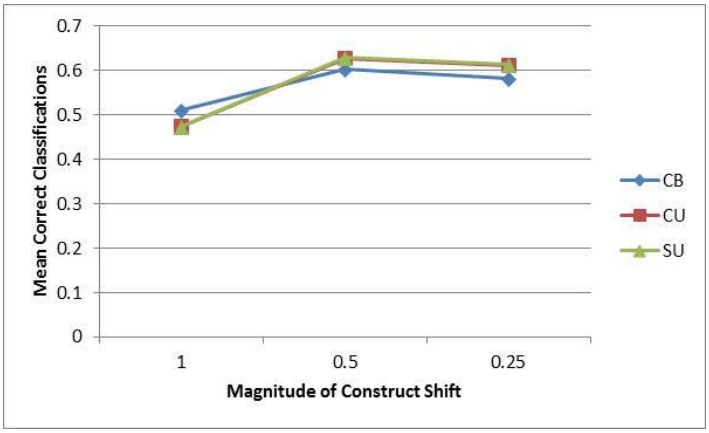

Figure 4.8q Mean Correct Classifications Based on Tercile Grouping for VP in Year 3

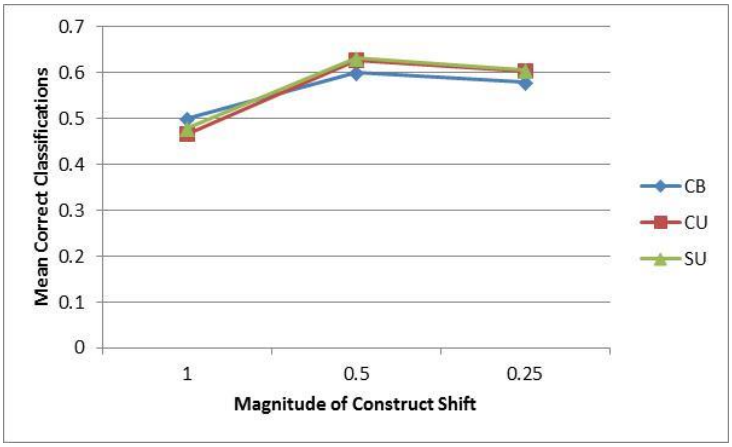

Figure 4.8r Mean Correct Classifications Based on Tercile Grouping for ZP in Year 3

4.3 Test of Between-Subject Effects (ANOVA)

It should be noted that assumptions of factorial ANOVA were checked and it was found that the normality assumption might be violated: some of the skewness and kurtosis values are significantly different from 0, although most of them are within the range of (-1, 1). It is determined that violation of normality assumption in this study does not pose serious threats to validity of the ANOVA results based on the following rationales:

1. The range of (-1, 1) of the skewness and kurtosis values does not constitute as an extreme departure from normality.
2. Considering the sample size within each cell (n=100), normality can be approximately assumed based on the central limit theorem.
3. The balanced design further alleviates the concern of threats to internal validity caused by violation of normality assumption.

Therefore, no efforts were made to transform the outcome variables. The results of three-way analysis of variance (ANOVA) of correlation and correct classifications for the three manipulated factors are presented in Tables 4.9, 4.10, 4.11, and 4.12. The p values and eta squared values are reported in the tables, and only the effects of the manipulated factors with significant p values ($p<.05$) and practical significance ($eta^2>.05$) are discussed.

Table 4.9 Tests of Between-subject Effects on Correlation between Teacher Effect Estimates and the Generating Values

Correlation	Year 1		Year 2		Year 3	
	p-value	eta^2	p-value	eta^2	p-value	eta^2
Persistence Pattern	**0.000**	**0.7816**	**0.000**	**0.4280**	0.000	0.0189
Variance	0.000	0.0069	**0.000**	**0.2796**	**0.000**	**0.9132**
Method	**0.000**	**0.0826**	**0.000**	**0.0604**	0.000	0.0017
P*V	0.000	0.0111	**0.000**	**0.0946**	0.000	0.0013
P*M	**0.000**	**0.1048**	**0.000**	**0.0835**	0.000	0.0022
V*M	0.000	0.0058	0.000	0.0260	**0.000**	**0.0614**
P*V*M	0.000	0.0072	0.000	0.0280	0.000	0.0012

Table 4.9 summarizes the ANOVA results when the correlation is compared. In terms of main effects, persistence pattern affects the correlations in year 1 and year 2 significantly and accounts for 78.16% (large effect) of the total variance in year 1 and 42.80% (large effect) of that in year 2; choice of vertical scaling methods also affects the correlation in year 1 and year 2 significantly and accounts for 8.26% (medium effect) of the total variance in year 1 and 6.04% (small effect) of that in year 2; magnitude of construct shift, which is represented by variance in the table, affects the correlation in year 2 and year 3 significantly and accounts for 27.96% (large effect) of the total variance in year 2 and 91.32% (large effect) of that in year 3. In terms of interaction effects, the combination of persistence pattern and vertical scaling methods affects the correlation in year 1 and year 2 significantly and accounts for 10.24% (medium effect) of the total variance in year 1 and 8.35% (medium effect) of that in year 2; the combination of persistence pattern and magnitude of construct shift affects the correlation in year 2 significantly and accounts for 9.46% (medium effect) of the total variance; the combination of magnitude of construct shift and vertical scaling methods method affects

the correlation in year 3 significantly and accounts for 6.14% (small effect) of the total variance.

