COMPARISONS BETWEEN EDUCATOR PERFORMANCE FUNCTION-BASED AND EDUCATION PRODUCTION FUNCTION-BASED TEACHER EFFECT ESTIMATIONS

By

Eun Hye Ham

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods – Doctoral of Philosophy

2014

ABSTRACT

COMPARISONS BETWEEN EDUCATOR PERFORMANCE FUNCTION-BASED AND EDUCATION PRODUCTION FUNCTION-BASED TEACHER EFFECT ESTIMATIONS

By

Eun Hye Ham

Challenging the current discordance in orientation between student assessment models and teacher/school value-added models, this study aims to present the educator performance function (EPERF)-based teacher effect estimation method which utilizes the nature of student criterion-referenced assessment, to evaluate its feasibility and usefulness by comparison with the currently prevailing methods – the education production function (EPROF)-based value-added model. Specifically, this study (1) investigated how different the teacher effect estimates of the EPERF-based method are from those of the EPROF-based method, (2) examined whether the model fit of the EPERF is acceptable, and (3) simulated whether the EPERF-based method is robust to the locations of cut-scores and number of performance levels. A northern state's student-teacher linked data set was used, and the student challenge index, which is defined as the degree of difficulty that teachers face in teaching a student to attain a desired/higher performance standard, was constructed as a summary quantity of individual students' characteristics.

The main findings from comparison between the two different teacher effect estimates – the educator performance level (EPL) from the EPERF-based method, and the value-added measure (VAM) from the EPROF-based method – were as follows: First, rank correlations between the two estimates were above .82 for mathematics. In particular, the EPL from the polytomous EPERF were very close to the VAM estimates in terms of ranking teachers, showing above .8 rank correlations. Second, in consistent and considerable ways, the relationship of the

teacher effect estimates to student and teacher characteristics did not differ between the EPL and VAM estimates. Third, intra-teacher rank correlations across different subjects and different grade levels were also similar between the EPL and VAM. These observations implied that the teacher ranking information resulting from the EPERF-based methods did not differ noticeably from the results of the EPROF-based method. The EPERF-based methods, however, produced several useful areas of information for understanding how average or individual teachers perform with their students.

For the second question, the EPERF showed a reasonable model fit in mathematics but not in reading. The conditional independence assumption of student success was violated. The amount of conditional dependency within each teacher was reasonable, and tended to be larger than in the EPROF-based models. Regarding the third question, it was found that, as a result of real-data simulations, the EPL based on the polytomous performance levels was quite robust to the location of cut-scores, and the number of performance levels also did not substantially change the teachers' ranking. These mixed results of model-fit and the robustness of the estimates bring into question on whether the EPL estimates change when student challenge index indicators are added, or when more generalized EPERF models are applied.

This study appraised a part of the validity evidence of using the EPERF-based method, including if the method is executable and if the estimated teacher effects are trustworthy, along with the comparison with the EPROF-based method. Implications of applying the EPERF-based teacher effect estimation and future directions for expanding the method are discussed.

My dedication goes my parents to whom I owe all the love and care.
My prayers are always with you, Mom and Dad!

ACKNOWEDGEMENTS

My deep gratitude goes first to my advisor, Dr. Mark Reckase who has given me cheerful encouragement and thoughtful prodding as needed. I much appreciate his willingness to support my studies, not only by offering creatively structured experience, but also by connecting me with various resources. My sincere appreciation also to my dissertation committee, Drs. Amita Chudgar, Spyros Konstantopolous, and Joseph Martineau for their supportive guidance as well as invaluable comments, and to Mr. Alexander Schwarz at the Michigan Department of Education for his help in my accessing data sets used for this dissertation study.

To the research team of value-added measures, including Drs. Cassandra Guarino and Jeffrey Wooldridge, and colleagues from the Economics of Education program, in particular, Francis Smart and Brian Stacy. I express my sincere gratitude for their considerable patience, encouragement, and help in that half of my doctoral years we spent together. And thank you to Dr. Douglas Campbell for his help with my academic writing, to Dr. Tenko Raykov for his intellectual challenges, and to my language partner Rose Jangmi Cooper for her patient and loving help. I also was greatly encouraged by receiving a Robert Craig Fellowship, an Anderson-Schwille Endowed Fellowship, and a Robert L. Ebel Endowed Scholarship from the College of Education.

My special appreciation goes to Dr. Sun-Gun Baek at Seoul National University who first sparked my interest in educational measurement, and to Dr. Richard Robert at Educational Testing Service who strongly encouraged me in my final doctoral year. I am also grateful to Drs.

Dongil Kim, Junyeop Kim, Youngbin Kim, Cheolil Lim, Soyoung Park, and Ki-Sun Sung for providing warm encouragement and opportunities to join in their works.

For their invaluable cheers and advice, thank you to my dear friends, Chi Chang, Inchul Choi, Yunjeong Choi, Emre Gonulates, Seung-Hwan Ham, Seunghyung Hwang, Hyesuk Jang, Unhee Ju, Hosun Kang, Chong Min Kim, Jay Lee, Liyang Mao, Tae Seob Shin, Veronica Son, and Anne Traynor at/from Michigan State University, who cheered me on warmheartedly and offered me thoughtful advice. And my profound gratitude to Pastor John Won, and to Jihye Jo and Hyunjung Byun for their continuing prayers.

No words can fully express my deep and heartfelt gratitude to my respected parents, Yeongsoo Ham and Jaesun Kim, my loving husband Wonjong Kim, and my precious sister Inji for their unfailing trust and support. My final acknowledgement goes to God Who blessed my endeavors with so many gifts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

CI ……...…… Challenge Index

EPERF ……...……... Educator Performance Function

EPERF-D1PL …… Random intercept EFPERF-based models using a dichotomous outcome

EPERF-P2PL … Random intercept and slope EFPERF-based models using a dichotomous outcome

EPERF-P1PL …… Random intercept EFPERF-based models using a polytomous outcome

EPERF-P2PL … Random intercept and slope EFPERF-based models using a polytomous outcome

EPL …… Educator Performance Level

EPL-D1PL …… EPL based on EPERF-D1PL

EPL-D2PL …… EPL based on EPERF-D2PL

EPL-P1PL …… EPL based on EPERF-P1PL

EPL-P2PL …… EPL based on EPERF-P2PL

EPROF ……...…… Education Production Function

VAM ……...……. Value-Added Model/Measure

VAM-AR ……....……....…… VAM based on Average Residuals

VAM-GA ……....…....…… VAM based on Student Gain Scores

VAM-RE ……....……....…… VAM based on Teacher Random Effects

CHAPTER 1. INTRODUCTION

1.1 Background

In the context of outcome-based accountability,[1] which has become a major subject that state governments now apply to their decision making, school or teacher effectiveness is generally regarded as how much of a contribution they make to their students' progress in educational outcomes, rather than teacher qualifications or the quality of their teaching performance itself (Carnoy, 2003; Jacob, 2007; Linn, 2000; O'Day, 2002). Studies to estimate individual school or teacher effects on students' growth in academic achievement by so-called 'value-added modeling' are increasing, and the movement to use the results for sanctioning and rewarding schools or teachers has been boosted (Harris et al., 2010; Sanders & Horn, 1994).

The implementation of outcome-based accountability requires not only a solid testing model to measure students' educational outcomes and an accurate value-added model to compute teachers' effects on those outcomes, but also effective coordination between the two models. Interestingly, however, it appears, as Koretz & Hamilton (2006) implied, that the way that value-added models have been developing is fundamentally incompatible with the way that testing models have been moving forward for the last 20 years: In a nutshell, testing models are criterion-oriented, whereas value-added models are norm-oriented.

[1] Despite the argument that outcome-based accountability should be conceptually distinguished from test-based accountability (for example O'Day, 2002; Spady, 1994), this study considers them interchangeable, under the condition that educational outcomes are mostly considered achievement test scores.

It is meaningful to recall why criterion-referenced tests have been substantially emphasized and expanded for building school and teacher accountability systems; a test needs to be designed in alignment with a curriculum, that is, what teachers are expected to teach, and the test results need to determine what a student knows and can do, and to guide what a teacher has to do with his/her students. This idea has necessitated major changes in the design of statewide tests, and in the ways of interpreting and using the results, which can be summarized as a shift from norm-referenced tests to criterion-referenced-tests and measurement-driven instruction (Koretz & Hamilton, 2006; Smith, O'Day, & Cohen, 1991). In contrast, current value-added models attempt to evaluate teachers based on their students' average learning gains over years, which are determined by what test scores other students in the state obtain. In this case, what teachers are expected to be accountable for regarding their students is not clearly stated until test results are released.

Challenging this discordance in orientation between the student-testing model and the teacher value-added model, this study aims to introduce a new teacher effect estimation model which supports the idea of criterion-oriented testing models, and to evaluate its feasibility and usefulness by comparison with an existing value-added model. To be specific, this study compared the new approach, the educator performance function-based teacher effect estimation, with the currently prevailing methods, the education production function-based value-added estimation.

The educator performance function (EPERF) is a non-linear probability model to describe the relationship between a teacher's proficiency, student characteristics, and students' success in reaching a certain performance level (Reckase, 2012); it applies mathematical models and main concepts of item response theory models (IRT models) to represent the relationship

between examinees' proficiency, item characteristics, and their correct responses to test items. Meanwhile, the education production function (EPROF), which is the basis of value-added models (VAM), formulates the relationship between student achievement test scores and relevant inputs such as teacher or school effects and student characteristics for determining best effectiveness as a linear regression model (Boardman & Murnane, 1979; Hanushek, 1979; Todd & Wolpin, 2003).

This EPERF-based teacher effect estimation model is distinguished intuitively from the EPROF-based, by the following points: (1) the way that the model uses achievement test results; (2) the way that student characteristics are incorporated in each model; (3) the way that the model defines teacher effectiveness; and (4) the statistical property of the teacher effect estimates that each model yields.

First, in the EPERF, the outcome variable is whether students pass or fail at a desired standard performance level, or is a categorical variable of performance standard classification. Currently, most state tests are designed and developed for criterion-referenced tests, aiming to determine into which performance category a student can be classified according to the state's performance target. In contrast, in the EPROF-based VAM, the outcome variable is a continuous test score and, basically, the method utilizes norm-referenced rank information of where a student stands compared to other students based on test scores.

Second, in the EPERF, student characteristic variables are used to determine individual students' challenge levels, i.e., the degree of difficulty teachers have in bringing a student to a desired performance standard, given that student's characteristics. The scale of the student challenge level decides the scale of the teacher effect estimates. In the EPROF, student

characteristic variables are considered as additional educational inputs other than teachers or schools.

Third, in the EPERF-based teacher effect estimation method, individual teacher's effect is treated as a latent trait, a proficiency in helping students achieve a desired performance standard. Each teacher's level of proficiency is estimated based on the challenge levels as well as on successes in the target performance of students the teacher taught. Meanwhile, in the EPROF-based method, the teacher effect is not explicitly defined but is estimated by the size of change in student-observed test scores after accounting for the effects of other inputs.

Fourth, related to the third point, teacher effect estimates from the EPERF-based method are sample-independent, once each student's challenge level is determined, whereas in the EPROF-based method, teacher effect estimates change depending on the sample involved in an analysis.

Although this new method seems comprehensible and advantageous in light of the above salient characteristics, since studies on the EPERF and its application are in their embryonic stage, relevant concepts and details of how it works rarely have been explicated. Accordingly, it still is open to question whether the results from the EPERF-based method provide useful information about teacher performance. At this stage, one of the basic questions that draws a great deal of attention is how teacher effects estimated from the EPERF-based method are consistent with those estimated from the EPROF-based VAM method that has been extensively used. This is the key question to be answered in this study.

1.2 Guiding Questions

The purpose of this study is to compare teacher effect estimates based on the educator performance function (EPERF) to those based on the education production function (EPROF), and to evaluate the feasibility of the EPERF-based teacher effect estimation. In this study, how the EPERF-based teacher effect estimation is distinct from the EPROF-based value-added estimation (VAM) with respect to their main concepts or assumptions was examined first. Along with this theoretical clarification of their similarities and differences, the following three specific questions were empirically answered.

The first question was how different the results from the EPERF-based method were from those of the EPROF-based VAM method. Using a northern state's student-teacher linked data set, individual teachers' effects were computed separately when using the EPERF-based method and when using the EPROF-based method. Then the estimates were compared, with the aim to scrutinize how they were consistent or different. It was also examined if some general features of the teacher effects from the EPERF-based method confirmed or opposed the findings from using the EPROF-based VAM method. To be specific, the relationship to student and teacher characteristics and the consistency of the estimates were monitored.

The second question was whether several basic assumptions of the EPERF were acceptable. The model fit of the EPERF, the conditional independence of student success, and the amount of dependency among student success were evaluated.

The third question was whether the results from the EPERF-based method were sensitive to the locations of cut-scores and the number of performance categories. Whether the teacher

effect estimates change depending on the locations of cut-scores or the number of performance categories was investigated by implementing small simulations with real data.

This study was intended to investigate thoroughly both the conceptual and empirical differences between the EPERF and the EPROF-based teacher effect estimation methods. Along with these comparisons, the inquiries on whether the new method is executable and whether the estimated teacher effects are trustworthy allowed us to appraise a part of the validity evidence for this new method. Taken together, the possibilities and limitations of the current EPERF-based teacher effect estimation are discussed with respect to the interpretation and use of the results, which provides some guidance on the conditions under which the EPERF-based methods are more feasible, and which points to future directions for developing and expanding the application of the EPERF.

CHAPTER 2. MODELS AND ASSUMPTIONS

Both the educational performance function (EPERF)-based models and the educational production function (EPROF)-based models are applied to estimate individual teachers' contributions to their students' learning, based on how the students performed on a large-scale assessment. Fundamental differences between the two models are (1) the type of student outcomes demonstrated in the tests to be used to evaluate teachers – performance standards (criterion-referenced) vs. scale scores (norm-referenced); and (2) the scale of teacher effectiveness measures – unit of student challenge index vs. unit of student test score. For the purpose of contrasting the two models in details, this chapter introduces the main concepts and assumptions of the EPERF, illustrates how the EPERF-based model works, and reviews findings of teacher effects from studies of the EPROF-based models.

2.1. Educator Performance Function-based Teacher Effect Estimation

2.1.1. The Educator Performance Function

The educator performance function (EPERF) represents the probability that a teacher or school succeeds in helping students with a certain level of challenge to achieve a given performance standard (Reckase, 2012). The main idea is to apply an analogy with item response models for estimating the latent teaching ability of educators, called the educator performance level (EPL): teachers differ in their level of proficiency in helping students progress in their academic achievement; students differ in the level of challenge that they pose for teachers, depending on their backgrounds as well as their prior achievement; a teacher's performance can

be determined by the level of challenge of the students with whom he/she has worked, as well as by his/her successes as demonstrated in students' performance. That is, each student is regarded as an item/task that each teacher takes on, and each teacher's EPL is estimated based on his/her successes in helping students reach given performance levels.

Equation 2-1 illustrates an EPERF for dichotomous performance categories, i.e., mastery coded as 1 and non-mastery coded as 0, as initially proposed by Reckase (2012). Basically, this is analogous to the two-parameter logistic IRT model (Lord, 1980).

$$P\left(s_{ij}=1\middle|\theta_j,a_j,X_i\right)=\frac{\exp\left[a_j\left(\theta_j-X_i\right)\right]}{1+\exp\left[a_j\left(\theta_j-X_i\right)\right]} \qquad \text{Equation 2-1}$$

Specifically, the probability that teacher *j* with θ level of teaching ability succeeds in helping student *i* to achieve a performance standard, $P(s_{ij}=1)$, is a function of the EPL of teacher *j*, θ_j, and the challenge level for student *i*, X_i, which is analogous to item/task difficulty. The slope parameter, a_j, indicates the strength of the relationship between the challenge level and the proficiency level of students within a teacher's classroom. The probability that a student who was taught by a teacher for a year is classified by a mastery level is modeled by the teacher's EPL as well as by the student's challenge level that could hinder the student from achieving an appropriate academic performance.

Unlike an IRT model, which estimates both item and person parameters simultaneously, the challenge level for each student is predetermined before fitting a logistic function. Ways to decide the challenge levels for students are discussed in detail in the following section. Once each student's challenge level is determined as reasonable and accurate, each teacher's EPL can

be estimated using the maximum likelihood estimation (MLE), or empirical Bayesian estimation, in the same way that regular item response models do. As a result, each teacher's EPL is his/her location on the scale of student challenge level, where the teacher has a .5 probability of bringing his/her students at that level of challenge to the desired performance.

This model can be generalized to cases with more than two categorical performance standards, such as basic, proficient, and advanced. Specific models either for dichotomous and polytomous performance categories used in this study are detailed in Chapter 3.

2.1.2. The Challenge Index for Students

Student challenge level is defined as the degree of difficulty that teachers face in teaching a student to attain a desired performance standard. It is assumed that students differ in how much challenge they pose to teachers expected to help their students achieve success in certain academic standards. It also is assumed that individual students' challenge levels can be estimated based on their observable characteristics; the quantity for this level is called the challenge index (CI). First consideration when constructing the student CI is what students are demanding from teachers to help them successfully fulfill a required performance level in each subject-matter, or which student characteristics can possibly impede their success. The rich previous literature on academic achievement up to now offers concrete ideas for the selection of reasonable indicators. Drawing on this literature, the main facets and relevant variables which can be potentially used to construct the student CI are listed as follows.

First, family background, such as parental socio-economic status, has been known widely as a predominant predictor of student achievement (Baker, Goesling, & Letendre, 2002; Chudgar & Luschei, 2009; Coleman, 1969; Hanushek, 1992; Sirin, 2005). For example, parental socio-

economic advantages relate closely to the amount of practical support for school work, as well as cultural resources for learning, that their children are given, which accordingly results in a positive influence on academic performance. By contrast, lack of support and resources for children's learning due to socially or economically disadvantaged families can be considered indicators of CI. Free or reduced lunch status is a well-known indicator for the latter. Parents' incomes or eligibility for housing support programs can be relevant also.

Second, some student characteristics directly hinder students from effective learning in schools. Students are dissimilar in what they bring into a new academic year and a new classroom. They differ in their cognitive abilities, and some are diagnosed as having learning disabilities. Since learning is accumulative process, the previous year's deficiency in curricular coverage also affects performance in following years. Also, previous achievement is likely to capture a substantial amount of variation in those factors (Ballou, Sanders, & Wright, 2004; Chetty, Friedman, & Rockoff, 2011; Konstantopoulos, 2014; Papay, 2011). Learning support program eligibility, including special education programs, can additionally be counted in constructing student CI. Limited English proficiency related to home language or migrant status also has been shown to be a significant predictor of achievement (e.g., Baker, Goesling, & Letendre, 2002; Buddin & Zamarro, 2009).

Third, relevant factors which can facilitate students' learning can inversely be counted in estimating their challenge levels. Students who actively participate in class or are highly motivated are more likely to learn better in their classroom than those who are not motivated or are often absent from classes (Brophy, 2010; Hulleman, Durik, Schweigert, & Harackiewicz, 2008). The latter is more challenging for teachers. Similarly, students who are more interested in or more value a particular subject-matter, and therefore spend more time on self-learning, are

likely to cooperate with their teachers to achieve a desired/higher performance standard. These characteristics will be inversely proportional to the challenge level.

Fourth, many studies have reported that some demographic variables such as gender or ethnicity also predict nontrivial amounts of variance in achievement (Aaronson, Barrow, & Sander, 2007; Ballou, Sanders & Wright, 2004; Buddin & Zamarro, 2009; Jacob, 2007). For instance, in particular in mathematics and science, male students tend to perform better on average than female students. Even though this general finding shows a phenomenon, it neither means that female and male students are expected to perform differently nor that teachers feel it is more difficult to work with female or minority students, for example, in mathematics, in order to achieve a desired performance level, thus whether to include those variables in a student CI is controversial.

Last, possibly disadvantageous school-level as well as classroom-level characteristics also can be taken into account. Some studies have shown that individual student performance depends on the dynamics of classroom interaction, and sometimes the school's environment has an effect on level of achievement (Burke & Sass, 2008; Card & Krueger, 1996).

Certainly, other unknown factors affect the degree of difficulty that teachers experience in their classrooms, and more indicators which are considered appropriate can be measured and added. It is also beneficial for teachers and/or school principals to participate in discussing and deciding which indicators ought be included in creating a student CI.

Another concern is how the indicators are weighted and combined to compose the challenge index which is fair to all teachers for evaluation purpose. Reckase (2012) proposed two methods to construct CI. One is to estimate the locations of the students based on an IRT model regarding student indicator variables as test items (IRT calibration). In this case, students

who achieve higher scores on the hypothetical latent trait represent those who impose more demands on their teachers. Another is to use the inverse (minus) of the predicted achievement, a linear combination of indicators, as the CI (OLS weighted sum score). A set of weights for the prediction is obtained from regressing achievement on selected student indicator variables using the previous cohort's data. In any case, it is critical to assure that the same indicator variables and the same weights are used to construct the CI for students who are involved in the analysis, so that all students are on the same CI scale, thus comparable. Both methods were tested and evaluated in this study, and specific procedures are described in Chapter 3.

2.1.3. Assumptions

Before presenting the assumptions required for the EPERF to be feasible, several prerequisite assumptions need to be recognized. First, cut-scores used to classify students into different performance categories are assumed to be precisely determined through an appropriate procedure. Accordingly, it has to be assured that the classification of each student based on the cut-score is reliable as well as valid. Second, it is also assumed that all key student characteristics or background variables that possibly influence their achievements are included in forming the challenge index, so that the CI accurately quantifies the level of challenge for students by considering out-of-school factors that possibly hinder students working with a teacher from reaching the desired performance standard. Finally, it needs to be assumed that the higher the challenge level for students, the more difficult it is for teachers to help them pass the desired standard. Otherwise, at least, there has to be an agreement that the success of a student whose challenge level is higher needs to be more heavily weighted for estimating EPL.

Basic assumptions that need to infer the causal relation between the estimated teacher capability and students' successes from the EPERF are identified, corresponding with those of regular item response models (de Ayala, 2009), as follows: (1) the fit of logistic regression models, (2) the uni-dimensionality of teacher capability, and (3) the conditional independence of students.

The most fundamental assumption is that the EPERF's functional form represents the relationship between teachers' capabilities and students' successes reasonably well. Specifically, the probability that a teacher succeeds in helping a student to reach a performance standard is assumed to increase monotonically with the level of teaching proficiency, $P(s_i = 1) \propto EPL$, along with a S-shape curve. That is, the more proficient a teacher, the higher the probability of success. The relationship is assumed to be non-linear, as is the logit function linking the teacher EPL and the probabilities of students' successes, as described in Equation 2-1. In order to check this assumption, the fit of the logistic regression models to the data was tested in several ways in Chapter 5.

The unidimensionality assumption of teacher capability is that there is only a single latent trait of the teacher that can explain the statistical dependence of students' successes, once their challenge levels are taken into account. This also can be stated in the way that unobservable student characteristics, which predict students' successes but are omitted from forming CI, are independent of teacher assignment once the CI takes account of the observed characteristics. That is, teacher assignment is an ignorable condition on the CI, which is similar to what Reardon & Raudenbush (2009) called the assumption of ignorability. This assumption brings back a concern over the quality of CI, because the teacher effect estimation procedure inherently relies on the student's CI.

The conditional independence of students' successes means that each student's probability of reaching a certain performance level is independent of each other after taking account of the single teacher trait. This assumption is close to that of no interference between units (Reardon & Raudenbush, 2009), or that of the stable unit treatment value assumption (SUTVA; Rubin 1986). If within a classroom there is interaction among students within a classroom that influences their successes, i.e., the so-called 'peer effect', then the conditional independency is challenged. This assumption also is closely associated with the unidimensionality assumption; if there is another dimension of teacher capability affecting students' successes, or if there is an unobservable variable interacting with teacher assignment, which is a violation of the unidimensionality assumption, the conditional independence assumption is violated.

When the conditional independence assumption is untenable, the joint probability of any performance pattern of N students who were taught by a teacher with a given level of θ cannot be equal to the product of the probabilities of individual students' successes when the teacher's EPL is set to θ; consequently, the teacher EPL estimates from the EPERF can be less dependable.

Additionally, it needs to be assumed that the distribution of CI for students per teacher has no significant impact on the estimation of teacher EPL. Because each teacher is sometimes exclusively assigned to different groups of students, in other words, a student is assigned to one or at most two teachers for each subject, the distribution of student CI is less likely to be identical across all teachers. This is a clear distinction from an item response function – a test item is given to most examinees. Therefore, admitting that the distribution of student CI is different across teachers, we need to confirm that the effect of the difference in the student CI distribution on the teacher effect estimation is minor or not in favor of a certain group of teachers.

2.2. Education Production Function-based Teacher Effect Estimation

2.2.1. Background and Models

The education production function (EPROF) has been popular and has been elaborated by economists and sociologists since the 1970s, in order to inquire into how much effect each input has on educational outcomes on average (Hanushek, 1989; Goldhaber & Brewer, 1997; Hill, et al., 2005), and how closely out-of-school inputs relate to inequality (Greenwald, et al., 1996; Murnane, et al., 1981). Recently, while emphasizing outcome-based accountability, the value-added approach based on the EPROF has drawn education policy makers' attention as a promising tool to evaluate individual schools' or teachers' effectiveness, rather than the average effects of school or teacher characteristics, such as the effect of teacher qualifications or school resources. Again, the key idea of VAMs is to isolate statistically from all other sources the contribution of individual schools or teachers to student achievement (Harris & McCaffrey, 2010; Meyer, 1997; Sander & Horn, 1994; Rockoff, 2004).

The EPROF relates observed student outcomes to student characteristics and educational inputs, such as teacher or school characteristics (Boardman & Murnane, 1979; Hanushek, 1979; Todd & Wolpin, 2003); this approach originated in the production function approach in industry. Economists and sociologists have employed the EPROFs to investigate how much school inputs or out-of-school inputs, such as family backgrounds (Murnane, et al., 1981), account for educational outcomes, and to determine whether school inputs are effectively invested and managed in ways to maximize student outcomes (Hanushek, 1989). Even though various types of educational outcomes, such as college entrance (Meyer, 1970), earnings or labor market

performance (Card & Krueger, 1998), and even socialization (Dee, 2003), have been used under the umbrella of studies using the EPROFs, most have focused on academic achievement through large-scaled achievement test scores. This is because the content that an achievement test covers is directly connected to what students formally learn most of the time in schools, and student data on academic achievement are in general collected every year by states; therefore, it is easier to observe their changes after years of schooling or a treatment. Also, test scores are regarded as less prone to subjectivity, compared to both teachers' observations and students' self-reported surveys; as less expensive than other measures of long-term outcomes; and as easily handled by quantitative analyses.

A general form of the EPROF is presented in Equation 2-2 (Hanushek, 1992, Harris & McCaffrey, 2010; Meyer, 1997). The achievement test score of student *i* at year *t* can be represented as a function of the student's family background, $F_i^{(t)}$; different inputs of the school the student attends, including the teacher, $S_i^{(t)}$; and student time-invariant characteristics, called innate ability, $\mathrm{I_i}$, and student time-varying student characteristics, μ_{it}. Note that cumulative inputs of families and schools reflect the cumulative nature of education.

$$\mathrm{A_{it}} = \mathrm{f}\left(\mathrm{F_i^{(t)}}, \mathrm{S_i^{(t)}}, \mathrm{I_i}, \mu_{\mathrm{it}}\right) \qquad \text{Equation 2-2}$$

When the change in achievement between two time points is considered, the EPROF is referred to as a value-added model specification, as in Equation 2-3. This function specifies the optimal relationship between educational resources and student growth in outcomes, and it has been applied to determine good or effective schools or teacher characteristics for student progress (Boyd, et al., 2009; Goldhaber & Brewer, 1997; Nye, et al., 2004).

$$A_{it} = f^{*}\left(F_i^{(t-t^{*})}, S_i^{(t-t^{*})}, I_i, \mu_{it}, A_{it^{*}}\right) \qquad \text{Equation 2-3}$$

Applying this general production function to measure the effect of each unit of educational resources on student growth in achievement brings about many challenges. Economists and educational statisticians (McCaffrey & Harris, 2009) have identified and have dealt with these challenges in different ways.

First, economists have given attention to how to isolate individual teacher or school effects, which allows us to make more rigorous inferences about their causal effects. Estimating individual unit effects is technically trickier than estimating average effects of a certain characteristic of units. More, because the result from the former is more likely to be used for high-stakes decisions than the latter, the former is practically riskier than the latter. Therefore, how to deal with unobserved effects is one of the main issues. Some studies have examined assumptions necessary for inferring a solid causal relationship from a model, have empirically tested if the assumptions are tenable in reality, and have evaluated how the estimators – when the assumptions are not held – can be potentially biased (Clotfelter, et al., 2007; Koedel & Betts, 2009; Rothestein, 2010).

Among economists, various forms of VAM have been refined according to the choice of input variables and the assumptions of the relationship between variables. One of the typical forms uses prior test scores as a covariate and includes a student-fixed effect term (Guarino, et al., in press; Harris &Sass, 2006; Koedel & Betts, 2009; McCaffrey, et al., 2004; Papay, 2011; Rockoff, 2004), as Equation 2-4 shows. Koedel & Betts (2009) labeled this model a "within-students approach"; Guarino et al. (in press) did "dynamic OLS (DOLS)"; and McCaffrey, et al.

(2004) labeled it "covariate adjustment models." The model with $\lambda = 1$ is equivalent to a gain score model, that is, a pooled OLS estimator (Guarino, et al., in press).

$$A_{it} = \lambda A_{i,t-1} + c_i + \beta X_{it} + \theta T_{it} + \varepsilon_{it}$$

A_{it} achievement test score of student *i* at year *t*
c_i time-invariant student fixed effect of student *i*
X_{it} time-varyiant student characteristics of student *i* at year *t*
T_{it} teacher indicator at year *t*
ε_{it} idiosyncratic error at year *t*

Equation 2-4

On the other hand, educational statisticians have developed the models to reflect the nature of student learning and schooling represented in longitudinal data, because the school or education setting differs from the industry setting in respect of the form or level of complexity. Specifically, they have been concerned with 1) intra-class dependency due to the hierarchical structures of data (McCaffrey, et al., 2004; Nye, et al., 2004); 2) intra-person dependency among repeated measures of each individual (McCaffrey, et al., 2004; Nye, et al., 2004); and 3) accumulated effects of educational inputs on student performance (Konstantopoulous & Chung, 2011; Lockwood, et al., 2007; Sanders & Horn, 1994). They have focused more on how accurately a model describes students' growth by taking into account the above three characteristics of data.

The layered model is one of the most popular; it is a constrained version of the general value-added model first described by McCaffrey, et al. (2004). Later it was named the *variable persistence model* by Lockwood, et al. (2007) (Briggs & Week, 2008), as Equation 2-5 shows. When all persistence parameters ($\alpha_{21}, \alpha_{31}, and\ \alpha_{31}$) are equal to 1, it is called a complete

persistence model, which is equivalent to the Tennessee Value Added Assessment System (Sanders & Horn, 1994).

$$
\begin{aligned}
&A_{it} = U_{it} + Z_{it} \\
&U_{it} = c_t' + \lambda A_{i,t-1} + \beta X_{it} + \delta T_{j(i,t)} + \varepsilon_{it} \\
&Z_{i1} = \theta_{j(i,1)} \\
&Z_{i2} = \alpha_{21} \cdot \theta_{j(i,1)} + \theta_{j(i,2)} \\
&Z_{i3} = \alpha_{31} \cdot \theta_{j(i,1)} + \alpha_{32} \cdot \theta_{j(i,2)} + \theta_{j(i,3)}
\end{aligned}
$$

Equation 2-5

Controversies have arisen continuously in using the EPROF-based VAMs to determine individual school or teacher effects and in using the results for high stakes decisions because of their strong required assumptions about student growth and test-scores (Baker, et al., 2010; Ballou, 2009; Harris, 2011; Kupermintz, 2003; Rothstein, 2010; Robin, et al., 2004) as well as the instability of value-added estimates (Briggs & Domingue, 2011; Martineau, 2006; McCaffrey, et al., 2009; Newton, et al., 2010); accordingly, a few alternative methods have been proposed (Betebenner, 2011; Reckase, 2012).

2.2.2. Assumptions

Several assumptions are required to compute causal teacher effects using the above models, as some studies have detailed (Harris, 2009; Reardon & Raudenbush, 2009; Rothstein, 2010) For example, Harris (2009) summarized the assumptions as follows: 1) the school system and teachers' teamwork do not significantly influence student achievement; 2) the impact of prior educational inputs (history) is captured sufficiently by prior achievement test scores; 3) student fixed-effect sufficiently accounts for the nonrandom assignment of students to teachers; 4)

all test scores are equivalent and the scales are interval; and 5) teachers are equally effective with all types of students.

In this study, assumptions about the interpretation and use of achievement test scores, which have been often overlooked in the studies of the EPROF-based VAM, are underlined and elaborated. The EPROF-based VAM is commonly based on the idea that a test score gives us the approximate achievement level of a student, thus teacher effectiveness can be approximated based on a set of test scores. These underlying assumptions supporting the ideas are reviewed with respect to the following three points: (1) test content, (2) test scale, and (3) the use of test results.

First, considering what students have learned, the equivalence of test contents needs to be assumed. That is, contents or constructs being assessed have to be the same between different tests over years, so that we can determine how much value is added on the content/constructs. However, content change or construct shift over years or grades is inherent in student growth; what students are supposed to know and be able to do changes depending on their ages or grade levels. Some studies have demonstrated that using single test scores from different tests not only misrepresents the amount of student growth (Reckase & Li, 2007), but also distorts the teacher value-added estimates based on the test scores (Martineau, 2006).

If one considers test scores only for norm-referenced interpretation – how one performed compared to other students in the tests – regardless of content or performance standards about what students are expected to learn in a certain grade, we might worry less about test contents. In this case, however, "growth" is defined as a change in a student's relative ranks between the prior and current years rather than as a change in the performance standards from the prior year to the current year.

Second, the assumption of interval scales is necessary (Harris, 2009; Reardon & Raudenbush, 2009) in order to interpret regression coefficients of teacher indicators from the EPROF-based VAMs in a meaningful and fair way. That is, one unit difference in a test's scores is assumed to be the same at every point on the test scale. When one uses only rank information of student test scores, a less strict version of this assumption can be applied: a test is equally accurate in terms of ranking students no matter where a student is located in the continuum of achievement. However, there exists no test which has the capacity to rank students equally precisely across every level of achievement, unless the test is very long. Tests are likely more precisely to place students in the middle of the distribution, but to do this less precisely with students in the extremes of the distribution.

Finally, even if we do not care about test contents and even if tests yield very accurate rank information about students, it is still in question what the number values from a test mean and whether teacher effect can be defined by using students' ranks without any reference to what students are expected to know and be able to do, and what teachers are expected to do. Current VAM does not consider the nature of criterion-referenced achievement tests when using test scores to estimate teacher effects. Note that criterion-referenced tests are designed to provide information of what students should know and be able to do (content standard), and how proficient the students are expected to be (performance standard), which is closely aligned to curriculum and guides teachers in what they are supposed to work for/with their students. Under the VAM ignoring these standards, the only resource remaining to teachers/schools to be responsive to their resulting VAM is to make their students obtain higher test scores than others in the state.

2.2.3. Findings from the Studies of the EPROF-based Teacher Effectiveness

Increasing numbers of studies on the EPROF-based VAM have yielded not only some common findings but also some conflicting conclusions about teacher effects on student achievement. This section briefly reviews findings from previous studies on the VAM, focusing on (1) the size of teacher effects, (2) the relationship with teacher qualifications or experience, (3) the relationship with student characteristics, and (4) the consistency of estimates.

Studies have found that 1% to 20% of variance in student test scores were due to differences in teachers (Chetty, et al., 2011; Rothstein, 2010; Condie, Lefgren, & Sims, 2012; Lockwood, et al., 2007; McCaffrey et al., 2004; Nye, et al., 2004; Sanders & Horn, 1997). Many studies have demonstrated that the size of the explained variance generally tended to be larger in mathematics test scores than in reading test scores (Condie, Lefgren, & Sims, 2012; Nye, et al., 2004; Rockoff, 2004), but some studies showed it depended on different grade levels (Konstantopoulos & Chung, 2011). It is known that children's reading ability is more likely to depend on their family backgrounds.

A few studies examined the relationship between teacher VAM estimates and teacher qualifications. Kane, Rockoff, & Staiger (2008) found little or no difference in average value-added measures among different certification status. Findings from many studies on the relationship between student achievement and teacher qualifications implied that the relationship between teacher effectiveness and traditional teacher qualifications is not transparent. Some studies reported no effect of the certification status or licensure on student achievement (Buddin, Zamarro, 2009; Croninger, Rice, Rathbu & Nishio, 2007; Palardy & Rumberger, 2008) but other reported a positive effect (Clotfelter, Ladd & Vigdor, 2007). Some concluded there was a significant impact of advanced degrees (Croninger, Rice, Rathbun & Nishio, 2007; Goldhaber &

Brewer, 2000), but others found there was no effect of advanced degrees (e.g., Clotfelter, Ladd & Vigdor, 2007).

The other strand of studies investigated the sensitivity of the VAM estimates on different statistical models (Lockwood et al., 2007; Newton, Darling-Hammond, Haertel, & Thomas, 2010), different sets of covariates (Ballou, et al., 2004; Papay, 2011), and different types of test scores (Corcoran et al., 2011, Jacob, 2007; Lockwood et al., 2007, Papay, 2011). It has been commonly demonstrated that while different statistical model specifications or different sets of covariates made no substantial change in the VAM estimates, those estimates were more sensitive to using different types of test scores. A few studies also evaluated the consistency of the VAM over years, and the result showed low to moderate correlation between different years (McCaffrey, et al., 2004).

2.3. General Comparisons between the EFERF-based and the EPROF-based Teacher Effect Estimations

Both the educator performance function (EPERF)-based teacher effect estimations and the educational production function (EPROF)-based value-added models (VAM) focus on estimating teacher capabilities, as demonstrated by student performance on achievement tests. This section elaborates the characteristics of the EPERF-based method by contrasting it with the EPROF-based method with respect to their statistical models, measurement aspects, and social consequences. These three aspects are interwoven rather than separate.

2.3.1. Statistical Modeling

Both the EPERF-based teacher effect estimations and the EPROF-based VAM basically use the same regression model to predict student outcomes from a set of student characteristics (see each regression model in sections 3.3.1 and 3.3.2). Evident distinctions between the EPERF-based and EPROF-based models are simply from (1) the types of student outcome variable, and (2) the ways to take into account student background variables in their regression models. These, however, follow the fundamental differences in (1) the definition of teacher effect of interest, and (2) the scale of teacher effect estimates, which are critical for the interpretation and use of the resulting teacher effect estimates.