Table 4.10 Tests of Between-subject Effects on Correct Classifications Based on Standard Error

Correct Classification	Year 1		Year 2		Year 3	
	p-value	eta^2	p-value	eta^2	p-value	eta^2
Persistence Pattern	**0.000**	**0.8954**	**0.000**	**0.3979**	**0.000**	**0.0564**
Variance	0.000	0.0378	**0.000**	**0.3341**	**0.000**	**0.7536**
Method	0.000	0.0009	0.000	0.0058	0.000	0.0114
P*V	0.000	0.0187	**0.000**	**0.0988**	0.000	0.0295
P*M	0.000	0.0202	0.000	0.0442	0.000	0.0055
V*M	0.000	0.0147	**0.000**	**0.0832**	**0.000**	**0.1353**
P*V*M	0.000	0.0123	0.000	0.0362	0.000	0.0083

Table 4.10 summarizes the ANOVA results when the number of correct classifications based on standard error is compared. In terms of main effect, persistence pattern affects the correct classifications in all three years significantly and accounts for 89.54% (large effect) of the total variance in year 1, 39.79% (large effect) of that in year 2, and 5.64% (small effect) of that in year 3; magnitude of construct shift affects the correct classifications in year 2 and year 3 significantly and accounts for 33.41% (large effect) of the total variance in year 2 and 75.36% (large effect) of that in year 3. In terms of interaction effect, the combination of persistence model and magnitude of construct shift affects the correct classifications in year 2 significantly and accounts for 9.88% (medium effect) of the total variance; the combination of magnitude of construct shift and vertical scaling methods affects the correct classifications in year 2 and year

3significantly and accounts for 8.32% (medium effect) of the total variance in year 2 and 13.53% (medium effect) of that in year 3.

Table 4.11 Tests of Between-subject Effects on Correct Classifications Based on Quintile Grouping

Correct Classification	Year 1		Year 2		Year 3	
	p-value	eta^2	p-value	eta^2	p-value	eta^2
Persistence Patten	**0.000**	**0.7478**	**0.000**	**0.2601**	0.000	0.0127
Variance	0.000	0.0106	**0.000**	**0.4554**	**0.000**	**0.9205**
Method	**0.000**	**0.1057**	**0.000**	**0.0514**	0.000	0.0088
P*V	0.000	0.0021	**0.000**	**0.0873**	0.082	0.0006
P*M	**0.000**	**0.1251**	**0.000**	**0.0616**	0.000	0.0034
V*M	0.000	0.0024	**0.000**	**0.0664**	**0.000**	**0.0535**
P*V*M	0.000	0.0063	0.000	0.0177	0.434	0.0005

Table 4.11 summarizes the ANOVA results when the number of incorrect classifications based on quintile grouping is compared. In terms of main effect, persistence pattern affects the correct classifications in year 1 and year 2 significantly and accounts for 74.58% (large effect) of the total variance in year 1 and 26.01% (large effect) of that in year 2; magnitude of construct shift affects the correct classifications in year 2 and year 3 significantly and accounts for 45.54% (large effect) of the total variance in year 2 and 92.05% (large effect) of that in year 3; vertical scaling method affects the correct classifications in year 1 and year 2 significantly and accounts for 10.57% (medium effect) of the total variance in year 1 and 5.14% (small effect) of that in year 2. In terms of interaction effect, the combination of persistence pattern and magnitude of construct shift affects the correct classifications in year 2 significantly and accounts for

8.73% (medium effect) of the total variance; the combination of persistence pattern and vertical scaling method affects the correct classifications in year 1 and year 2 significantly and accounts for 12.51% (medium effect) of the total variance in year 1 and 6.16% (small effect) of that in year 2; the combination of magnitude of construct shift and vertical scaling methods affects the correct classifications in year 2 and year 3 significantly and accounts for 6.64% (small effect) of the total variance in year 2 and 5.35% (small effect) of that in year 3.

Table 4.12

Tests of Between-subject Effects on Correct Classifications Based on Tercile Grouping

Correct Classification	Year 1		Year 2		Year 3	
	p-value	eta^2	p-value	eta^2	p-value	eta^2
Persistence Pattern	**0.000**	**0.7251**	**0.000**	**0.3333**	0.000	0.0143
Variance	0.000	0.0212	**0.000**	**0.2957**	**0.000**	**0.9201**
Method	**0.000**	**0.1205**	**0.000**	**0.0680**	0.000	0.0109
P*V	0.000	0.0054	**0.000**	**0.1408**	0.060	0.0006
P*M	**0.000**	**0.1210**	**0.000**	**0.0873**	0.000	0.0034
V*M	0.000	0.0020	0.000	0.0411	0.000	0.0493
P*V*M	0.000	0.0047	0.000	0.0338	0.010	0.0014

Table 4.12 summarizes the ANOVA results when the number of incorrect classifications based on tercile grouping is compared. In terms of main effect, persistence pattern affects the correct classifications in year 1 and year 2 significantly and accounts for 72.51% (large effect) of the total variance in year 1 and 33.33% (large effect) of that in year 2; magnitude of construct shift affects the correct classifications in year 2 and year 3 significantly and accounts for 29.57% (large effect) of the total variance in year 2 and

92.01% (large effect) of that in year 3; choice of vertical scaling methods affects the correct classifications in year 1 and year 2 significantly and accounts for 12.05% (medium effect) of the total variance in year 1 and 6.80% (small effect) of that in year 2. In terms of interaction effect, the combination of persistence pattern and magnitude of construct shift affects the correct classifications in year 2 significantly and accounts for 14.08% (medium effect) of the total variance; the combination of persistence pattern and vertical scaling methods affects the correct classifications in year 1 and year 2 significantly and accounts for 12.10% (medium effect) of the total variance in year 1 and 8.73% (medium effect) of that in year 2.

4.4 Summary of the Main Findings

The accuracy of the SU method is largely influenced by persistence pattern, variance, and time. In terms of the correlation between the teacher effect estimates and the generating values, CP produced the highest correlation with a mean value of 0.80 across different variances and different years while ZP leads to the lowest with a mean value of 0.67. With the decrease of the magnitude of construct shift, the correlation for all models tends to increase, with the most dramatic increase occurring from variance 1.0 to variance 0.5 in year 3.

In terms of incorrect classifications based on standard errors, the SU method performs extremely well: in year 2 and year 3 not a single teacher is incorrectly classified; in year 1, incorrect classification occurs with VP and ZP at variance 1.0, and it should be noted that the values are negligibly small. In terms of correct classifications based on

standard errors, the performance of the SU method depends on persistence pattern, variance, and time. In year 1 and year 2, number of correct classifications seems to decrease with the decrease of the persistence of teacher effect in the sense that CP tends to perform the best and ZP the worst; in year 3, this pattern seems to be reversed. Another finding is that in year 2 and year 3, with the decrease of magnitude of construct shift the SU method tends to correctly classify more teachers, and when the variance is equal to 1, it performs poorly.