First, the two models use different outcome variables: a categorical performance standard for the EPERF and a continuous test score for the EPROF. Accordingly, the teacher effect of interest is different between the two. For instance, in the EPERF, one is concerned with how teachers work to help students pass the cut-off scores of the test, while in the EPROF, one focuses on how teachers help their students achieve higher scores on the test than do other students. For the former, what teachers are expected to achieve with their students is obvious, in alignment with the student performance standards based on the state benchmark.

Second, the way to take into account student background variables, including prior test scores, is also distinctive between the two models. In the EPROF, they are directly included as covariates in a regression model, in which regression coefficients associated with teacher indicators are regarded as unique teacher effects, after ruling out the effect of those covariates. Meanwhile, for the EPERF-based method, student characteristics are used to determine student challenge levels before computing teacher effects. Then, the student challenge level is used in the

process of estimating teacher effect based on the EPERF. Basically, the EPERF-based method is a two-step procedure.

The big advantage of this two-step procedure of the EPERF-based method is to produce sample-independent estimates of teacher effects, once construction of the student challenge index, including selection of relevant indicator variables and their weights, is decided. Provided that the weights of student characteristics for forming the student challenge indices are appropriately and fairly determined, teacher capability estimates are independent of which sample of students, teachers, and schools is involved in the analyses. By contrast, in the EPROF-based method, teacher effect estimates are dependent on which sample of students is included in the analyses, and individual teachers' ranks also move around relying on the sample of students, schools, or districts the data contain. This causes ambiguity in the interpretation of teacher effect estimates as a stable teacher attribute in the EPROF-based method.

Further advantage of the scale of the EPERF-based teacher effect estimates is delineated in the following section.

2.3.2. Measurement Aspects

The EPROF-based method uses a student test score as an outcome variable, which implies defining educational success as obtaining a higher test score than do other students. Most state tests, however, are designed to assess if students meet a state benchmark of student performances; this has been an important and dominant principle for developing achievement tests as a part of forming a school accountability system. One of the fortes of the EPERF-based method is to allow us to maintain the nature of the criterion reference for evaluating teachers by using student performance categories as an outcome variable.

Since the EPERF is based on the idea and mathematical form of item response models, the teacher effect estimate has several useful properties, as a measure of teacher attributes, that person parameters in item response theory have if all required assumptions can be upheld. For example, teacher performance level is scaled on the same scale as student challenge level, which makes the interpretation of the estimated teacher effects straightforward and sensible. That is, the EPERF-based method basically intends to make comparable the two different but interdependent measures related to student outcome, namely teacher attribute and student attribute. In the language of Wright and Stone (2004), the two measures are how hard a student was to teach, and how effective a teacher was in that student's success.

2.3.3. Social Consequences

There have been many controversies in using the EPROF-based VAMs for high stakes decisions (Baker, et al., 2010; Ballou, 2009; Martineau, 2006; Harris, 2011; Kupermintz, 2003; Rothstein, 2010; Robin, et al., 2004). One concern is that achievement test scores can be misinterpreted and misused. Even if assumptions of interval scale and construct comparability are plausible, a one-unit difference in an achievement test without any reference to what students know and can do cannot necessarily be regarded as a qualitative difference in learning outcomes, nor does it define teacher effects.

Also, for schools and teachers who want to improve their practice, value-added measures are less informative than are other possible measures. Because individual teacher or school effects are measured by how many units of achievement test scores they increased for one year, the only practical information of what they can do for their value-added measure is to make their students' test scores higher than the others, regardless of what the tests assess. Moreover, the

models are very complicated and difficult to use as tools to communicate with many other stake-holders in practice.

The EPERF-based teacher capability evaluation places less emphasis on the informativeness of an achievement test score itself. Instead it turns educators' and students' attentions to benchmarks of student performance that a test assesses. The results can guide what teachers and schools seek for in alignment with state performance standards. Consequently, this provides teachers with space to work on other desirable student outcomes that are not formally assessed but remain highly valued in education, such as non-cognitive outcomes, rather than focusing only on getting their students to obtain higher gains than do other students.

Furthermore, teachers could be less reluctant to teach disadvantaged students when using the EPERF-based method. The EPERT-based method takes students' characteristics that impede their academic achievement into account when evaluating teacher capability.

CHAPTER 3. DATA AND METHODOLOGY

This chapter outlines specific procedure for data analyses: Data properties for this study are described briefly in the first section. Each step of constructing student challenge index (CI) and of applying the educational performance function (EPERF) to compute teachers' educator performance levels (EPL) is detailed in the second and third sections respectively. Also, the steps of computing teachers' value-added measures (VAM) based on the educational production function (EPROF) are summarized at the end of the third section. Each plan of analyzing data for answering the three research questions posed in section 1.2. is delineated in the last three sections: (1) comparisons of the teacher effect estimates between the EPERF-based and the EPROF-based methods; (2) examination of the model fits of the EPERF-based method; and (3) the effect of the locations of cut-scores and the number of performance standards categories on the EPERF-based teacher effect estimates.

3.1. Description of the Data

Student-teacher linked data of the 2010-2011 academic year from the Michigan Department of Education, in particular approximately 300,000 students from Grades 4 to 7 and their 12,000 teachers, were used for this study. The data contained several student background characteristics known to predict achievement levels, such as economically disadvantaged group, limited English proficiency, and attendance. Several teacher backgrounds, including credential types and education, were also available.

The state test, Michigan Educational Assessment Program (MEAP), is administered during every October, the beginning of the new academic year, for the purpose of providing teachers with diagnostic information on students newly assigned to teachers. This assessment administration schedule, however, may not be the best practice for using the results to evaluate newly assigned teachers. This is because the test results on October are likely to reflect how well their previous academic year's teachers taught. Accordingly, in this study, 2010-2011 Academic year's teachers were evaluated based on their students' performance on the October 2011 test not on the October 2010 test, assuming that students' test results from October 2011 are likely to be due to 2010-2011 Academic year's instead of 2011-2012 Academic year.

Descriptive statistics of several student characteristics by school-level are displayed in Table 3-1. The state test classifies students into the following four performance categories based on the test scores through their standard setting procedure: (1) Basic, (2) Partially proficient, (3) Proficient, and (4) Advanced. For applying the EPERF with dichotomous outcomes, 'basic' and 'partially proficient' were merged into non-mastery, coded as 0; and 'proficient' and 'advanced' were merged into mastery, coded as 1. For the polytomous case, the four categories were used as specified in the data. For VAM estimation, students' IRT scale scores were used.

3.2. Constructing the Student Challenge Index (CI)

3.2.1. Selection of Indicators and the Weights

The process of selecting challenge index indicators depends on which one of the two methods – OLS weighted sum score or IRT calibration, proposed in section 2.1.2– would be applied. In either case, it starts with all candidates being available in the given data.

Table 3-1. Distributions of student-level selected indictors (2010-2011 Academic year cohort)

	Elementary (Grades 4-5)					Secondary (Grades 7-8)				
	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
Prior math score	103,828	1.30	1.09	-3.85	7.32	191,487	1.16	1.00	-4.58	6.68
Prior math proficiency	103,828	0.86	0.35	0	1	191,487	0.86	0.35	0	1
Prior reading score	103,414	0.84	1.17	-5.16	4.89	191,223	0.99	1.21	-5.08	5.26
Prior reading proficiency	103,414	0.86	0.35	0	1	191,223	0.83	0.37	0	1
Economically disadvantaged	103,983	0.50	0.50	0	1	191,950	0.46	0.50	0	1
Free/reduced lunch eligibility	105,027	0.50	0.60	0	2	194,007	0.47	0.60	0	2
Free/reduced lunch	105,027	0.45	0.50	0	1	194,007	0.41	0.49	0	1
Targeted assistant program	105,027	0.11	0.31	0	1	194,007	0.06	0.23	0	1
Special education	103,983	0.12	0.32	0	1	191,950	0.11	0.31	0	1
Disability	105,027	0.12	0.33	0	1	194,007	0.11	0.32	0	1
Limited English proficiency	103,983	0.03	0.18	0	1	191,950	0.03	0.16	0	1
Proportion of attendance	105,027	0.98	0.08	0	1	194,007	0.98	0.10	0	1
Above 80% of attendance	105,028	0.97	0.16	0	1	194,008	0.96	0.20	0	1
Female	105,027	0.49	0.50	0	1	194,007	0.49	0.50	0	1
Asian	103,075	0.02	0.15	0	1	189,911	0.03	0.16	0	1
Black	103,075	0.15	0.36	0	1	189,911	0.15	0.36	0	1
Hispanic	103,075	0.05	0.22	0	1	189,911	0.06	0.23	0	1

When using the OLS weights, it is first necessary to have an optimal regression model in which individual indicators uniquely contribute to explaining the variance of student performance. They should be correlated with achievement but not be highly correlated with each other; this yields a set of optimal weights. Starting with a model to regress the following year's achievement on all possible indicator variables, some were associated with very small and insignificant *t*-values, or were showed directions opposite from the expected, which is possibly due to collinearity, and they were deleted stepwise.

Data from the 2009-2010 cohort, the previous cohort of the target cohort of 2010-2011, comprising approximately 110,000 students per grade, were used to select a set of appropriate indicators and to obtain their regression coefficients by grade and subjects. The weights were to be used for constructing the 2010-2011 academic year students' challenge index. In the process, the disability indicator was deleted because no additional unique explanatory power was found. A set of remaining indicator variables were regarded as more effective in explaining the variation in student achievement, and the final selection of the indicator variables is shown in the second column of Table 3-2.

For indicators of socio-economic background, free or reduced lunch eligibility and economically disadvantaged groups which were identified by the state government were available. In particular, the economically disadvantaged subgroups were identified based on a combination of their free- or reduced- meal eligibility, and their immigrant status and homeless status.[2] For indicators of learning disadvantage, targeted assistant school program eligibility, special education program, and limited English proficiency, were used. For the relevant factor of learning advantage, proportion of attendance and prior achievement were used, and their weights

[2] Michigan Department of Education (2012). Community Eligibility Option: Frequently Asked Questions.

were expected to be negative. Finally, depending on whether to include gender and ethnicity as indicator variables, two different sets of indicators, full and compact sets, were determined in order to monitor whether the two sets yield different weights. Assuming that the weights would be dissimilar between grade as well as subject-matter, all regression models were separately estimated by grade and subject-matter.

Table 3-2. Final sets of selected indicators for constructing the student challenge index

Facet	OLS weighted sum score	IRT calibration
Social/economic background	• Economically disadvantaged group • Free/reduced lunch eligibility	• Economically disadvantaged group • Free/reduced lunch eligibility
Learning disadvantaged	• Targeted assistant school program • Special education program • Limited English proficiency	• Targeted assistant school program • Special education program • Limited English proficiency • Disability
Learning advantaged (reverse)	• Proportion of attendance • Previous year's achievement in the same subject	• Proportion of attendance • Previous year's achievement in mathematics and reading
Demographic	• Gender • Ethnicity (Asian, Black, Hispanic)	

Fit indexes of regression models using each final set of indicators and the resulting regression coefficients are shown in Table 3-3 for mathematics and Table 3-4 for reading. The models' Adjusted R-square fell above .6 for mathematics and .5 for reading. More variance in mathematics achievement was explained by the final set of indicators than that in reading. The unexplained variance in student achievement, approximately 40-50% of the total variance, does not imply that the statistical models for obtaining the CI weights are of questionable value. Rather, it is fair to assume that part of the unexplained variance in student achievement is due to

their teachers. In other words, significant room exists for teachers to help students, given their background or characteristics, work better. While for mathematics, R-square slightly increases as grade increases, this does not hold for reading. The inclusion of gender and ethnicity made no substantial difference.

For applying IRT calibration, because student challenge level is considered as a latent trait, indicators are supposed to be substantially correlated, assuming that there is one common latent trait to explain the dependency among the indicators. Also, given that all indicators for IRT calibration were dichotomous, it was attempted to include as many as possible indicators to guarantee a reasonable amount of variation in the estimated challenge level across students. For initial examination, tetrachoric correlations among indicators and Cronbach's alpha (internal consistency between indicators) were observed while each indicator was added or deleted. The final set of indicators for IRT calibration was determined as shown in Table 3-2. No gender and ethnicity indicators were considered because it seems unfair to take gender or ethnicity into account when approximating their level of the latent trait.

3.2.2. Weights of Indicators

A set of weights for the indicator variables was determined by regression coefficients of indicator variables predicting achievement; this is shown in Table 3-3. As expected, prior test scores and proportion of attendance were positively associated with the following year's achievement in either mathematics or reading. Also, they were most heavily weighted compared to other indicators. Students who performed better in the previous year scores and attended more

classes are likely to be less demanding of teachers to teach them to perform better. Weights of proportion of attendance, however, tended to decrease in secondary school years, Grades 6-7, compared to elementary school years, Grades 4-5, in both mathematics and reading.

Table 3-3. OLS weights of selected challenge index indicators (Mathematics)

	Full Set				Compact Set			
Indicators	G4	G5	G6	G7	G4	G5	G6	G7
Pre-test	0.79	0.57	0.68	0.64	0.81	0.58	0.70	0.66
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Economically disadvantaged	-0.10	-0.07	-0.08	-0.08	-0.11	-0.09	-0.09	-0.08
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Free/reduced lunch	-0.05	-0.03	-0.04	-0.03	-0.06	-0.04	-0.05	-0.03
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Targeted assistant	-0.13	-0.06	-0.13	-0.05	-0.11	-0.04	-0.12	-0.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Special education	-0.17	-0.12	-0.18	-0.11	-0.14	-0.11	-0.17	-0.09
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01	(0.01)	(0.01)
Limited English proficiency	-0.01	-0.05	-0.06	-0.05	0.05	-0.01	-0.02	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Proportion of attendance	0.38	0.32	0.19	0.15	0.37	0.34	0.21	0.15
	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
Female	-0.09	0.06	0.03	-0.04				
	(0.00)	(0.00)	(0.00)	(0.00)				
Asian	0.42	0.24	0.26	0.23				
	(0.01)	(0.01)	(0.01)	(0.01)				
Black	-0.13	-0.14	-0.11	-0.09				
	(0.01)	(0.01)	(0.01)	(0.01)				
Hispanic	-0.06	-0.04	-0.05	-0.03				
	(0.01)	(0.01)	(0.01)	(0.01)				
Adjusted R^2	0.64	0.66	0.67	0.68	0.63	0.65	0.67	0.68
N	108,989	108,133	109,232	109,842	108,378	107,763	109,025	109,768

- These weights were obtained from multiple linear regressions to predict mathematic achievement using 2009-2010 Academic year cohort data
- Each cell contains the associated regression coefficient and standard error

Table 3-4. OLS weights of selected challenge index indicators (Reading)

Indicators	Full Set				Compact Set			
	G4	G5	G6	G7	G4	G5	G6	G7
Pre-test	0.57	0.64	0.78	0.58	0.58	0.66	0.81	0.60
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Economically disadvantaged	-0.11	-0.08	-0.13	-0.10	-0.13	-0.11	-0.16	-0.11
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Free/reduced lunch	-0.05	-0.05	-0.08	-0.05	-0.06	-0.07	-0.10	-0.05
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Targeted assistance	-0.15	-0.15	-0.25	-0.14	-0.14	-0.13	-0.23	-0.14
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Special education	-0.29	-0.28	-0.35	-0.29	-0.29	-0.27	-0.34	-0.30
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Limited English proficiency	-0.17	-0.23	-0.22	-0.09	-0.10	-0.18	-0.14	-0.07
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
Proportion of attendance	0.67	0.39	0.30	0.21	0.68	0.42	0.36	0.23
	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
Female	0.05	0.08	0.14	0.12				
	(0.00)	(0.00)	(0.01)	(0.00)				
Asian	0.27	0.09	0.33	0.20				
	(0.01)	(0.01)	(0.01)	(0.01)				
Black	-0.12	-0.24	-0.29	-0.08				
	(0.01)	(0.01)	(0.01)	(0.01)				
Hispanic	0.02	-0.09	-0.08	-0.04				
	(0.01)	(0.01)	(0.01)	(0.01)				
Adjusted R^2	0.55	0.58	0.54	0.53	0.54	0.57	0.53	0.52
N	107,129	106,447	107,545	108,240	106,583	106,084	107,331	108,1169

- These weights were obtained from multiple linear regressions to predict reading achievement using 2009-2010 Academic year cohort data
- Each cell contains the associated regression coefficient and standard error

Table 3-5. IRT-based Indicator Parameters and Standard Errors

	1PL								2PL							
	G4		G5		G6		G7		G4		G5		G6		G7	
Indicators	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a
Economically disadvantaged	-0.38 (.01)	1.81 (.01)	-0.05 (.01)	1.64 (.01)	-0.04 (.01)	1.71 (.01)	0.02 (.01)	1.69 (.01)	-0.28 (.01)	1.72 (.03)	0.26 (.01)	1.29 (.02)	0.28 (.01)	1.32 (.02)	0.37 (.01)	1.26 (.02)
Free/reduced lunch eligibility	0.38 (.01)		0.10 (.01)		0.12 (.01)		0.17 (.01)		0.19 (.01)	2.75 (.03)	0.44 (.01)	1.25 (.02)	0.47 (.01)	1.24 (.02)	0.54 (.01)	1.20 (.02)
Targeted assistant	1.45 (.01)		1.77 (.01)		2.17 (.01)		2.34 (.01)		6.38 (.35)	0.33 (.02)	5.70 (.23)	0.40 (.02)	6.54 (.23)	0.44 (.02)	7.10 (.27)	0.43 (.02)
Special education	2.36 (.01)		1.68 (.02)		1.70 (.02)		1.73 (.02)		1.13 (.01)	8.43 (.17)	1.44 (.01)	11.1 (.36)	1.56 (.01)	9.93 (.15)	1.58 (.00)	10.20 (.14)
Limited English Proficiency	2.34 (.02)		2.79 (.02)		2.79 (.01)		2.88 (.01)		3.59 (.14)	1.00 (.05)	5.63 (.32)	0.70 (.05)	5.46 (.20)	0.73 (.03)	6.02 (.25)	0.66 (.03)
Proportion of attendance	2.89 (.02)		2.82 (.02)		2.51 (.01)		2.46 (.01)		2.09 (.02)	3.37 (.10)	2.27 (.02)	3.94 (.12)	2.11 (.01)	4.99 (.13)	2.07 (.01)	5.22 (.12)
Disability	1.22 (.02)		1.64 (.02)		1.66 (.02)		1.68 (.02)		1.10 (.01)	8.19 (.19)	1.35 (.04)	16.0 (4.6)	1.42 (.05)	16.7 (4.5)	1.44 (.03)	16.7 (2.4)
Prior test score on Mathematics	1.94 (.01)		1.21 (.01)		1.46 (.01)		1.49 (.01)		2.41 (.04)	1.18 (.03)	1.69 (.02)	1.29 (.02)	1.97 (.01)	1.34 (.02)	1.94 (.01)	1.42 (.02)
Prior test score on Reading	1.46 (.01)		1.52 (.01)		1.45 (.01)		1.16 (.01)		1.57 (.02)	1.40 (.02)	1.85 (.02)	1.58 (.03)	1.85 (.01)	1.50 (.02)	1.55 (.01)	1.51 (.02)
-2LL	311400.24		352511.24		534237.58		573515.28		288475.89		340229.00		512686.19		548803.58	
Cronbach's α	*.65*		.68		.71		.72		*.65*		.68		.71		.72	

Other indicators including economically disadvantaged group, free or reduced lunch eligibility, targeted assistant school-program eligibility, special education eligibility, and limited English proficiency were negatively associated with achievement. Students who are members of these categories are expected to be more challenging for teachers. Negative coefficients associated with the limited English proficiency group and special education eligibility, tend to be larger for reading achievement than for mathematics achievement. Of note is that the effect of limited English proficiency on mathematics was positive in Grade 4 unless controlling for ethnicity. It may be because younger Asian students who newly migrated to the US tend to perform well in mathematics despite being less proficient in English language, and as grade goes up, the discrepancy between mathematics achievement and English proficiency would lessen. The direction of female weights is opposite between mathematics and reading in Grade 4 and Grade 7.