Classifications based on quintile and tercile groupings have similar patterns, and are summarized as follows. The number of incorrect classifications decreases with the decrease of the magnitude of construct shift regardless of the persistence pattern and year; in year 1 and year 2 the number of classifications decreases with the increase of the persistence of teacher effect, and in year 3 this persistence has no noticeable impact upon the number of incorrect classifications. In terms of correct classifications, while the persistence of teacher effect seems not to be influential, the general pattern is that the correct classification rate increases with the decrease of the magnitude of construct shift regardless of the persistence pattern and year.

Comparison of the three vertical scaling methods is conducted and there are several findings. The first one is that the SU method tends to perform better than both the CB and the CU methods in most simulation conditions. For some conditions, the CB method performs better, although the difference tends to be negligible. The next finding is that despite the fact that the SU method is devised to tackle the issue of construct shift, its performance is still influenced by the magnitude of construct shift. The third finding is that the CB method does not perform considerably better than the CU method as expected.

When the magnitude of construct shift is large (variance = 1), it tends to perform marginally better than the CU method; when the variance is smaller than 1, the CU method tends to be marginally better. It should be noted that the differences between those two methods are rather small and it is reasonable to conclude that they perform similarly. The last main finding is that when classification is the purpose, the standard error based approach misclassifies almost no teachers at the expense of correctly classifying considerably fewer teachers.

CHAPTER 5: REAL DATA ANALYSIS

Mariano et al. (2010) fit the GP model in a real data set and found that while the GP model had the best model fit among all persistence models, the proximal year teacher effect estimates between the GP model and the VP model were extremely highly correlated. Therefore, they concluded that for that particular data set they used, choosing the GP model over the VP model might not make a difference in terms of teacher effect estimation. Aware of the fact that their conclusions were based on a single data set, they suggested that other real data sets should be used to investigate the generalizability of their findings. This chapter addresses this question and provides another example of applying the GP model to a real data set. Section 5.1describes the data set, section 5.2 lists the specific questions related to this data set, and section 5.3 provides the results and the analysis.

5.1 Data

The data were three years of math scores (2008, 2009, and 2010) on a state achievement test from grade 3 to grade 8. This test is not vertically scaled, which aligns well with Mario et al.'s suggestion of evaluating the performance of the GP model with test data that does not have a development scale. In the dataset there were four cohorts: cohort1 (grade 3 through grade 5), cohort2 (grade 4 through grade 6), cohort3 (grade 5 through grade 7), and cohort 4 (grade 6 through grade 8). Table 5.1 summarizes the sample size of each cohort:

Table 5.1 Cohort Sample Size

	Year		
Cohort	2008	2009	2010
Cohort1	7246	7336	7273
Cohort2	7251	7337	7107
Cohort3	7321	7095	7052
Cohort4	7374	7282	7201

One common phenomenon of longitudinal data is missing data, and the data set used in this study was no exception. If the missing data issue is ignored, the sample size becomes 8522 for cohort 1, 8610 for cohort2, 8656 for cohort3, and 8617 for cohort4. If the observations with missing data are deleted, the sample size becomes 6074 for cohort 1, 5850 for cohort2, 5842 for cohort3, and 6089 for cohort4. Considering that from about 42% to 80% of students have at least one year of missing score out of four to five year testing (McCaffrey & Lockwood, 2011), approximately 30% of students having missing data out of three year testing in this data is not surprising. Moreover, the GP model is flexible enough to accommodate the missing data issue and those observations with missing scores need not to be deleted. Table 5.2 lists the descriptive statistics of each cohort's score at each of the three years:

Table 5.2 Descriptive Statistics of Scores

Cohort	Year	Mean	SD	Min	Max
Cohort1	2008	421	38	310	585
	2009	431	40	297	650
	2010	433	36	309	650
Cohort2	2008	430	41	317	584
	2009	432	38	329	650
	2010	427	35	335	650
Cohort3	2008	430	38	327	589
	2009	423	36	240	650
	2010	422	36	321	568
Cohort4	2008	428	36	314	566
	2009	424	35	309	650
	2010	429	35	320	572

5.2 Research Questions

Mariano et al. (2010) fit the GP model to an empirical data set, which contains vertically scaled mathematic test scores of a cohort of students progressing from grade 1 to grade 5 in academic years 1997-1998 through 2001-2002. They obtained three main findings:

1. The correlation between proximal year effects and future year effects is about 0.5 to 0.6, while the correlations among the future year effects are higher than 0.9.

2. With the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), the GP model is the best fitting model comparing to the CP model and VP model.

3. Despite the best mode fit provided by the GP model, estimates of the proximal year effects from the GP model and the VP model are extremely highly correlated, which makes the computational intensity of the GP model not worthwhile if the interest is only in the estimate of the proximal year effects. In contrast, the choice between the VP model and the CP model has a much greater impact upon the teacher effect estimates.

Based on their findings, the corresponding research questions are:

1. Are the correlation values between proximal year effects and future year effects and among the future year effects in the current data set similar to the values Mariano et al. (2010) found?
2. Does the GP model provide the best model fit to the current data set comparing to the CP, VP, and ZP model?
3. How does the proximal year teacher effects of the GP model compare to those of the CP, VP, and CP model in the current data set?

5.3 Results

The R package GPvam is used to answer each of the above research questions.

5.3.1 Research Questions 1

Tables 5.3a to 5.3d present the correlation values between the proximal year effects and the future year effects and between the future year effects for the four cohorts. The correlation between the proximal year effects and future year effects listed in the above four tables are all above 0.84, which are much higher than those values found by Mariano et al. (2010). The correlation among future year effects is all above 0.99, which is consistent with their finding that those values are above 0.9.

Table 5.3a Correlation in Cohort1

Year	Teacher Effect	Year 1			Year 2	
		Proximal	Future 1	Future 2	Proximal	Future 1
Year 1	Proximal	1				
	Future 1	.988	1			
	Future 2	.990	.999	1		
Year 2	Proximal				1	
	Future 1				.896	1

Table 5.3b Correlation in Cohort 2

Year	Teacher Effect	Year 1			Year 2	
		Proximal	Future 1	Future 2	Proximal	Future 1
Year 1	Proximal	1				
	Future 1	.995	1			
	Future 2	.986	.998	1		
Year 2	Proximal				1	
	Future 1				.846	1

Table 5.3c Correlation in Cohort 3

Year	Teacher Effect	Year 1			Year 2	
		Proximal	Future 1	Future 2	Proximal	Future 1
Year 1	Proximal	1				
	Future 1	.984	1			
	Future 2	.981	.999	1		
Year 2	Proximal				1	
	Future 1				.977	1

Table 5.3d Correlation in Cohort 4

Year	Teacher Effect	Year 1			Year 2	
		Proximal	Future 1	Future 2	Proximal	Future 1
Year 1	Proximal	1				
	Future 1	.998	1			
	Future 2	.996	.999	1		
Year 2	Proximal				1	
	Future 1				.957	1

5.3.2 Research Questions 2

In order to compare different persistence models, Mariano et al. (2010) used DIC, a model fit index often seen in hierarchical Bayesian models. When models are estimated with maximum likelihood estimation, the Akaike Information Criterion (AIC) is often used to compare model fit, and the GPvam package also uses this index. AIC is used as the model fit index in this study. Table 5.4 lists the AIC values for all the persistence models fit in each of the four cohorts, and the minimum value within each cohort is bolded. The pattern is similar to Mariano et al.'s finding that the GP model provides the best model fit comparing to the other persistence models.

Table 5.4 AIC for Different Models in Different Cohorts

Cohort	Model			
	ZP	VP	CP	GP
Cohort 1	202556.3	199101.6	199649.4	**198905.5**
Cohort 2	201260.1	197560	198140.3	**197275.5**
Cohort 3	195955.7	193154.8	193716.1	**192795.5**
Cohort 4	196928.4	194391.5	194731.8	**94292.5**

5.3.3 Research Questions 3

Tables 5.5a to 5.5d present the correlation among the proximal year teacher effect estimation of four different persistence models for the four cohorts. Consistent with Mariano et al.'s finding, the correlation between the VP and the GP models is extremely high regardless of the year, and the correlation between the CP and the VP model seem to be less in year 2 and year 3. An interesting finding is that in year 1, the teacher effect estimates of the CP and the VP model seem to be also extremely highly correlated.

Table 5.5a Correlation among Teacher Effects Estimates of Different Models in Cohort 1

Year		Year 1				Year 2				Year 3			
	Model	CP	VP	ZP	GP	CP	VP	ZP	GP	CP	VP	ZP	GP
	CP	1											
Year	VP	.997	1										
1	ZP	.893	.902	1									
	GP	.996	.999	.912	1								
Year	CP					1							
2	VP					.945	1						
	ZP					.518	.739	1					
	GP					.886	.976	.799	1				
Year	CP									1			
3	VP									.844	1		
	ZP									.322	.749	1	
	GP									.776	.974	.812	1

Table 5.5b Correlation among Teacher Effects Estimates of Different Models in Cohort 2

Year		Year 1				Year 2				Year 3			
	Model	CP	VP	ZP	GP	CP	VP	ZP	GP	CP	VP	ZP	GP
	CP	1											
Year	VP	.996	1										
1	ZP	.801	.824	1									
	GP	.994	.999	.834	1								
Year	CP					1							
2	VP					.868	1						
	ZP					.225	.645	1					
	GP					.749	.951	.762	1				
Year	CP									1			
3	VP									.888	1		
	ZP									.560	.852	1	
	GP									.883	.986	.855	1

Table 5.5c Correlation among Teacher Effects Estimates of Different Models in Cohort 3

Year		Year 1				Year 2				Year 3			
	Model	CP	VP	ZP	GP	CP	VP	ZP	GP	CP	VP	ZP	GP
	CP	1											
Year	VP	.991	1										
1	ZP	.848	.890	1									
	GP	.989	.999	.895	1								
Year	CP					1							
2	VP					.911	1						
	ZP					.650	.887	1					
	GP					.909	.988	.880	1				
Year	CP									1			
3	VP									.913	1		
	ZP									.668	.889	1	
	GP									.917	.993	.879	1

Table 5.5d Correlation among Teacher Effects Estimates of Different Models in Cohort 4

Year		Year 1				Year 2				Year 3			
	Model	CP	VP	ZP	GP	CP	VP	ZP	GP	CP	VP	ZP	GP
	CP	1											
Year	VP	.995	1										
1	ZP	.798	.824	1									
	GP	.994	.999	.830	1								
Year	CP					1							
2	VP					.937	1						
	ZP					.609	.822	1					
	GP					.925	.992	.827	1				
Year	CP									1			
3	VP									.855	1		
	ZP									.443	.815	1	
	GP									.832	.989	.831	1

CHAPTER 6: DISCUSSION

In this final chapter, the main findings of the simulation study and the real data analysis are summarized in section 6.1 and discussed in section 6.2. In section 6.3 the implications for practice of teacher evaluation are presented. Section 6.4 focuses on the limitations of the current study and suggestions for future studies.

6.1 Summary of Findings

6.1.1 Accuracy of the SU Method

The accuracy of the SU method was investigated using the Spearman correlation between the teacher estimates and the true values and teacher classification accuracy. The Spearman correlation value ranges from 0.47 to 0.91, and is influenced by the choice of persistence pattern and magnitude of construct shift. CP has a mean correlation of 0.80, VP has a mean correlation of 0.74, and ZP has the lowest correlation value with a mean of 0.67. With the increase of the persistence parameter, the correlation increases. The mean correlation value is 0.61 when variance is equal to 1, 0.79 when the variance is equal to 0.5, and 0.81 when the variance is equal to 0.25. With the increase of the magnitude of construct shift, the correlation value decreases. This is somewhat surprising considering that the GP model used in the SU method was devised to deal with the issue of construct shift.

The number of incorrect classifications based on standard errors is negligibly small regardless of the model and variance. The number of correct classification based on standard errors ranges from 0 to 13.56, and is influenced by the choice of persistence

pattern and magnitude of construct shift. For CP the mean of correct classifications is 6.78, for VP it is 3.95, and for ZP it is 2.14. The mean of correct classifications is 2.83 when the variance is 1, 3.04 when then variance is 0.5, and 7 when the variance is 0.25. As with the Spearman correlation, the number of correct classifications is also influenced by the choice of persistence pattern and magnitude of construct shift.