For IRT calibration, one-parameter logistic models and two-parameter logistic models were applied to calibrate the selected indicators using 2010-2011 academic year data. As a results of IRT calibration using BILOG-MG, indicators' parameters and models' fit indexes are displayed in Table 3-5. Note that several indicators of relevant factors, such as proportion of attendance and prior performance levels in mathematics and reading, were reverse coded. For both mathematics and reading, the economically disadvantaged group indicator was the most frequently applied indicator; compared to other student characteristics, the economically disadvantaged group indicator does not contribute much information for locating students at the higher level of the challenge index.

For one-parameter logistic models, limited English proficiency and proportion of attendance show high challenge level for any grade-level. However, notice that the *b*-parameters

of the one-parameter logistic models are likely to reflect in part the proportion of students who are applicable for each index (see Table 3-1); if the proportion of students who are categorized as economically disadvantaged is large, then the b-parameter tends to be lower. For the two-parameter logistic models, targeted assistant program eligibility, in addition to limited English proficiency and proportion of attendance, seem to contribute more to being located at the higher level of the challenge index. However, the standard errors of targeted assistant program eligibility and limited English proficiency are relatively large. Special education and disability show higher discrimination parameters, compared to the others.

3.2.3. Distributions of the Challenge Index

First, for the OLS weighted sum score, using weights in Tables 3-3 and 3-4, each individual student's predicted achievement was computed based on his/her indicator variables, and the minus (negative value) of the predicted achievement was taken for his/her challenge index for each subject. That is, the higher the predicted achievement, the lower the challenge level represented in the challenge index. Weights were obtained from the previous cohort's data and were used to compute the following cohort's CI. The challenge index was standardized with a 0 mean and a 1 standard deviation by grade-level.

Second, for the IRT calibration, using the fixed indicator-parameters in Table 3-4, each individual student's latent trait, which represents the degree of difficulty that he/she poses to his/her teacher, was estimated for both mathematics and reading in common. The higher the score on the latent trait, the higher the challenge index. The estimated students' challenge index was also standardized with a 0 mean and a 1 standard deviation by grade-level.

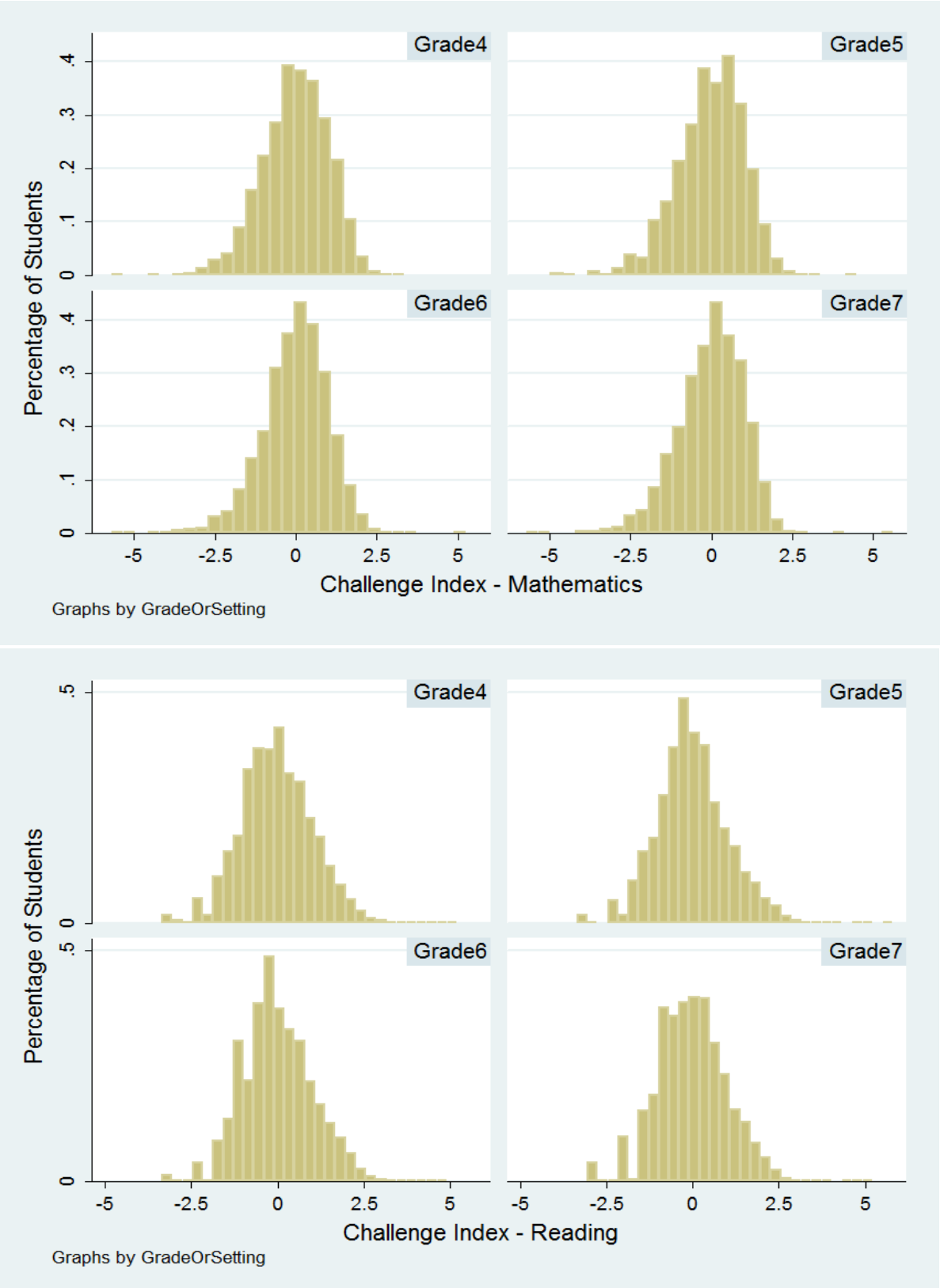


Figure 3-1. Distributions of student challenge index from the OLS weighted sum scores by grade (Compact set of indicators)

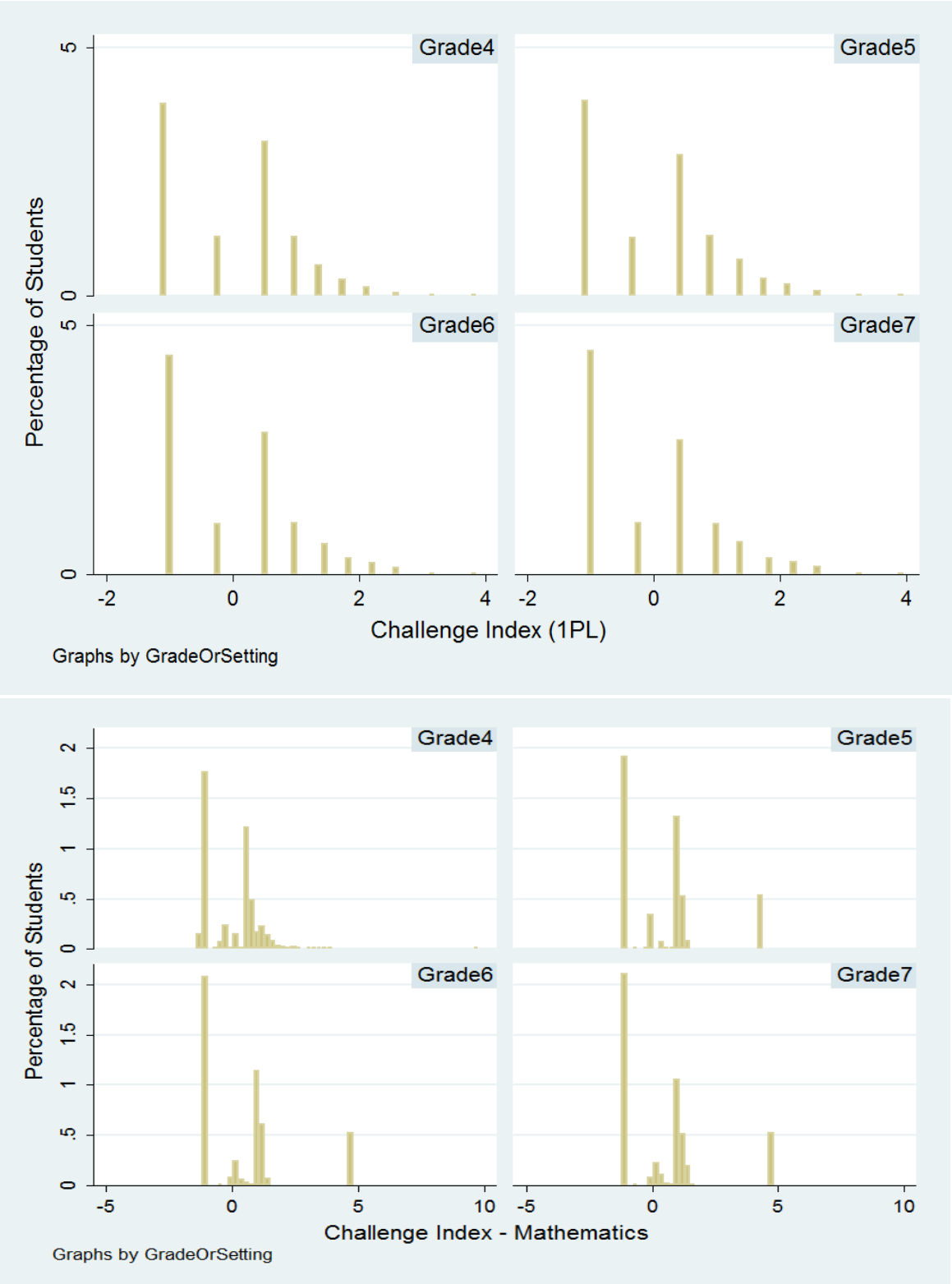


Figure 3-2. Distributions of student challenge index from the IRT calibration by grade (1PL on the top; 2PL on the bottom)

Table 3-1 displays the distributions of the selected indicators for 2010-2011 academic year students whose teachers' effects will be estimated. Distributions of the resulting challenge index, yielded from the OLS weights and IRT calibration, are shown in Figures 3-1 and 3-2 respectively. Based on the observed distributions, the OLS weighted sum score is preferred over the IRT calibration. While the OLS weighted sum scores are bell-curved, with substantial variation to discriminate students according to their challenge level, the results from the IRT calibration show that there to be only a few score categories available. This suggests more indicators are necessary to obtain more variation of the CI across students when using the IRT calibration.

Distributions of student CI by performance level, that is, basic, partially proficient, proficient, and advanced, are displayed in Table 3-6 and in Figures 3-3, 3-4, and 3-5. As shown in Table 3-6, as the performance level increased from basic to advanced, the average CI decreased, which is true for all grade-levels, for both mathematics and reading, and for both the OLS weighted sum scores and IRT calibration. Standard deviations in the OLS weighted sum scores tended to be larger in higher performance levels, proficient and advanced, whereas those in the IRT calibration tended to be larger in lower performance levels, basic and partially proficient. A series of one-way ANOVA and post-hoc analyses were conducted in order to examine whether the CI distributions are statistically different depending on the four performance levels. Every pair of distributions of adjacent performance levels significantly differed in both mathematics and reading as well as in the OLS weighted sum scores and the IRT calibration.

Looking at the OLS weighted sum score distribution in Figure 3-3 (elementary school) and 3-4 (secondary school), as the performance level goes up from basic to advanced, the

Table 3-6. Means and standard deviations of student challenge index by performance level (2010-2011 Academic year)

	Grade 4			Grade 5			Grade 6			Grade 7		
						OLS weights (CS)						
Mathematics	M	SD	N	M	SD	N	M	SD	N	M	SD	N
Basic	0.77	0.63	16,824	0.73	0.62	19,304	0.77	0.59	30,263	0.74	0.60	33,766
Partially proficient	0.06	0.60	10,592	0.07	0.59	13,846	0.10	0.53	21,533	0.00	0.60	29,652
Proficient	-0.72	0.72	15,768	-0.76	0.75	17,495	-0.70	0.72	29,359	-0.87	0.69	22,798
Advanced	-1.90	0.77	1,735	-2.14	0.90	1,589	-2.07	0.93	3,569	-2.00	0.82	4,126
Reading												
Basic	1.27	0.68	4,773	1.22	0.74	7,248	1.21	0.69	12,997	1.28	0.67	9,258
Partially proficient	0.62	0.64	8,687	0.54	0.66	9,063	0.46	0.63	19,597	0.52	0.67	25,661
Proficient	-0.27	0.74	26,074	-0.20	0.67	24,281	-0.35	0.67	39,643	-0.36	0.73	48,769
Advanced	-1.07	0.75	5,770	-0.96	0.69	11,185	-1.08	0.70	13,393	-1.19	0.77	11,036
						IRT calibration (1PL)						
Mathematics												
Basic	0.49	0.94	17,086	0.53	0.96	19,618	0.58	0.97	30,816	0.53	0.99	34,487
Partially proficient	-0.11	0.88	10,670	-0.14	0.86	13,977	-0.14	0.84	21,754	-0.20	0.83	29,924
Proficient	-0.46	0.82	15,900	-0.51	0.75	17,631	-0.52	0.71	29,565	-0.57	0.67	22,951
Advanced	-0.77	0.66	1,740	-0.81	0.57	1,597	-0.79	0.52	3,608	-0.77	0.51	4,146
Reading												
Basic	1.00	0.84	4,945	0.98	0.89	7,491	1.00	0.91	13,415	1.12	0.94	9,631
Partially proficient	0.39	0.88	8,799	0.32	0.89	9,182	0.24	0.89	19,829	0.32	0.95	26,035
Proficient	-0.25	0.87	26,305	-0.23	0.84	24,519	-0.33	0.78	39,964	-0.33	0.78	49,158
Advanced	-0.65	0.73	5,802	-0.60	0.71	11,257	-0.66	0.63	13,490	-0.65	0.60	11,101

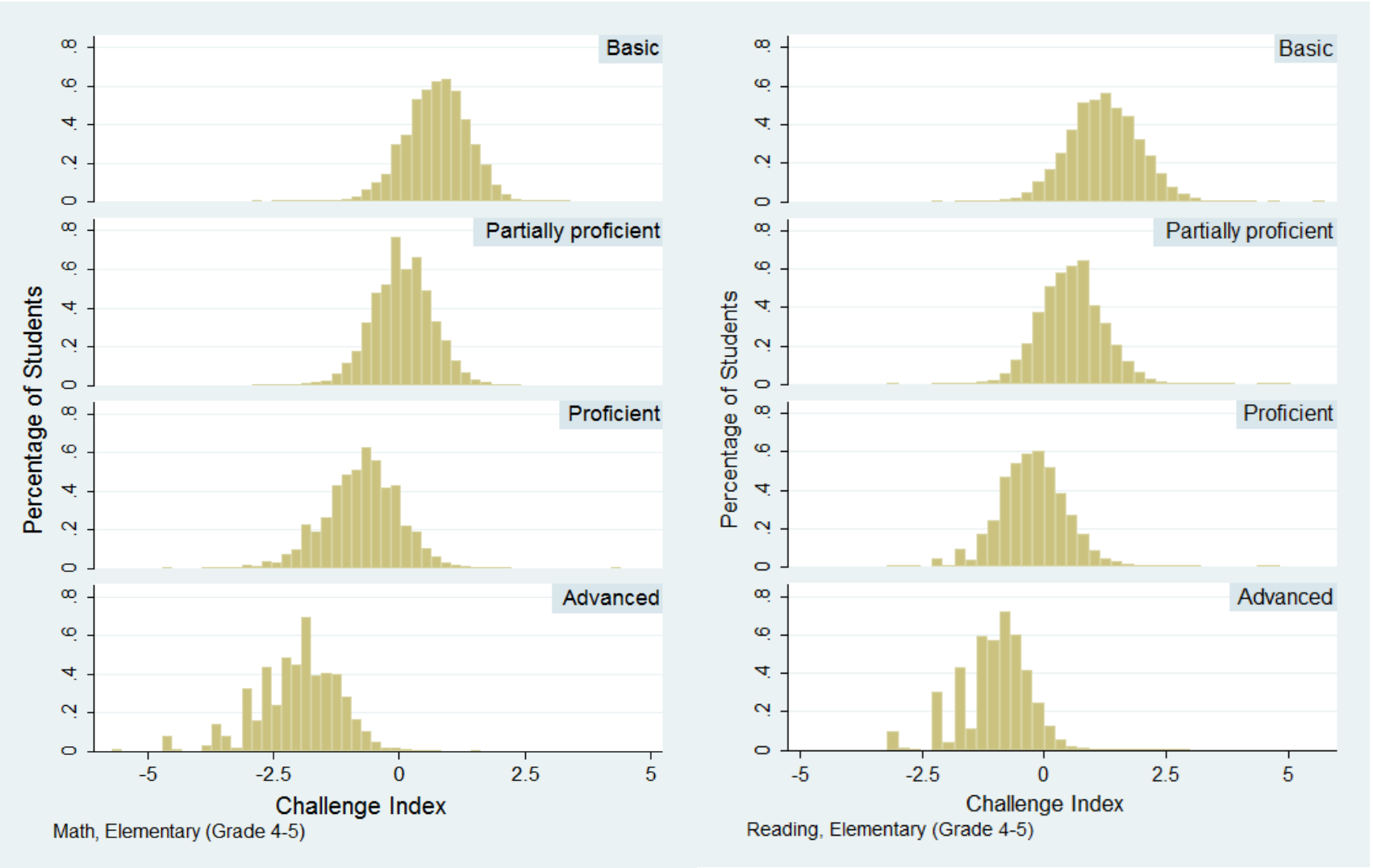


Figure 3-3. Distributions of student challenge index by performance level (OLS weighted sum score; elementary school; mathematics on the left; reading on the right)

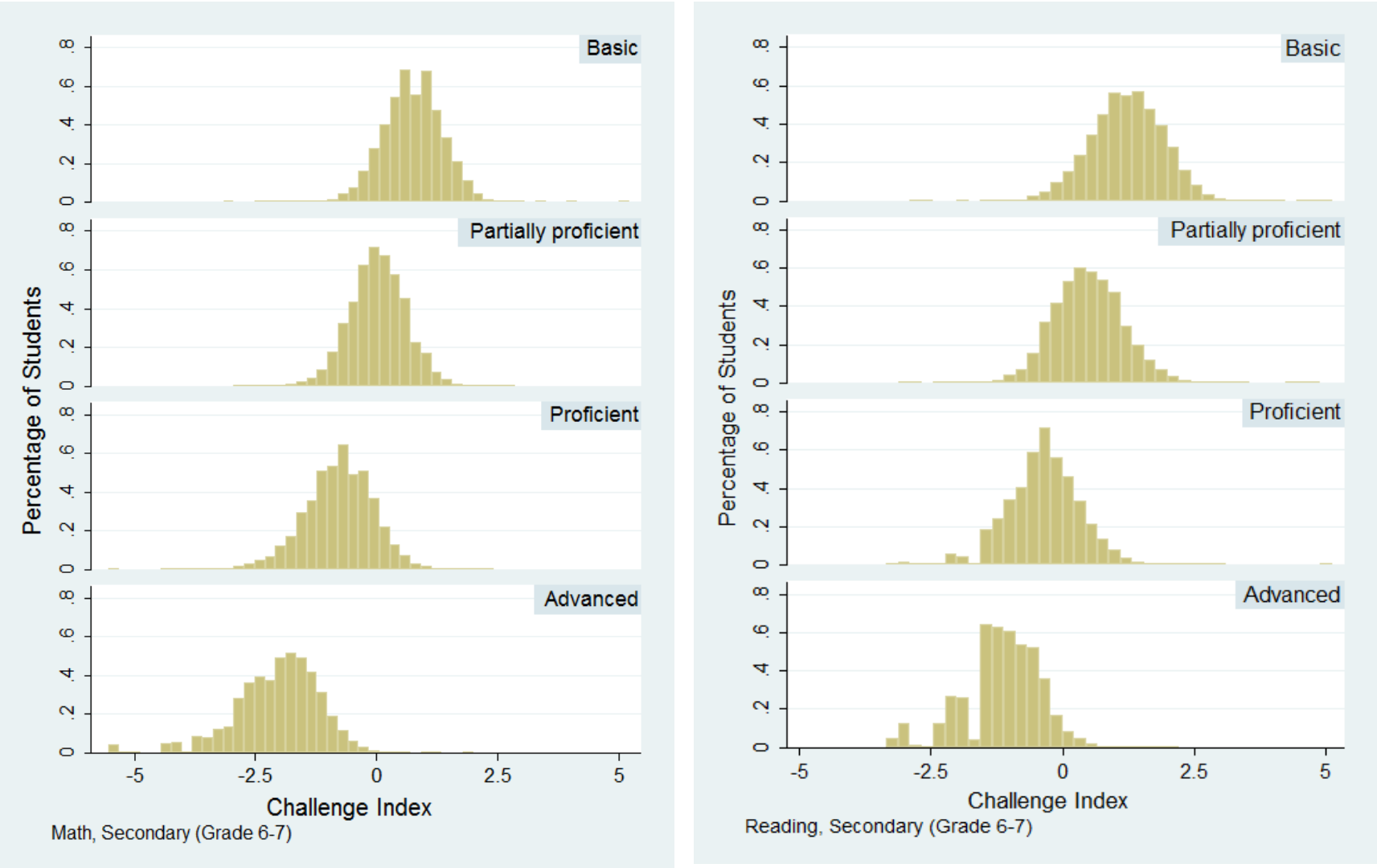


Figure 3-4. Distributions of student challenge index by performance level (OLS weighted sum score; secondary school; mathematics on the left, reading on the right)

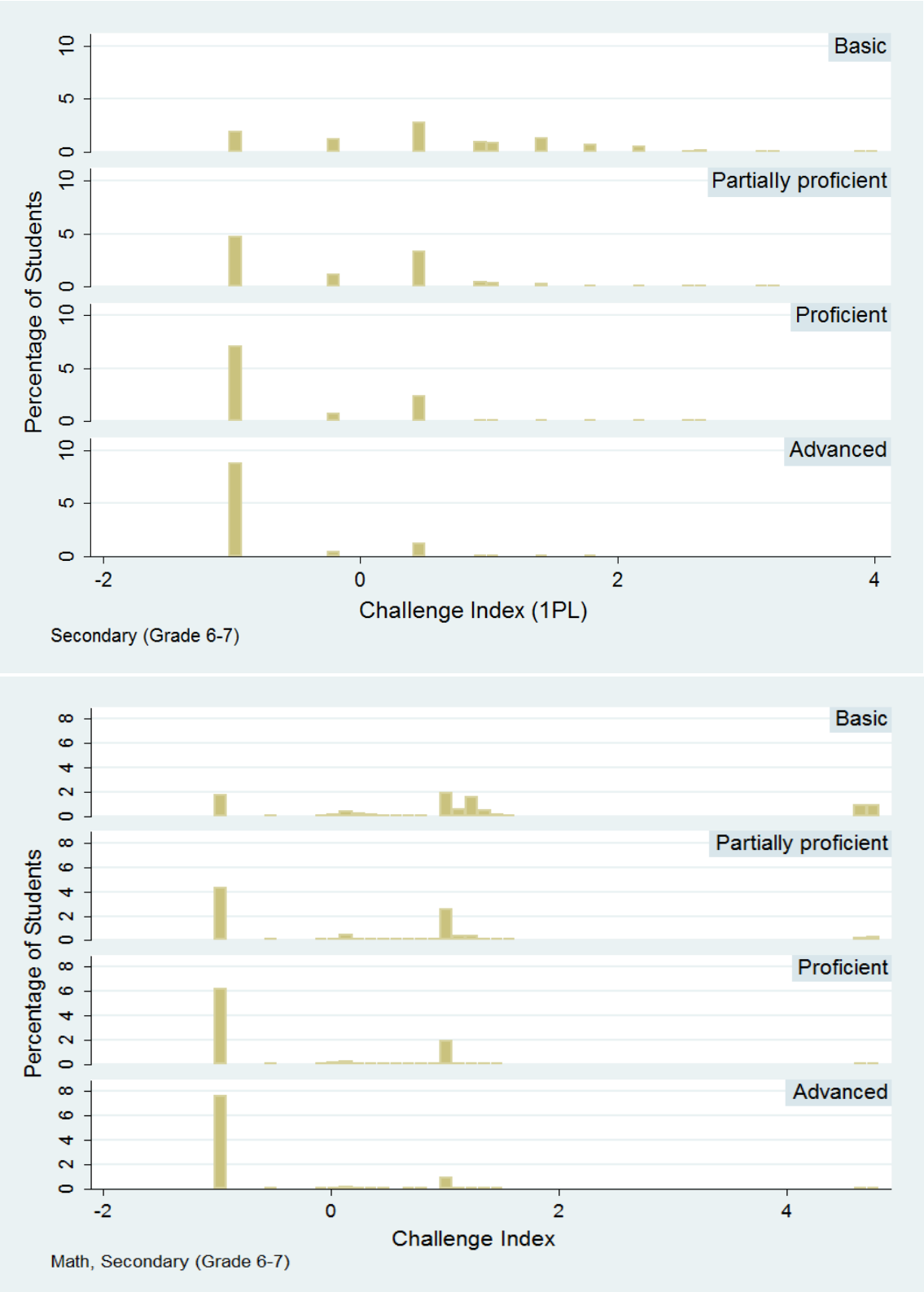


Figure 3-5. Distributions of student challenge index by performance level (IRT calibration; secondary school; 1PL on the top, 2PL on the bottom)

Table 3-7. Correlations between different types of student challenge index and between the challenge index and achievement

	Correlation									
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Mathematics										
1. CI-OLS-CS	1.00	0.98	0.61	-0.97	-0.77	0.67	0.71	0.61	-0.68	-0.64
2. CI-OLS-FS	0.95	1.00	0.62	-0.99	-0.78	0.69	0.71	0.62	-0.69	-0.65
3. CI-IRT-CS	0.56	0.59	1.00	-0.53	-0.49	0.66	0.68	1.00	-0.57	-0.52
4. Prior test score	-0.89	-0.93	-0.47	1.00	0.76	-0.65	-0.68	-0.53	0.66	0.63
5. 2011 test score	-0.76	-0.79	-0.47	0.71	1.00	-0.61	-0.64	-0.49	0.64	0.61
Reading										
6. CI-OLS-CS	0.66	0.68	0.66	-0.60	-0.60	1.00	0.97	0.66	-0.95	-0.71
7. CI-OLS-FS	0.64	0.68	0.66	-0.60	-0.60	0.99	1.00	0.68	-0.98	-0.73
8. CI-IRT-CS	0.56	0.59	1.00	-0.46	-0.47	0.66	0.66	1.00	-0.57	-0.51
9. Prior test score	-0.61	-0.65	-0.54	0.55	0.59	-0.97	-0.98	-0.54	1.00	0.70
10. 2011 test score	-0.58	-0.61	-0.50	0.54	0.61	-0.72	-0.72	-0.50	0.71	1.00

- Elementary school on the lower diagonal; secondary on the higher diagonal
- CS: compact set of indicators; FS: full set of indicators including gender and ethnicity

distributions of CI based on the OLS weighted sum score move toward the left; the higher the performance level, the lower the CI on average. Looking at IRT scores in Figure 3-5 (secondary school) and A-2 (elementary school), the number of students in the lowest challenge level increased, as the performance level increased from basic to advanced.

Correlations between different types of CI, and between CIs and achievement were monitored (see Table 3-7). First, correlations of the OLS weighted sum scores between using the full set of indicators and using the compact set of indicators - including neither gender nor ethnicity - are above .9; no substantial difference existed between the two sets of indicators. Second, the resulting CI using OLS weighted sum scores are moderately correlated with the results of the IRT calibration, showing .56 to .68 correlations. Third, looking at the correlations between the CI and test scores, while the correlations between OLS weighted sum scores and achievement test scores ranged from -.89 to -.99, those between IRT score and achievement scores were between -.47 and -.68 within the same subject; this makes sense because prior test scores were used to compute the OLS CI, while prior dichotomous performance category were used for the IRT CI. Also, the correlations to current achievement were around -.7 for the OLS CI and above -.5, which implies some room for teachers to work better given student CI. Last, the correlation of CI between mathematics and reading ranged from .5 to .7.

3.3. Computing Teachers' Educator Performance Levels and Value-added Measures

The primary goal of this study is to compare the results from the educator performance function (EPERF)-based teacher effect estimation method to those from the education production function (EPROF)-based method. First, for the EPERF-based method, six different teacher effect

estimates were computed depending on the number of performance categories – dichotomous or polytomous performance categories, depending on the number of estimated teacher parameters – 1PL or 2PL, and depending on ways to compose the challenge index (CI) – OLS weights or IRT calibration. Second, for typical examples of the EPROF-based VAM, three different teacher effect estimates were computed under the frame of the covariate adjustment model, shown in Equation 2-4: (1) random effect model, (2) average residual; and (3) gain score model. The same set of student variables was included in the EPROF as covariates, and was also be used for constructing a student challenge index in the EPERF. Model specifications to be compared in this study are shown in Table 3-8, and each specification is illustrated in the following sections.

Table 3-8. Summary of model specifications

	EPERF-based method	EPROF-based method
LHS Student outcome	1. Probability of becoming proficient (compared to non-proficient) 2. Probability of becoming proficient (using 4 different performance categories)	1. Continuous scores – IRT scale scores
RHS Student characteristics	Student challenge index 1. OLS weighted sum scores 2. IRT calibration	A set of student covariates
RHS Teacher effect	1. Random effect of intercept 2. Random effect of slope	3. Random effect of teacher indicators 4. Average residuals by each teachers 5. Gain scores

3.3.1. EPERF-based Teacher Effect Estimation

The general form of educator performance function (EPERF) for dichotomous student performance categories – proficiency or non-proficiency – is presented in Equation 3-1. This is analogous to the two-parameter logistic IRT model (Equation 2-1), which is equivalent to the generalized latent linear and mixed model with random intercept and random slope (Rabe-Hesketh & Skrondal, 2012; Raykov & Marcoulides, 2011).

$$\begin{aligned} \text{logit}[\left(\text{P}\left(\text{Y}_{\text{ij}}=1\right)\right] &= \beta_{0j}+\beta_{1j}X_{ij} \\ \beta_{0j} &= \gamma_{00}+u_{0j}\ ,\ \ u_{0j}\sim N(0,\psi_1) \\ \beta_{1j} &= \gamma_{10}+u_{1j}\ ,\ \ u_{1j}\sim N(0,\ \psi_2\ ) \end{aligned} \qquad \text{Equation 3-1}$$

where Y_{ij} is performance level of student *i* who was working with teacher *j* (0 or 1);
X_{ij} is challenge level of student *i* who was working with teacher *j*;
u_{0j} is random intercept of teacher *j;*
u_{1j} is random slope of teacher *j*

For the one-parameter EPERF (EPERF-D1PL), an individual teacher's effect is defined only by the random effects of intercept, u_{0j}, while the variance in random effects of slope, ψ_2 , is set to be 0, which means that the slope of student CI on the probability of student success, u_{1j}, is assumed to be identical across all teachers and estimated in common with all teachers. The average teacher effect is γ_{00}, and the variance of individual teacher effects, ψ_1, is to be estimated.

In the two-parameter EPERF (EPERF-D2PL), β_{1j} is assumed to vary across teachers, and u_{1j} is estimated as a slope-parameter for each teacher. The resulting β_{1j} will be mostly negative because the probability of students' successes decreases as their challenge index increases, and corresponds to the minus of *a*-parameter in Equation 2-1. Note that the varying

slope for each teacher in the EPERF-D2PL is based on the idea that the impact of CI on their students' successes can be differentiated among teachers. That is, the statistical significance of ψ_2 – the variance of u_{1j} – implies whether individual teachers are differentiated in mediating the relationship between student CI and success. As a result, the two different parameters for each individual teacher, u_{0j} and u_{1j}, were obtained from this model. For identification ψ_1 was set to 1 in this model.

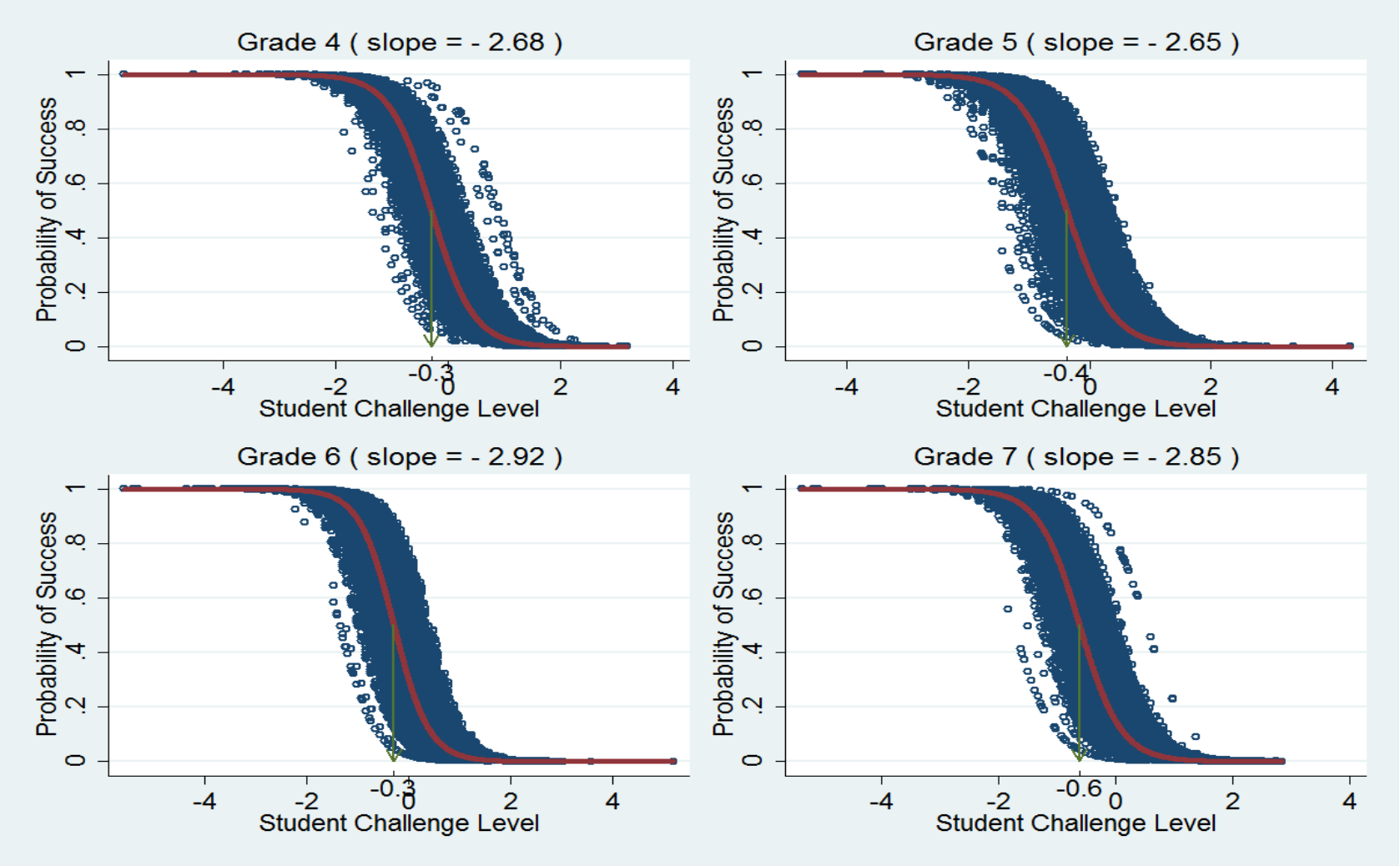
The interpretation of u_{0j} is straightforward; each individual teacher's deviation from the average logit of the probability that students whose CI is average success. Accordingly, supposing that the estimate of u_{0j} is larger for teacher A than for teacher B, it is concluded that teacher A's expected or average probability of success is higher than teacher B regardless of student CI. The interpretation of u_{1j} is somewhat complicated. In a simple way, the regression coefficient, β_{1j}, represents the amount of decreasing probability of success when one unit of student CI increases; u_{1j} is each individual teacher's deviation from the average decreasing amount of probability of success associated with one unit increase in the student CI, γ_{10}. This also can be considered an interaction term. Supposing that we compare the two teachers whose u_{0j} estimates are identical but u_{1j} estimates are different from each other: teacher A's u_{1j} is -.3; teacher B's is -.5. While the probability of getting their students whose CI were below the teachers' u_{0j} estimate success, is larger in teacher A than in teacher B, the probability of getting their students whose CI were above their u_{0j} estimate success, is large in teacher B than in teacher A. In this case, the teachers are mediating the relationship between the students' CI and their successes.

In both models, individual teachers' random effects of intercept were transformed to IRT-typed person parameters using the following transformation: $\theta_j = \beta_{0j}/-\beta_{1j}$, which rescales the teacher random effect of intercept onto the scale of student CI, X_{ij}. Again, while in the 1PL model, β_{1j} is common across teachers, it varies across teachers in the 2PL model. The transformed individual teacher's educator performance level (EPL), θ_j, indicates the point of student CI corresponding to a probability of 0.5 that a teacher has a student succeed to achieve a target performance level.