The number of incorrect classifications based on quintile grouping ranges from 1.63 to 10.33, and the number of correct classifications ranges from 21.94 to 30.37. For CP the mean of incorrect classifications is 3.97 and the mean of correct classifications is 28.03; for VP the mean of incorrect classifications is 5.01 and the mean of correct classifications is 26.99; for ZP the mean of incorrect classifications is 5.59 and the mean of correct classifications is 26.41.

The number of incorrect classifications based on tercile grouping ranges from 0.31 to 7.24, and the number of correction classifications ranges from 18.76 to 25.69. For CP the mean of incorrect classifications is 2.13 and the mean of correct classifications is 23.87; for VP the mean of incorrect classifications is 2.97 and the mean of correct classifications is 23.03; for ZP the mean of incorrect classifications is 3.53 and the mean of correct classifications is 22.40.

To sum up, the classification accuracy heavily depends on the specific approach used for classification. When the standard error based approach is used, the chance of incorrect classification seems to be negligible, although it is at the expense of correct classifications. In most cases the number of correct classifications is below 10, and in extreme cases not a single teacher is correctly classified. When quintile or tercile

grouping are used as the classification method, the correct classification rate is much higher with on average more than 20 teachers correctly classified, which comes with the price of increased number of incorrectly classified teachers. In extreme cases, the number of incorrect classifications can be higher than 10.

6.1.2 Comparison of Different Methods

The comparison of different methods was also conducted using the Spearman correlation between the teacher estimates and the true values and teacher classification accuracy. For the CB method, the Spearman correlation value ranges from 0.24 to 0.83 with a mean of 0.66; for the CU method, it ranges from 0.05 to 0.87 with a mean of 0.66; for the SU method, it ranges from 0.47 to 0.91 with a mean of 0.74.

In terms of classification accuracy based on standard errors, for the CB method, the number of incorrect classifications ranges from 0 to 0.31 with a mean of 0.05, and the number of correct classifications ranges from 0.01 to 11.81 with a mean of 3.80; for the CU method, the number of incorrect classifications ranges from 0 to 0.30 with a mean of 0.01, and the number of correct classifications ranges from 0 to 11.92 with a mean of 3.93; for the SU method, the number of incorrect classifications ranges from 0 to 0.19 with a mean of 0.01, and the number of correct classifications ranges from 0 to 13.56 with a mean of 4.29.

In terms of classification accuracy based on quintile grouping, for the CB method, the number of incorrect classifications ranges from 3.49 to 12.62 with a mean of 6.39, and the number of correct classifications ranges from 19.38 to 28.51 with a mean of

25.61; for the CU method, the number of incorrect classifications ranges from 1.94 to 15.28 with a mean of 6.21, and the number of correct classifications ranges from 16.72 to 30.06 with a mean of 25.78; for the SU method, the number of incorrect classifications ranges from 1.63 to 10.33 with a mean of 4.86, and the number of correct classifications ranges from 21.67 to 30.37 with a mean of 27.14.

In terms of classification accuracy based on tercile grouping, for the CB method, the number of incorrect classifications ranges from 1.9 to 9.75 with a mean of 4.32, and the number of correct classifications ranges from 16.25 to 24.1 with a mean of 21.68; for the CU method, the number of incorrect classifications ranges from 0.9 to 11.76 with a mean of 4.15, and the number of correct classifications ranges from 14.04 to 25.1 with a mean of 21.85; for the SU method, the number of incorrect classifications ranges from 0.31 to 7.24 with a mean of 2.87, and the number of correct classifications ranges from 18.76 to 25.69 with a mean of 23.10.

Overall, the CB and the CU methods perform similarly, while the SU method is consistently superior to them. The finding that the SU method is superior to the CB method is fairly surprising considering that the CB method does not involve model mis-specification and therefore was expected to be superior. It turns out that only in year 3 with variance equal to 1 is the CB method better than the SU method. An even more surprising finding is that the CB and the CU method perform similarly: when the variance is 1 the CB method is consistently better than the CU method, while the CU method is better than or equal to the CB method when the variance is 0.5 or 0.25. The CB method was expected to be superior to the CU method since the only difference between those

two methods is model misspecification, and the subsequent VAM analysis is identical between them.

6.2 Discussion

6.2.1 Correct Model Specification vs. Post Hoc Adjustment

Correct model specification is one of the key assumptions of statistical analysis. It is often expected that when the parameter estimates out of a correctly specified model should be more accurate than those out of an incorrectly specified model. Li's finding (2011) confirmed that when the generating model is a bifactor one, the parameter estimates out of the bifactor model are more accurate than the parameter estimates out of the UIRT model, which basically specifies the model due to its ignorance of the multidimensional structure. It had been expected that the more accurately estimated parameters in the bifactor model should translate into more accurate estimation of teacher effect, even though Li's study focuses on parameter recovery while in VAM, the emphasis is usually on the rank ordering of teachers and their classifications. In other words, the expectation was that the bifactor model based teacher effect estimates should have higher correlation values with the generating values than those estimates based on the UIRT model, and they should have lower incorrect classification rates and higher correct classification rates.

However, this expectation is only correct when the variance is equal to 1, a scenario considered with a large magnitude of construct shift. The difference between the CB method based estimates and the CU method based estimates are small, except at year

2 with the ZP model. When the variance decreases, the estimates based on those two methods become similar and in some cases, the UIRT model even performs slightly better than the bifactor model. This counterintuitive finding seems to contradict Li's finding (2012) that the parameter estimates out of the UIRT model is always more inaccurate than those out of the bifactor model. It is believed this inconsistency is probably caused by different focuses of Li's study and the current one: Li's study focuses on parameter recovery while the emphasis of this study is the recovery of the rank ordering of teachers. Another possible explanation is that since the bifactor model is more complicated, it is difficult to separate the effects from the general and secondary factors when the variance of the secondary factor is smaller. This difficulty might introduce more estimator errors.

The SU method is similar to the CU method in the sense that it also ignores the multidimensional structure of the data and hence mis-specifies the model. Such misspecification is compensated by an ad hoc procedure in the GP model that relaxes the perfect correlation assumption of teacher effects across years. Such an adjustment seems effective since the SU method outperforms the CB method in most of the simulation conditions, except when the magnitude of construct shift is large (variance = 1).