Some examples of the results of fitting student performance data to the above models are displayed in Figure 3-6 and 3-8 for the EPERF-1PL, and Figure 3-8 for the EPERF-2PL; variation in individual teachers' slopes in the EPERF-2PL is observed in Figure 3-8. They represent individual students' expected probabilities of success in reaching proficient level given their teachers' estimated EPL. The fitted line, called student characteristics curves (SCC), represents conditional probability of success given student CI, when their teachers' EPL is equal to 0. As students' challenge levels increase, the probability of success in achieving the proficient level, which the state benchmark requires, tends to decrease. Note that the point on the scale of student CI corresponding .5 probability of success varies across different grade levels as well as different subjects. For instance, while for Grade 4, the probability that students with -.3 CI would achieve the proficient level in mathematics when teachers' EPL is fixed at 0, was .5, for Grade 6, the probability that students with -.3 CI obtain the proficient level was smaller than .5; the probability of success for students with -.6 CI was .5. In reading (see Figure 3-8), CI points associated with .5 probability of success were significantly higher than in mathematics: all CI points corresponding to .5 probability were negative in mathematics, whereas those were all

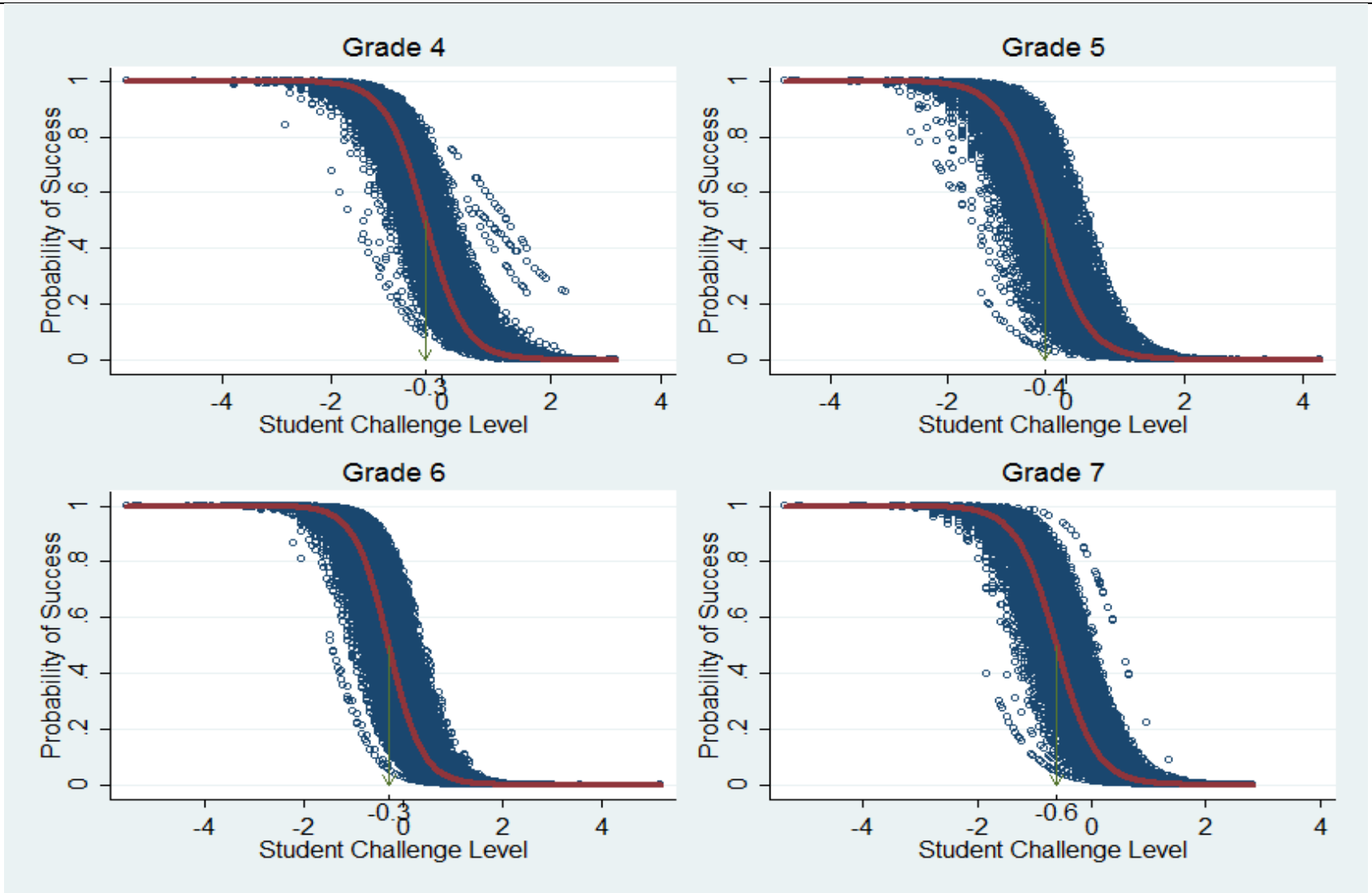
positive in reading. This suggests that the likelihood for students with higher CI to achieve proficiency is higher in reading than mathematics. Another observation is that variance in the probability of success is smaller in reading than in mathematics. Details of the interpretation are discussed in Chapter 4.

Figures 3-10 to 3-11 display the teacher characteristics curves (TCC), which describe the relationship between teachers EPL and the probability of their students' successes. The red line indicates the expected probability of success according to teachers' EPL when student challenge index is equal to 0. And each dot represents each teacher. As teachers' estimated performance level escalates, the average of the probability that their students' success in reaching the proficient level tends to increase. The brightness of circles indicates the average challenge index of the students who worked with each teacher; the darker the circles, the more challenging the teacher's students. As shown, when teacher EPL is similar, the probability of success tends to decrease along with their students' average challenge levels increase. Variance in the estimated teacher EPL in reading appeared not substantial as Figure 3-12 shows.



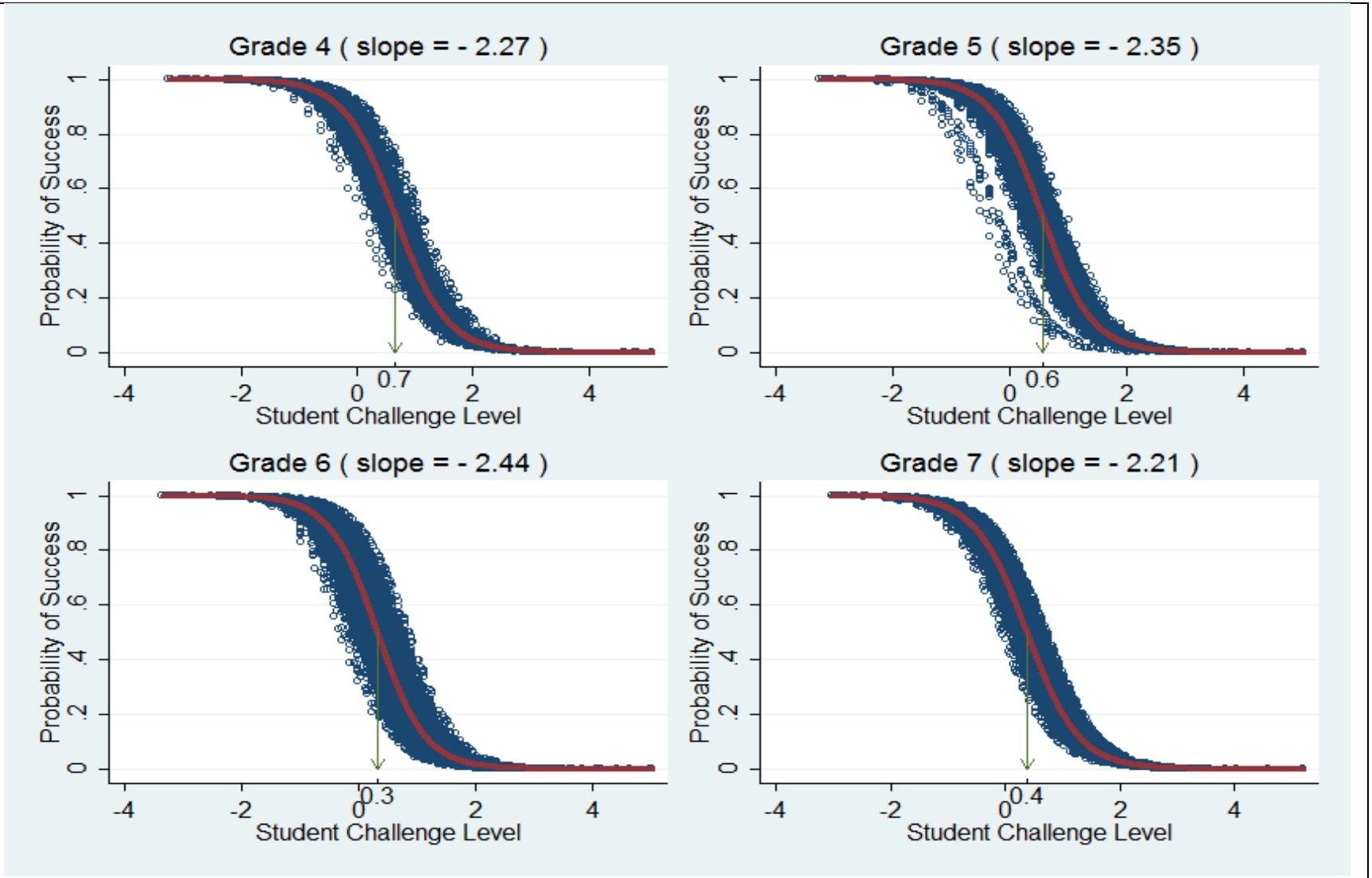
- For interpretation of references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.
- Each blue mark represents the estimated probability of individual student's success; and the center red line represents the conditional probability of success when teacher EPL=0

Figure 3-6. Student characteristic curve of the EPERF-1PL with the compact set of indicators by grade (Mathematics)



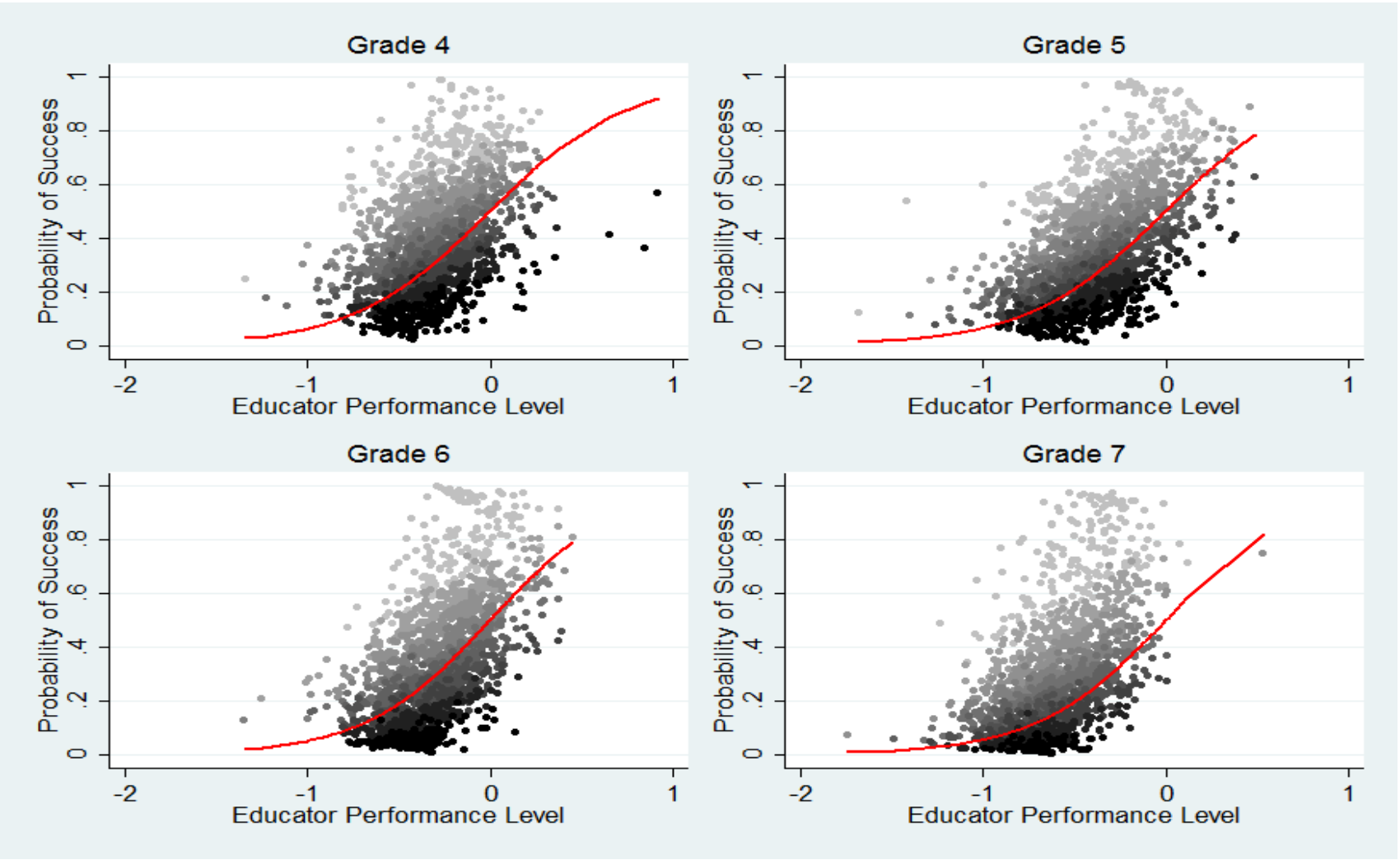
- Each mark represents the estimated probability of individual student's success; and the center line represents the conditional probability of success when teacher EPL=0

Figure 3-7. Student characteristic curve of the EPERF-2PL with the compact set of indicators by Grade (Mathematics)



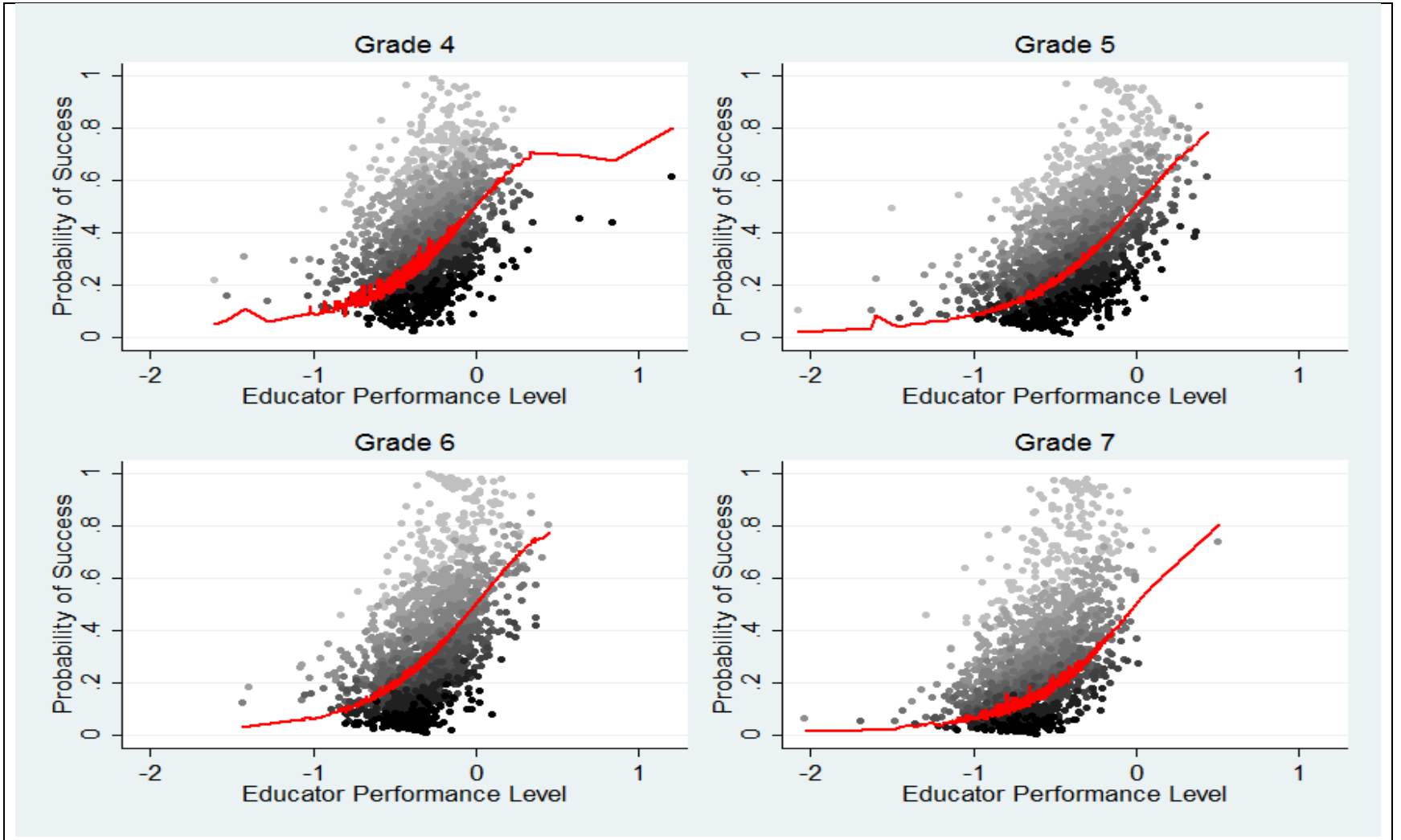
- Each mark represents the estimated probability of individual student's success; and the center line represents the conditional probability of success when teacher EPL=0

Figure 3-8. Student characteristic curve of the EPERF-1PL with the compact set of indicators by grade (Reading);



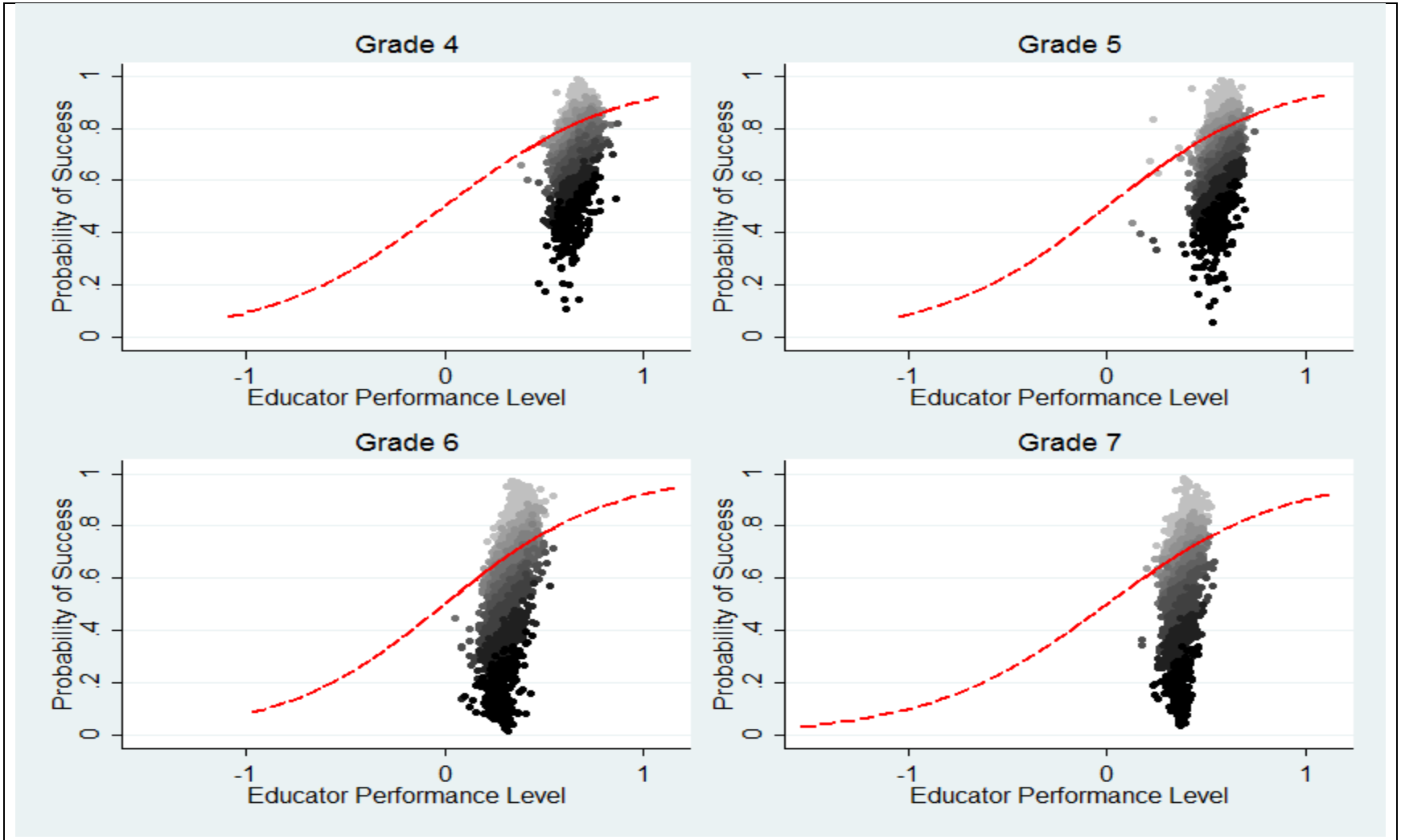
- Each mark represents individual teacher;
- The brightness of the marks represents the average students' challenge index (The more saturated the gray, the higher the average challenge levels);
- Red lines represent the expected probability of success when student challenge index=0