6.2.2 Practicality of Each Method

The CU method represents an approach commonly employed in practice, in which a UIRT model is assumed to fit the test data across grades and hence used for calibration with a common item nonequivalent groups design. While easy to implement, this method can be fairly inaccurate when the magnitude of construct shift is large. The most extreme scenario is at year 2 with ZP model, when the CB based teacher effect estimates has a

correlation value of 0.05 with the generating value, and an average of 15.28% teachers are incorrectly classified.

The CB method represents an ideal scenario in which the estimating model and the generating model are the same and no model misspecification is involved. In this study a simulation is used and the true model is known, while in practice, researchers do not have such luxury and the true model is never known. Considering the arbitrary and somewhat unlikely assumption of the CB method that students only grow on a common dimension, it is hard to believe that the bifactor model assumed in the CB method is a good approximation of the reality. More importantly, this study shows that even when the bifacotor model is the correct model, the CB method still does not have a clear advantage over other methods. In addition, the implementation of the CB method can only be done in IRTPRO and the setup is quite cumbersome, which further restricts its applicability.

The SU method is probably superior to the other two methods considering both the performance and applicability. With the R package GPvam available, its implementation is as easy as the CU method and it gives the best performance overall among three methods. Although estimation of fitting the GP model with the SU method is considerably longer than fitting other persistence models with the other two methods, the SU method does not involve vertical scaling and therefore avoids the issues and problems inherent in vertical scaling.

6.2.3 Teacher Effect Persistence Pattern Matters

Based on the results of ANOVA at the end of Chapter 4, teacher effect persistence pattern accounts for a substantial portion of the overall variance. When the persistence

pattern is CP, which means that teacher effects persist undiminished into the future years, on average the accuracy of teacher effect estimation is the best. When the persistence pattern is ZP, which means that the teacher effects do not persist, the teacher effect estimation is the worst. While in this study the true persistence pattern is always specified, it is believed that the incorrect specification of the true persistence pattern would impact the teacher effect estimation. While it is always desirable to fit the data with the corresponding persistence model that aligns with the persistence pattern, in reality the persistence pattern is unknown. Therefore, the ZP model or the CP models are not recommended to use in practice unless there is strong evidence that ZP or CP persistence pattern exist. Instead, the more versatile VP model should be used since the ZP model and the CP model are special cases of the VP model. Following the same vein of reasoning, probably the GP model should be considered an ideal choice. However, the GP model and the VP model seem to produce extremely similar results, as shown in Chapter 5 as well as in Mariano's study (2010), but computation time for the GP model is considerably longer than that for the VP model.

6.2.4 How Much Should We Trust VAM

For VAM to be used for high-stake decisions, it is necessary to show in simulation studies that accurate teacher effect estimates can be obtained which ultimately support accurate teacher classifications. Even if the results of simulation studies suggest so, however, the researcher should realize that simulation studies are idealized in many aspects and the results should be treated with caution if they were ever used to support high-stake applications.

In the current simulation study, the other inherent issues in VAM such as non-random assignment of students and correct VAM model specification are assumed not to exist and the only focus is how the construct shift impacts the VAM results. In this case, it is concluded that the best performing SU method performs satisfactorily: the spearman correlation between the estimated teacher effects and the generating values has a mean of 0.74, although in some cases it can be as low as 0.47 and the average number of incorrect classifications can be as high as 10.06.

The classification accuracy varies depending on which method is used. With the standard error method almost no incorrect teacher classifications occur, although the number of correct teacher classifications can be as low as 0 or close to 0. In other words, when the standard error is used to classify teachers, most of the teachers are classified into the middle group with very few or none in the effective or ineffective group. When quintile grouping or tercile grouping are used, while the correct teacher classifications increase considerably so that more than half of the teacher are correctly classified in most simulation conditions, the incorrect teacher classifications also increases and in the extreme cases more than 10 teachers are incorrectly classified.

Considering that this simulation study represents a nearly ideal scenario in which the inherent issues in VAM such as non-random assignment of students, correct VAM model specification, and missing data or missing teacher student linkage are assumed not to exist, it might be argued that the above results are not particularly promising. Even if the results are considerably better and the accuracy of the parameter estimation is much higher, it can still be argued that in an empirical data set with non-random assignment of students, correct VAM model specification, and missing data, the accuracy of the best

performing SU method would be expected to deteriorate and it would become hard to generalize the findings. With the current findings, it is reasonable to expect that the results from real data are going to be less reliable than what was found in this study and extreme caution has to be exercised if those results will ever, if not never, be the sole basis in high-stake decision making process.

However, it is argued that VAM should not held responsible when it is used as a single measure to evaluate something as complicated as teaching. There is an increasing realization that teaching might be too complex to be measured accurately by any single measure, as concluded in a research brief (2012) of the famous three-year study known as the Measures of Effective Teaching (MET). Sponsored by the Bill & Melinda Gates Foundation, MET was designed (random assignment of participating teachers) to investigate how to systematically evaluate teachers using multiple measures, which include the student achievement gains, classroom observations, and student surveys. One of its key findings is that an index using the weighted sum of those three measures is superior to the student achievement gains alone on almost every dimension including predictive power, reliability, and stability, and it is concluded that this index supports high-stake decisions based upon it.

To sum up, VAM may not perform impressively when it is used as the single measure to evaluate teaching. Teaching is a complex process requiring multiple measures, and VAM, when properly combined with other measures, performs better than any single measure alone.

6.3 Limitations and Directions for Future Research

When comparing the GP model with other persistence models that require vertical scaling, only one vertical scaling design (common item design with concurrent calibration) was used in this study. In a more comprehensive simulation study, various vertical scaling designs with different combination of data collection, calibration method, and percentage of common items can be used to investigate their performance in VAM in comparison to the GP model. Due to the scope limitation, only the current design was implemented.

In terms of IRT models, only the 2PL model was investigated in this study. Future studies can extend to 1PL model and 3PL model to compare their performance with the 2PL model. In terms of item type, only dichotomous items were used in the current study. To better mimic tests used in really testing programs, the mixture of dichotomous and polytomous items can be used. Future studies can manipulate the proportion of dichotomous and polytomous items to investigate its impact upon the teacher effect estimation.