Figure 3-9. Teacher characteristic curve of the EPERF-1PL with the compact set of indicators by grade (Mathematics)



- Each mark represents individual teacher;
- The brightness of the marks represents the average students' challenge index (The more saturated the gray, the higher the average challenge levels);
- Red lines represent the expected probability of success when student challenge index=0

Figure 3-10. Teacher characteristic curve of the EPERF-2PL with the compact set of indicators by grade (Mathematics)



- Each mark represents individual teacher;
- The brightness of the marks represents the average students' challenge index (The more saturated the gray, the higher the average challenge levels);
- Red lines represent the expected probability of success when student challenge index=0

Figure 3-11. Teacher characteristic curve of the EPERF-1PL with the compact set of indicators by grade (Reading)

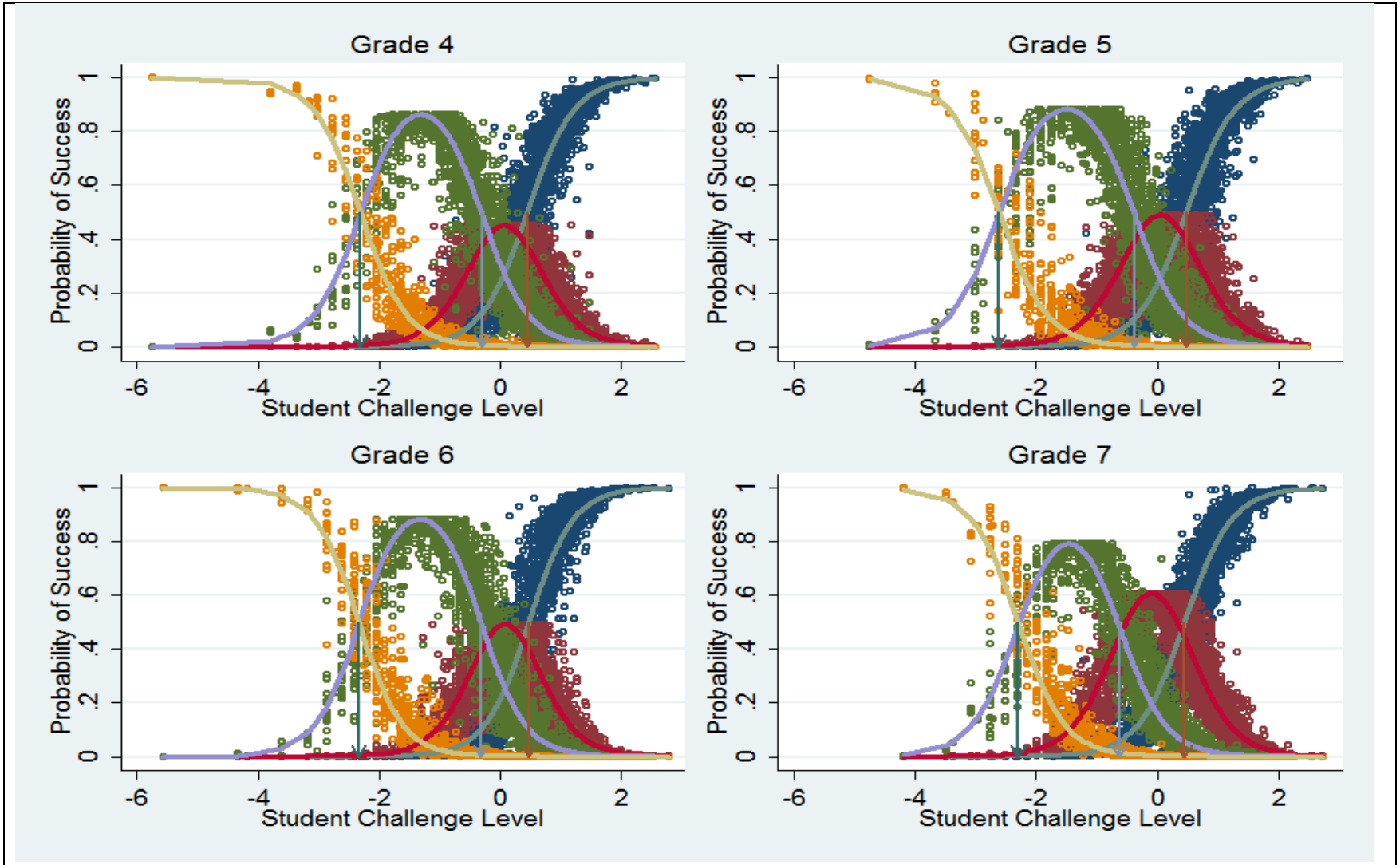
Polytomous performance categories also can be utilized likewise, as polyotomous IRT models for more than three category responses. The general form of EPERF for polytomous student performance categories – for example, advanced, proficient, partially proficient, and basic – applying ordinal logistic regressions based on cumulative distribution function (CDF) is shown in Equation 3-2.

$$\text{logit}[(\mathrm{P}(\mathrm{Y_{ij}} > \mathrm{k})] = \beta_{0j} + \beta_{1j}(X_{ij} - \tau_k)$$
$$\beta_{0j} = \gamma_{00} + u_{0j}\ ,\quad \theta_{0j} \sim N(0, \psi_1) \qquad \text{Equation 3-2}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}\ ,\quad \theta_{1j} \sim N(0,\ \psi_2\ )$$

where $\mathrm{Y_{ij}}$ is performance level of student *i* who was working with teacher *j* (1, 2, 3 or 4);
X_{ij} is challenge level of student *i* who was working with teacher *j*;
u_{0j} is random intercept of teacher *j;*
u_{1j} is random slope of teacher *j;*
τ_k is threshold for category *k*

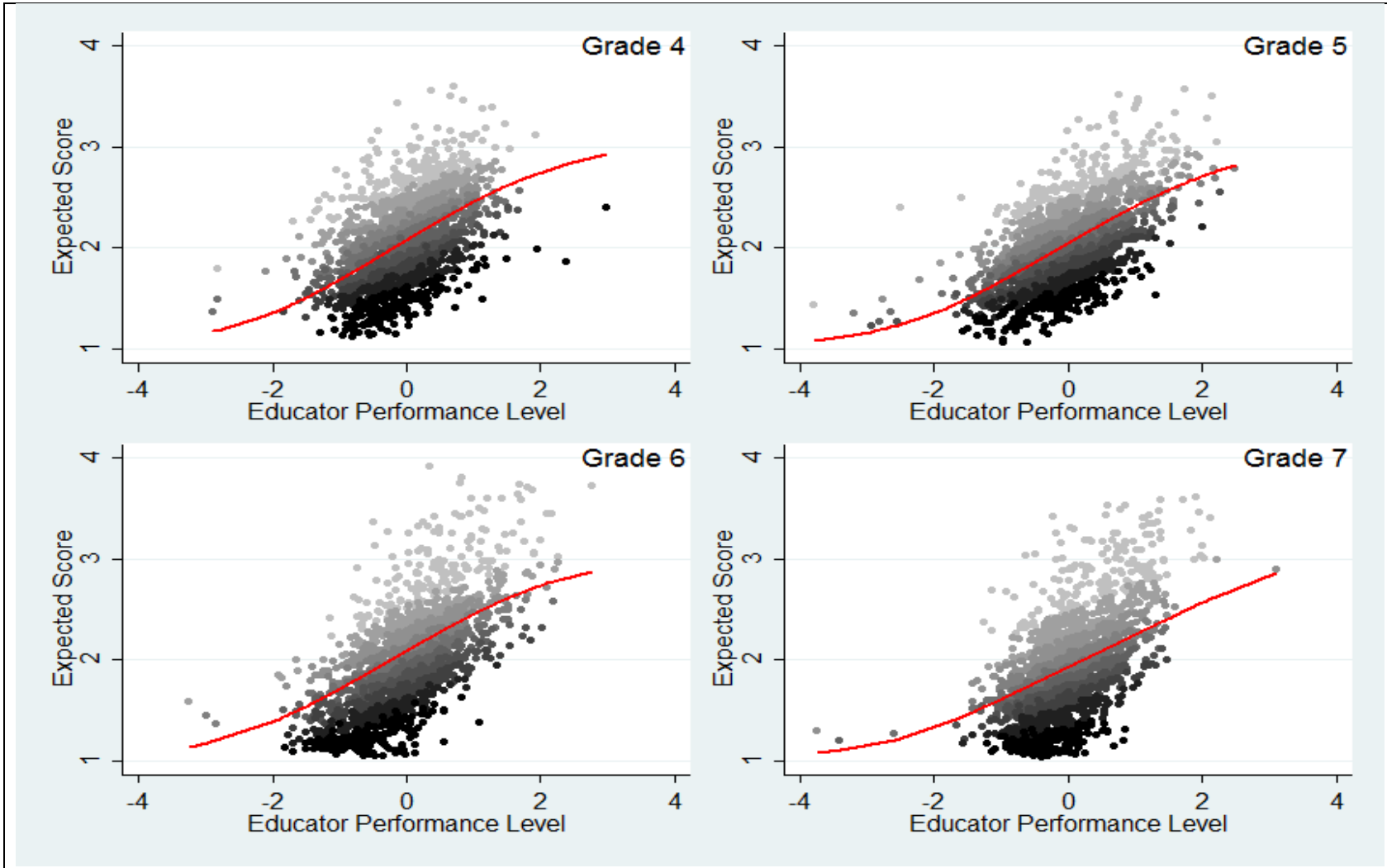
β_{1j} is constant across categories, which implies that linear predictors for different categories are parallel. This restriction is equivalent to that of Samejima's grade response model (1974). In this model, each threshold, τ_k, is the point where $\mathrm{P}(\mathrm{Y_{ij}} > \mathrm{k})$ is equal to 0.5, and can be converted into the scale of student CI using the following transformation: τ_k/γ_{10}. The probability of attaining k-th level, $\mathrm{P}(\mathrm{Y_{ij}} = \mathrm{k})$, is obtained from $\mathrm{P}(\mathrm{Y_{ij}} > \mathrm{k} - 1) - \mathrm{P}(\mathrm{Y_{ij}} > \mathrm{k})$. As in the dichotomous models, while for the EPERF-P1PL, β_{1j} is assumed to be equal across all teachers, for the EPERF-P2PL, it is estimated for each teacher. STATA gllamm command (Zheng & Rabe-Hesketh, 2007) was used for all the estimations.

Figure 3-12 illustrates the fitted models using the polytomous performance category for mathematics by grade. In each grade-level, the very left curve monotonically decreasing indicates the probability of achieving "advanced" level (category 4); the second left uni-modal curve represents the probability of reaching "proficient" level (category 3); the third left uni-modal curve is the probability of becoming "partially proficient" level (category 2); and the very right curve monotonically increasing represents the probability of being "basic" level (category 1). While the probability of attaining basic level increases along with student challenge level, the probability of achieving advanced level decreases as student challenge level increases when teacher EPL is set to 0. The probability of being diagnosed as either partially proficient or proficient levels goes up to the threshold, and goes down after the threshold. Figure 3-13 shows the expected average scores of students depending on their teachers' estimated EPL. The expected score was calculated as follows: $\sum_{k=1}^{k=m} \mathrm{P}\left(\mathrm{Y_{ij}} = k\right) \cdot k$ (*m* is the number of performance levels). According to the fitted functions, as the estimated teacher EPL increases, their students' average expected scores also increase up to 3 which indicates "proficient level", not 4, "advanced level". Further information of the fitted models and individual teachers' EPL are evaluated and discussed in Chapter 4.



- From the left to the right, each curve represents advanced, proficient, partially proficient, and basic respectively.
- Intersections of the curves represent the thresholds
- 3,000 random sample

Figure 3-12. Student characteristic curve of the EPERF-P1PL with the compact set of indicators by grade (Mathematics);



- Each mark represents an individual teacher
- Brightness of the marks represents the average students' challenge index (The more saturated the gray, the higher the average challenge levels)
- Red lines represent the expected score given EPL when student challenge index=0

Figure 3-13. Teacher characteristic curve of the EPERF-P1PL with the compact set of indicators by grade (Mathematics)

3.3.2. EPROF-based Teacher Effect Estimation

Based on the covariate adjustment model in Equation 2-3, individual teachers' random effects and average residuals were computed respectively. For random effect estimation, multi-level models were constructed considering students as level-1 and teachers as level-2. Each teacher's random effect of intercept, after controlling student background variables, was estimated using maximum-likelihood estimation. For average residual estimation, student-level residuals when predicting achievement on a set of student background variables were averaged by each teacher. The mean of the student residuals was regarded a teacher's value-added measure. Last, using student gain scores as outcome variables, teacher fixed effects were computed. A large set of teacher dummy variables was included, and the coefficients associated with each teacher dummy variables were considered individual teachers' value-added measures on their students' gain scores. Students' IRT-scale scores from the state test were used as the outcome variable of the EPROF-based model. IRT-scale scores' distributions are shown in Table 3-9.

Table 3-9. Student IRT-scaled scores on the 2011 state test (by grade and subject)

	Mathematics					Reading				
	M	SD	Min	Max	Skew-ness	M	SD	Min	Max	Skew-ness
Grade 4	.97	1.10	-5.09	6.24	.64	1.06	1.15	-5.12	4.91	.01
Grade 5	1.07	.92	-2.11	6.55	.89	1.16	1.17	-4.85	5.26	.11
Grade 6	1.20	1.04	-3.35	6.68	.76	.96	1.23	-5.12	4.82	.26
Grade 7	.68	.93	-4.78	6.15	.78	1.29	1.02	-4.44	5.44	.04

3.4. Comparison between the EPL and VAM estimates

The teacher effect estimates from the educator performance function (EPEFR)-based method, educator performance levels (EPL), and those from the education production function (EPROF)-based method, value-added measures (VAM), were compared using the following three aspects: 1) distributions and rank of teacher effect estimates; 2) relationship of the estimates to different student characteristics; and 3) consistency of the estimates between different subjects.

First, the distributions and rank correlations of different teacher effects were compared in order to check how consistently the models rank the individual teachers. It is expected that the EPL and VAM are moderately and positively correlated. If the correlation is too high, it may be concluded that the EPL is not differentiated from the VAM in terms of their resulting teacher ranking, and that the unique information of the teacher capability that the EPERF-based method provides over the EPROF-based method is small. Contrarily, if the correlation is too low, criterion validity evidence of the EPL may be threatened.

Second, how the EPL and VAM estimates were associated with major student background variables or with teacher background variables, were investigated. For example, the better teacher effect estimates are expected to not be highly correlated with any student characteristic, and to be correlated with relevant teacher qualifications or characteristics.

Last, intra-person (teacher) rank correlations were monitored within each model as measures of consistency. The extent to which estimates of each individual teacher change across different subjects (mainly in elementary schools) or different classrooms (mainly in secondary schools), was evaluated. Correlations are expected to be moderate or high rather than very low,

under the assumption that teacher capability is stable or gradually changes over the years rather than dramatically improves or declines.

3.5. Examination of the Model Fits of the EPERF

For evaluation of the model fit of the educator performance function (EPERF), the following four subtopics are addressed: (1) fit of the logistic regression models; (2) conditional independence of students; and (3) dependency of student success for the same teacher.

First, it is necessary to check on how well data on students' successes or failures fit the EPERFs. The following three indicators were used to evaluate the goodness of fit of the EPERFs: (1) student-level residuals; (2) teacher-level residuals; and (3) error rate (Gelman & Hill, 2007; Hosmer et al., 1997). For student-level residuals, individual students' deviances between their observed performance and predicted probabilities of succeeding were averaged within each homogenous group; this group was determined according to either similar challenge levels or similar predicted probability of succeeding, as shown in Equation 3-3. The residuals for *g* student groups were plotted.

$$residual_g = \frac{\sum_{n=1}^{N_g}\left[O_{ig} - P_{ig}\left(\hat{\theta}_j\right)\right]^2}{N_g} \qquad \text{Equation 3-3}$$

N_g, number of students in group *g*
O_{ig}, observed performance of *i*-th student in group *g* (0 or 1)
$P_{ig}\left(\hat{\theta}_j\right)$, expected probability that *i*-th student in group *g* successes given teacher *j*'s estimated EPL, $\hat{\theta}_j$ which is predicted from teacher *j*'s fitted EPERF

For teacher-level residuals, the mean of deviations between students' observed performance and their expected probability of succeeding based on the EPERF was computed for each teacher as shown in Equation 3-4. Basically, it is inversely proportional to how many students per teacher on average showed the expected performance predicted from each EPERF, given the student's CI. If a teacher's EPERF perfectly fits the data, the residual will be '0'. Teacher-level residuals by the level of EPERF also were plotted and checked to see if any noticeable pattern exists. In line with that, the assumption of monotonically increasing probability along with the increasing estimated EPL can be monitored.

$$residual_j = \frac{\sum_{n=1}^{N_j}\left[O_{ij} - P_i\left(\hat{\theta}_j\right)\right]^2}{N_j} \qquad \text{Equation 3-4}$$

N_j, number of students worked with teacher *j*
O_{ij}, observed performance of student *i* worked with teacher *j* (0 or 1)
$P_i\left(\hat{\theta}_j\right)$, expected probability that student *i* in teacher *j*'s class successes given teacher *j*'s estimated EPL, $\hat{\theta}_j$ which is predicted from its fitted EPERF

The error rate indicates the proportion of cases for which the fitted model's prediction is wrong. That is, for the dichotomous case, if O_i=1 but the predicted probability is smaller than 0.5, or if O_i=0 but the predicted probability is larger than 0.5, the case is counted as a misclassification. This rule is represented in Equation 3-5. The error rate was compared to that of the null model, which is simply to assign the same probability to each case without any predictor, and so the error rate of the null model is the proportion of 1's in the data, P=$\sum_{i=1}^{N} O_i/N$ or its compensation of 1, whichever is smaller. For the polytomous case, the expected scores were used for obtaining the error rate, shown in Equation 3-6.

$$ER_j(dichotomous) = \sum_{i=1}^{N} m_{ij} \Big/ N_j \qquad \text{Equation 3-5}$$

$m_{ij} = 1$, if $P_i(\hat{\theta}_j) > 0.5$ & $O_{ij} = 0$ or
$P_i(\hat{\theta}_j) < 0.5$ & $O_{ij} = 1$
$m_{ij} = 0$, otherwise

$$ER_j(polytomous) = \sum_{i=1}^{N} m_{ij} \Big/ 4N_j \qquad \text{Equation 3-6}$$