In terms of sample size, this study assumed that the class is mid-sized and each class has 25 students, which is an ideal scenario. Future studies can manipulate the number of students in each class to include small-sized and large-sized classes, and the constraint of equal class sizes can be relaxed to have a mix of small-sized, middle-sized, and large-sized classes. Since the class size is directly related to the standard error of the teacher effect estimate of the particular teacher in charge of the class, it is possible that the change of sample size may present a different picture in terms of the classification accuracy when the standard error based approach is used.

Absence of missing data is another limitation, which is ubiquitous in each possible real data set. Although it was deliberately chosen to avoid the possible confounding effects of missing data with those of construct shift, in future studies missing data can be introduced to make the results more generalizable.

Another possible extension in future studies is the inclusion of covariates. In practice when VAM is used to evaluate teachers, covariates at the student level and the school level are usually used. This study did not include covariates out of the same concern of possible confounding effects. The estimation of persistence based VAM models with covariates should not be a problem since the R package GPvam conveniently offers such capabilities, although the computation time may be longer.

Probably the main limitation of this study is the use of the bifactor model to generate scenarios of construct shift. As mentioned in chapter 2, the bifactor model is an innovative and convenient framework to generate different magnitudes of construct shift, and vertical scaling method is available in this framework. However, the conceptualization that teachers only cause students to grow on the general dimension and the orthogonality assumption between the general and the grade specific dimension may seem not to approximate the reality accurately. A more realistic framework should be a more general multidimensional IRT model, where different dimensions can be correlated and the number of dimensions in each grade and the proportion of each can be different depending on the curriculum change. The main challenge of using a general multidimensional IRT model is the lack of vertical scaling methods when construct shift occurs across different grades. Until such methods become available, the comparison between the GP model and other persistence model requiring vertical scaling cannot be

made and caution should be used with the generalization of the findings of the current study.

Last but not the least, this study uses a two-stage process to estimate the teacher effect: the first stage is the estimation of the IRT ability parameters, and the second stage is the use of the ability parameter estimates in the VAM models to estimate teacher effects without taking into consideration measurement errors inherent in those ability estimates. The next step could combine this two-stage process into one concurrent estimation process to simultaneously estimate the IRT ability parameters and teacher effect parameters.

REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement (2nd ed.)* Washington, DC: American Council on Education.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics* 29: 37–66.

Beguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000). *Effect of Multidimensionality on Separate and Concurrent Estimation in IRT Equating.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, IL.

Beguin, A. A., & Hanson, B. A. (2001). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating.* Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison: University of Wisconsin Press.

Bergman, L. R., Eklund, G., & Magnusson, D. (1991). Studying individual development: Problems and methods. In D. Magnusson, L.R. Bergman, G. Rudinger, & B. TÄorestad (Eds.), *Matching problems and methods in longitudinal research* (pp. 1-28). Cambridge: Cambridge University Press.

Braun, H. & Wainer, H. (2007). Value-added modeling In C. R. Rao and S. Sinharay (Eds.), *Handbook of Statistics, vol. 27: Psychometrics* (pp. 867-893). Amsterdam: Elsevier Science. (2007).

Briggs, D. C., Weeks, J. P., & Wiley, E. W. (2008). *The Sensitivity of Value-Added Modeling to Vertical Scaling*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Briggs, D. C. & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical scale. *Education Finance & Policy*, 4(4), 384-414.

Briggs, D., & Domingue, B. (2011). *Due Diligence and the Evaluation of Teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times*. Boulder, CO: National Education Policy Center. Retrieved February 2, 2012, from http://nepc.colorado.edu/publication/due-diligence

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling: An International Journal*, 1, 103–124.

Camilli, G., Yamamoto, K., & Wang, M. –M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, *17*, 379-388.

Carlin, B., & Louis, T. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC Press.

Chin, T. Y., Kim, W., & Nering, M. L. (2006, April). *Five statistical factors that influence IRT vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

DePascale, C. A. (2006). *Measuring Growth with the MCAS Tests: A Consideration of Vertical Scales and Standards*. National Center for the Improvement of Educational Assessments.

Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1996). *Analysis of longitudinal data.* New York: Oxford University Press.

Fitz-Gibbon, C. T. (1997). *The value added national project: Final report, feasibility studies for a national system of value added indicators*. London: SCAA.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika, 57*(3), 423-436.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Goldhaber, D. (2010). *When the Stakes are High, Can We Rely on Value-Added? Exploring the Use of ValueAdded Models to Inform Teacher Workforce Decisions*. Center for American Progress. Accessed on January 9, 2012, from http://www.americanprogress.org/issues/2010/12/pdf/vam.pdf

Gordon, R., Kane, T., & Staiger, D. (2006). *Identifying effective teachers using performanceon the job* (Technical report; White Paper 2006-01). Washington, DC: The Brookings Institution.

Harris, D., & Sass, T. (2006). *Value-added models and the measurement of teacher quality*. Unpublished manuscript.

Harris, D. J., & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*, 151-159.

Harris, D.N. (2008). *The policy uses and "policy validity" of value-added and other teacher quality measures*. Paper presented at the National Conference on Value-Added Modeling, University of Wisconsin–Madison, April.

Hendrickson, A. B., Cao, Y., Chae, S. E., & Li, D. (2006 April). *Effect of base year on IRT vertical scaling from the common-item design*. Pater presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Hibpshman, T.L. (2004). *A Review of Value-Added Models*. Kentucky Education Professional Standards Board.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement,* 19, *139-147.*

Kadane, J.B., & Schum, D.A. (1996). *A probabilistic analysis of the Sacco and Vanzetti evidence*. New York: Wiley.

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003, April). *Separate versus concurrent calibration methods in vertical scaling*. Paper presented at the

annual meeting of the National Council on Measurement in Education, Chicago, IL.

Karkee, T., Wang, Z., & Green, D. R. (2006, April). *Exploring the effects of dimensionality on three vertical scaling procedures.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Karl, A., Yang, Y. and Lohr, S. (2012) Efficient Maximum Likelihood Estimation for Multiple Membership Mixed Models Used in Value-Added Modeling. *Computational Statistics and Data Analysis,* 59 (2013), 13–27.

Karl, A.T., Yang, Y., Lohr, S., 2012. *GPvam: maximum likelihood estimation of multiple membership mixed models used in value-added modeling*. R Package Version 2.0-0. http://cran.r-project.org/web/packages/GPvam/index.html.

Kim, J., Lee, W., & Kim, D. (2008 March). *The effect of choosing a base grade on the vertical scale using various IRT calibration methods.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, NY.

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

Koretz, D. (2008*). Measuring Up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press. 2009 Outstanding Book Award, American Association of Colleges for Teacher Education.

Li, Y., & Lissitz, R (2012). Exploring the Full-Information Bifactor Model in Vertical Scaling With Construct Shift. *Applied Psychological Measurement* January 2012 vol. 36 no. 1 3-20.

Linn, R. L. (2001). *The Design and Evaluation of Educational Assessment and Accountability Systems* (No. CSE Techinical Report 539). Los Angeles, CA: Center for the Study of Evaluation (CSE), National Center for Research on Evualtion, Standards, and Student Testing (CRESST).

Lisstz, R.W., & Huynh, H. (2003). Vertical equating for state assessments: issues and solutions indetermination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, 8.

Lissitz, R. (Ed.). (2005). *Value added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

Lockwood, J., McCaffrey, D., Mariano, L., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, 32, 125–150.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983) A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing.* New York: Academic Press. (Reprinted from the Computerized Adaptive Testing Conference, 1979, April, Minneapolis)

Mariano, L., & McCaffrey, D., Lockwood, J. (2010). A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling. *Journal of Educational and Behavioral Statistics*, 35, 253–279.

Martineau, J. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for value-added accountability. *Journal of Educational and Behavioral Statistics*, 31, 35–62.

McCaffrey, D., Lockwood, J., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability (MG-158-EDU)*. Santa Monica, CA: RAND.

McCaffrey, D., Lockwood, J., Koretz, D., Louis, T., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29, 67–101.

Meng, H., Kolen, M. J. & Lohman, D. (2006, April). An empirical investigation of IRT scaling methods: How different IRT models, parameter estimation procedures, proficiency estimation methods, and estimation programs affect the results of

vertical scaling for the Cognitive Abilities Test. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23, 13–23.

Meyer, R. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16,* 183–301.

Mislevy, R. J. , & Levy, R.(2007). Bayesian psychometric modeling from an evidence-centered design perspective. In C. R. Rao, & S. Sinharayc (Eds.), *Handbook of Statistics 26: Psychometrics* (pp. 607-641). Amsterdam: North-Holland.

Murphy, D. (2012) *Where is the Value in Value-Added Modeling* [White paper]? Retrieved from http://educatoreffectiveness.pearsonassessments.com/downloads/ViVa_v1.pdf

Oshima TC, Davey TC, Lee K (2000) Multidimensional linking: Four practical approaches. *Journal of Educational Measurement 37*:357–373

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales.* (pp. 253-272). New York, NY US: Springer Science + Business Media.

Pomplun, M., Omar, H., & Custer, M. (2004). Comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model. *Educational and Psychological Measurement, 64,* 600-616.

R Development Core Team, 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasbash, J., & Browne,W. (2001). Modelling non-hierarchical structures. In A. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics* (pp. 93–103). West Sussex, England: Wiley.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* William H. Angoff Memorial Lecture Series. Available www.ets.org/Media/Research/pdf/PICANG9.pdf. Accessed 10 March 2009.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401-412.

Reckase, M. (2004). The Real World is More Complicated than We Would Like *Journal of Educational and Behavioral Statistics*, 29(1): 117-120.

Rothstein, J. (2009). Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, 4(4), 537–571.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record*, 104, 1525–1567.

Rubin, D., Elizabeth, S., & Elaine Z. (2004). A potential outcome view of value-added assessment in education. *Journal of Educational and Behavioral Statistics* 29 (1): 103–16.

Sanders, W., Saxton, A., & Horn, B. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluational measure?* (pp. 137–162). Thousand Oaks, CA: Corwin.

Schum, D.A. (1987). *Evidence and Inference for the Intelligence Analyst*. University Press of America, Lanham, MD.

Schum, D.A. (1994). *The Evidential Foundations of Probabilistic Reasoning*. Wiley, New York.

Schmidt, W.H., Houang, R.T., McKnight, C.C. (2005). Value-added research: Right idea but wrong solution? In: Lissitz, R. (Ed.), *Value Added Models in Education: Theory and Applications*. JAM Press, Maple Grove, MN, pp. 145–165.

Shkolnik, J., Hikawa, H., Suttorp, M., Lockwood, J., Stecher, B., & Bohrnstedt, G. (2002). Appendix D: The relationship between teacher characteristics and student achievement in reduced-size classes: A study of 6 California districts. In G. W. Bohrnstedt, & B. M. Stecher (Eds.), *What we have learned about class size reduction in California Technical Appendix.* Palo Alto, CA: American Institutes for Research.

Simon, M. K. (2008). *Comparison of concurrent and separate Multidimensional IRT linking of item parameters.* ProQuest Information & Learning, US.

Skaggs, G., & Lissitz, R. W. (1986a) An exploration of the robustness of four test equating models. *Applied Psychological Measurement, 10,* 303-317.

Slinde J.A. & Linn R. L (1979) A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement,* 16*, 159-165.*

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210. doi: 10.1177/014662168300700208.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29(1), 11–36.

Tong, Y. (2005). *Comparisons of methodologies and results in vertical scaling for educational achievement tests.* Unpublished doctoral dissertation, University of Iowa, Iowa City.

Tong, Y., & Kolen, M. J. (2006 April). *Scale shrinkage*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wang, M. (1986). *Fitting a unidimensional model to multidimensional item response data*. Paper presented at the Office of Naval Research contractors meeting.

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. Journal of Educational Measurement, 35, *93-107.*

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal Analyses of Academic Achievement. *Journal of Educational Measurement, 28*(1), 61-76.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50,* 399-410. 25

Yen, W.M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299-325.

Yen, W.M.(2007). Vertical Scaling and No Child Left Behind. In N.J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp.273-283). New York, NY: Springer.