$m_{ij} = 1$, if $(E_i(\hat{\theta}_j) < 1.5)$ & $(O_{ij} > 1)$ or
$\left(1.5 \le E_i(\hat{\theta}_j) < 2.5\right)$ & $\left(O_{ij} = 1, 3, or\ 4\right)$ or
$\left(2.5 \le E_i(\hat{\theta}_j) < 3.5\right)$ & $\left(O_{ij} = 1, 2, or\ 4\right)$ or
$(E_i(\hat{\theta}_j) \ge 3.5)$ & $(O_{ij} < 4)$
$m_{ij} = 0$, otherwise

Second, conditional independence of student successes within each teacher was evaluated applying the concept of Yen's Q_3 (1984). Basically, the Q_3-index is a measure of dependency among the item-level residuals. When conditional independence is true, the expected distribution of Q_3 is normal with a mean of 0 and a variance of $1/(N-3)$, where *N* is the number of students. In the context of the EPERF, because each student is assigned to only one teacher or at most to two or three, in each subject-matter, Q_3 cannot be computed in the usual way such as when each item is administered to all examinees or at least multiple examinees. In order to approximate the Q_3-like-index in this context, students were divided into 50 quantile groups or 25 quantile groups according to their challenge index, so that students in the group were regarded as the same difficulty items, that is, approximately equally-challenging students. As shown in Equation 3-7, each student's residual, d_{ij}, was obtained first, and the residuals were averaged by

each teacher and each quantile group. Then, the correlations of the teacher-level residuals between different quantile groups were computed, and the distribution of the resulting 1,225 (=$50 \times 45/2$), or 300 (=$25 \times 24/2$) correlation coefficients was observed.

$$d_{ij} = O_{ij} - P_i(\hat{\theta}_j)$$
$$d_{gj} = \sum_1^{N_{gj}} d_{ij} \,/N_{gj} \qquad \text{Equation 3-7}$$
$$Q_{3gg'} = corr\,(d_g, d_{g'})$$

d_{ijg} is a residual for a student *i* working with teacher *j* in the *g*-th quantile group

Lastly, dependency among the student success for the same teacher was approximated by conditional intra-class correlation or residual intra-class correlation ρ of the student success (Rabe-Hesketh & Skrondal, 2011).

$$\rho \equiv Cor(\xi_{ij}, \xi_{i'j}) = \frac{\psi}{\psi + \pi^2/3} \qquad \text{Equation 3-8}$$

For the VAM-RE, regular conditional intra-class correlation (Rabe-Hesketh & Skrondal, 2011; Raudenbush & Bryk, 2002) was computed according to Equation 3-9.

$$\rho \equiv Cor(y_{ij}, y_{i'j}) = \frac{\psi}{\psi + \theta} \qquad \text{Equation 3-9}$$

3.6. Effects of Locations of Cut-scores and the Number of Performance Categories

How the locations of cut-scores and the number of performance categories in a test affect the educator performance function (EPERF)-based teacher effect estimation were simulated with real data. First, in order to examine the effect of the locations of cut-scores, several scenarios were planned depending on different proportions of students divided into each category. In the case of dichotomous performance categories, only one cut-score for each scenario was fixed to a score-point, which breaks the students into the two groups, i.e., non-mastery and mastery, in the ratio of 7:3, 6:4, 5:5, 4:6, or 3:7 (five scenarios) in order of their rank of achievement scores. To be concrete, at first students were grouped in 10 quantile categories based on their mathematics scores, the quantile categories were merged to make the two groups according to the above ratios. Likewise, in the case of four performance categories, three cut-scores for each scenario were set to the three score-points, which divides the students into the four groups, i.e., not proficient, partially proficient, proficient, and advanced, in the ratio of 4:3:2:1, 3:3:2:2, 3:3:3:1, 2:3:4:1, 1:4:4:1, 1:3:5:1, 2:2:4:2, or 1:2:5:2: (eight scenarios) in order of their rank of achievement scores. Proportions were decided considering practicality. The major concern in these simulations is how the ranks of the teacher effect estimates change across different scenarios. The intra-person correlation of teacher effect estimates between different scenarios, and whether there are teachers whose estimates substantially shift across different scenarios, were observed.

Regarding the number of performance categories, the four scenarios – the cases of existing six, seven, eight, and ten performance categories – were evaluated, in addition to the cases of two and four performance categories in the original data. Data already defined the four performance categories, and each two adjacent categories, that are basic and partially proficient,

and proficient and advanced, respectively were merged into one to create the two performance category case. For the six-category scenario, students grouped according to the original dichotomous categories were divided into three groups within each category in equal proportion, which created six groups. For the eight-category scenario, students grouped according to the original four categories were divided into two groups within each category in equal proportion, which produced eight groups. For the seven-category scenario, the two highest performance categories in the eight-category scenario, were merged into one group. For the ten-category scenario, students grouped in the two-category case were divided into five groups within each category in the same proportion, which created ten groups. The secondary interest is correlation between estimated teacher effects and student prior test scores in each scenario. That is, whether the increasing number of performance categories mitigates any potential association between teacher effect estimates and assigned students' prior achievement, was investigated.

CHAPTER 4. COMPARISONS OF THE TEACHER EFFECT ESTIMATES

In this chapter, results from the educator performance function (EPEFR)-based method and the educational production function (EPROF)-based method, are compared with respect to the following three aspects: (1) distributions and rank of teacher effect estimates; (2) relationship of the estimates to different student and teacher characteristics; and (3) consistency of the estimates between different contexts. At the end of this chapter, supplementary information on the estimated educator performance level (EPL) that the EPERF-based method produced, which is potentially advantageous over using the EPROF-based method, is also elucidated.

4.1. Distribution and Rank Correlation

Descriptive statistics of the estimated teacher effects - the educator performance levels (EPLs) from the four EPERF-based models and the value-added measures (VAMs) from the three EPROF-based models - are displayed in Table 4-1 for mathematics and Table 4-2 for reading. Also, rank correlations of estimates between models were observed to check how consistently the models rank the teachers. The EPL estimates resulting from using the OLS weighted challenge index (CI) with a compact set of indicators were presented in this chapter. They almost matched those resulting from using the OLS weighted CI with a full set of indicators. And the IRT calibrated CI was not considered as a reasonable approximation of student challenge level (see Figure 3-2).

The means of the EPL from dichotomous outcomes (EPL-D1PL and EPL-D2PL) are smaller than 0 in mathematics, and larger than 0 in reading, while other estimates' means are 0. It

needs to be noticed that the means of EPL-D1PL and EPL-D2PL were adjusted by the scale of the students' CI, instead of being set arbitrarily by each regression model. In the EPERF-based methods with dichotomous outcomes, for example, the -.3 mean of teachers' EPL in Grade 4 mathematics refers to the .5 probability that average teachers will have success in helping their students with CI of -.3 to pass a desired proficiency level. Likewise, the .3 mean of teachers' EPL in reading indicates that the probability that average teachers will have success in helping students with CI of .3 to pass a desired standard is .5. Accordingly, supposing that students' CIs are fixed, the probability that average teachers are successful in helping students obtain the proficiency level in mathematics is higher than in reading. For the EPERF-based methods with polytomous outcomes (EPL-P1PL and EPL-P2PL), the interpretation can be made by each category based on the estimated thresholds, as discussed further in section 4.5.

Meanwhile, the mean of VAM is assigned to the average teachers whose average students' test scores increased as much as did other teachers' average students.[3] From regression models used for obtaining VAMs, the average students represent those whose values on a set of background variables used as covariates are average for all students involved; and the average gain of test scores indicates average difference scores in the tests.

Estimated variances of teacher effects are very similar among the different models. Variance in mathematics tended to be marginally larger in the EPL, particularly in the case of using polytomous categories. Variances in reading were substantially smaller compared to mathematics, and tended to be larger in the VAM than in the EPL. The amount of variation in the EPL using the four performance categories as an outcome was close to those in the VAM using

[3] Raudenbush & Jean (October, 2012) How should educators interpret value-added scores? Carnegie Knowledge Network Knowledge Brief. See http://www.carnegieknowledgenetwork.org/briefs/value-added/interpreting-value-added/

test scores treated as a continuous variable, which suggests the possibility of using categorical performance standards as reasonable approximation of teachers' outcomes resulting from working with students.

Rank correlations between the EPL estimates and the VAM estimates were fairly high for all the cases (see Table 4-1 for mathematics and Table 4-2 for reading). Since some previous studies have shown that gain scores are not reliable measures of growth (Bonate, 2000), and that using gain scores for computing VAM is not consistent (Papay, 2011), VAM estimates using gain scores (VAM-GA) were not seriously considered for the comparison. Rank correlations between the VAM estimates and the EPL estimates using the dichotomous performance categories ranged from .81to .86 in mathematics, and ranged from .59 to .79 in reading. For the EPL estimates using the four performance categories, rank correlations with the VAM estimates were at least .88 in mathematics and .82 in reading. In addition, no noticeable outliers were found according to the scatter plots in all grades (see Figure 4-1 and Figure 4-2 for example in Grade 5). Again notice that variances of the EPL estimates when using dichotomous categories in reading were smaller than the VAM estimates. The EPERF-based method using polytomous outcomes and the EPROF-based method appeared to produce consistent rankings of teachers both in mathematics and reading.

Ranking of teachers was almost perfectly correlated between the EPL-D1PL and EPL-D2PL, and between the EPL-P1PL and EPL-P2PL. Differences in model fit between the 1PL and 2PL models were also small in consideration of the large number of students in each model (see Table 5-1 in section 5.1), although they were statistically significant. As this marginal random effect of the slope implies that individual teachers were not very different from each other in

Table 4-1. Descriptive statistics and rank correlations of teachers' EPL and VAM estimates in mathematics

Elementary	Grade 4 (N=1,750)				Grade 5 (N=1,758)				Rank correlations (Grade 4 on the lower diagonal; Grade 5 on the upper diagonal)						
	M	SD	Min	Max	M	SD	Min	Max	1.	2.	3.	4.	5.	6.	7.
1. EPL-D1PL	-.30	.23	-1.34	.92	-.37	.27	-1.68	.49		.99	.90	.91	.85	.84	.63
2. EPL-D2PL	-.30	.23	-1.61	1.21	-.37	.28	-2.08	.44	.99		.92	.85	.85	.84	.64
3. EPL-P1PL	.00	.23	-1.12	1.15	.00	.28	-1.52	1.00	.88	.88		.99	.92	.91	.68
4. EPL-P2PL	.00	.22	-1.06	.96	.00	.26	-1.22	1.05	.89	.89	.99		.92	.91	.68
5. VAM-RE	.00	.20	-1.12	1.03	.00	.20	-1.01	.91	.83	.82	.92	.92		.99	.73
6. VAM-AR	.00	.26	-1.41	1.23	.00	.24	-1.12	1.11	.82	.82	.92	.92	.99		.76
7. VAM-GA	.00	.26	-1.58	1.26	-.01	.27	-1.62	1.08	.81	.81	.90	.90	.97	.98	
N. of students	31.6	16.7	11	115	40.4	22.9	11	156							
Secondary	**Grade 6 (N=1,731)**				**Grade 7 (N=1,698)**				**Rank correlations (Grade 6 on the lower diagonal; Grade 7 on the upper diagonal)**						
	M	SD	Min	Max	M	SD	Min	Max	1.	2.	3.	4.	5.	6.	7.
1. EPL-D1PL	-.30	.23	-1.34	.45	-.62	.23	-1.75	.53		.99	.87	.89	.83	.82	.66
2. EPL-D2PL	-.29	.23	-1.43	.45	-.62	.24	-2.04	.51	.99		.86	.89	.83	.81	.66
3. EPL-P1PL	.00	.25	-1.18	1.01	.00	.24	-1.44	1.20	.90	.89		.99	.92	.90	.70
4. EPL-P2PL	.00	.24	-1.06	.82	.00	.23	-1.19	1.16	.91	.91	.99		.91	.89	.69
5. VAM-RE	.00	.21	-.93	.99	.00	.18	-.86	.97	.86	.85	.92	.92		.99	.77
6. VAM-AR	.00	.23	-1.11	1.13	.00	.20	-.90	1.14	.83	.83	.89	.88	.99		.81
7. VAM-GA	.00	.23	-1.33	.91	.00	.21	-1.16	1.12	.80	.80	.85	.85	.92	.94	
N. of students	82.2	42.4	11	282	90.3	41.9	11	224							

1. EPL-D1PL: EPL using dichotomous outcome with random intercept;
2. EPL-D2PL: EPL using dichotomous outcome with random intercept and random slope;
3. EPL-P1PL; EPL using polytomous outcome with random intercept;
4. EPL-P2PL: EPL using polytomous outcome with random intercept and random slope;
5. VAM-RE: VAM with random teacher effects;
6. VAM-AR: VAM with average residual by teacher;
7. VAM-GA: VAM with gain scores

Table 4-2. Descriptive statistics and rank correlations of teachers' EPL and VAM estimates in reading

Elementary	Grade 4 (N=1,940)				Grade 5 (N=2,039)				Rank correlations (Grade 4 on the lower diagonal; Grade 5 on the upper diagonal)						
	M	SD	Min	Max	M	SD	Min	Max	1.	2.	3.	4.	5.	6.	7.
1. EPL-D1PL	.67	.05	.39	.88	.57	.06	.13	.75		.88	.81	.80	.70	.68	.57
2. EPL-D2PL	.66	.08	.43	1.3	.56	.07	-.34	.80	.90		.70	.68	.60	.59	.50
3. EPL-P1PL	.00	.15	-.68	.81	.00	.15	-1.07	.42	.82	.72		.99	.84	.82	.70
4. EPL-P2PL	.00	.14	-.65	.79	.00	.11	-1.02	.39	.84	.73	.99		.84	.82	.70
5. VAM-RE	.00	.14	-.64	.84	.00	.11	-079	.37	.70	.60	.88	.88		.99	.80
6. VAM-AR	.00	.24	-1.07	1.12	.00	.21	-1.35	.96	.69	.60	.87	.87	.99		.83
7. VAM-GA	.00	.25	-1.38	1.40	.00	.23	-1.81	.81	.54	.54	.70	.69	.78	.81	
N. of students	36.0	37.6	11	336	39.9	28.3	11	214							
Secondary	Grade 6 (N=2,226)				Grade 7 (N=1,986)				Rank correlations (Grade 6 on the lower diagonal; Grade 7 on the upper diagonal)						
	M	SD	Min	Max	M	SD	Min	Max	1.	2.	3.	4.	5.	6.	7.
1. EPL-D1PL	.33	.06	.06	.55	.38	.05	.18	.54		.98	.87	.88	.78	.77	.37
2. EPL-D2PL	.33	.07	.07	.68	.38	.04	.17	.53	.99		.84	.85	.75	.73	.36
3. EPL-P1PL	.00	.20	-.91	.70	.00	.14	-.60	.62	.88	.87		.99	.91	.88	.41
4. EPL-P2PL	.00	.20	-.89	.66	.00	.14	-.59	.59	.89	.88	.99		.90	.88	.41
5. VAM-RE	-.02	.18	-.79	.75	.00	.11	-.42	.55	.79	.79	.91	.92		.98	.46
6. VAM-AR	-.01	.25	-1.19	.85	-.01	.25	-1.19	.85	.77	.75	.89	.89	.99		.50
7. VAM-GA	-.01	.24	-1.54	.90	.02	.21	-.96	.71	.64	.63	.76	.76	.85	.88	
N. of students	79.5	46.4	11	283	93.8	46.7	11	267							

1. EPL-D1PL: EPL using dichotomous outcome with random intercept;
2. EPL-D2PL: EPL using dichotomous outcome with random intercept and random slope;
3. EPL-P1PL; EPL using polytomous outcome with random intercept;
4. EPL-P2PL: EPL using polytomous outcome with random intercept and random slope;
5. VAM-RE: VAM with random teacher effects;
6. VAM-AR: VAM with average residual by teacher;
7. VAM-GA: VAM with gain scores

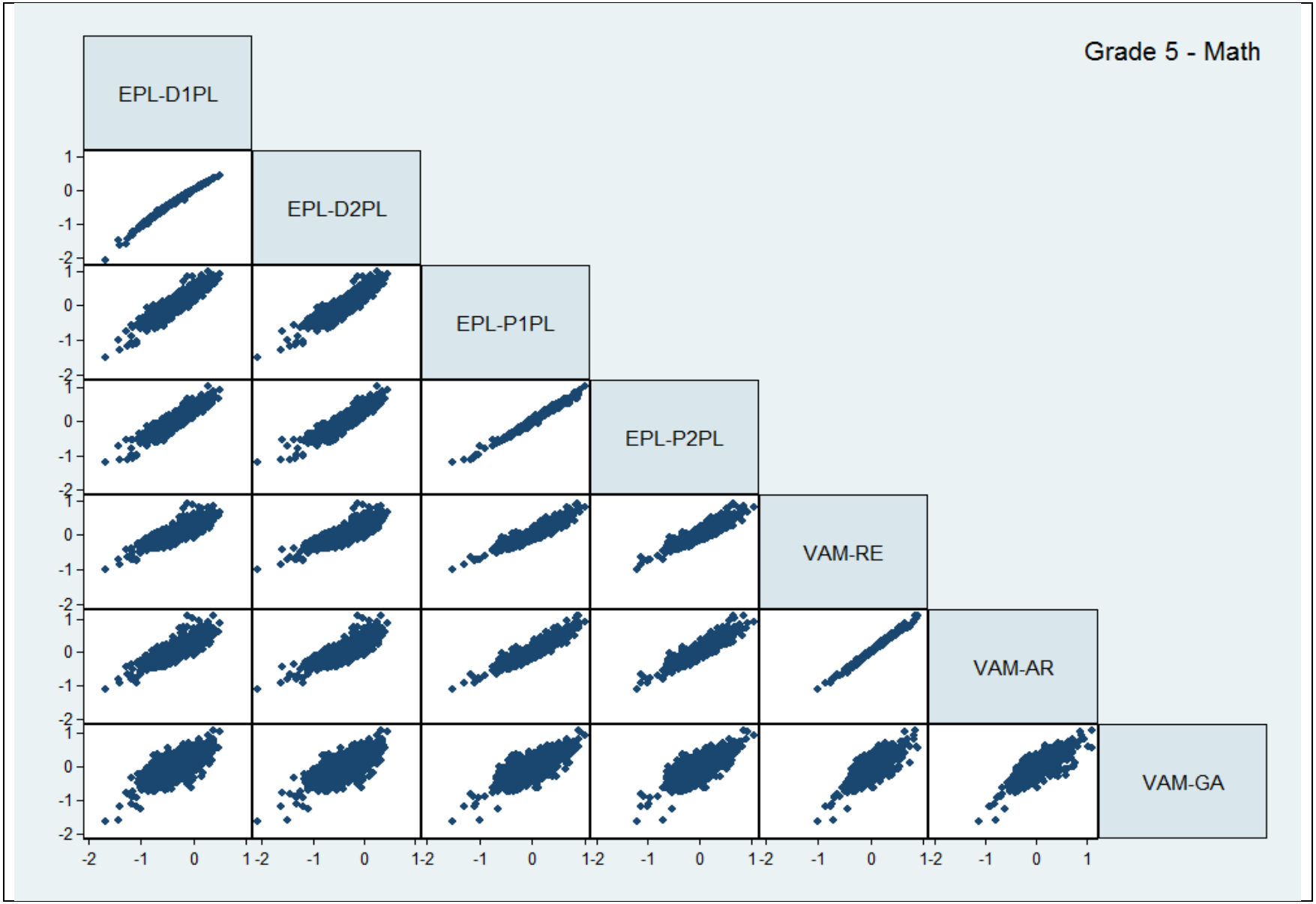


Figure 4-1. Scatter plots of the EPL and VAM estimates (Grade 5, Mathematics)

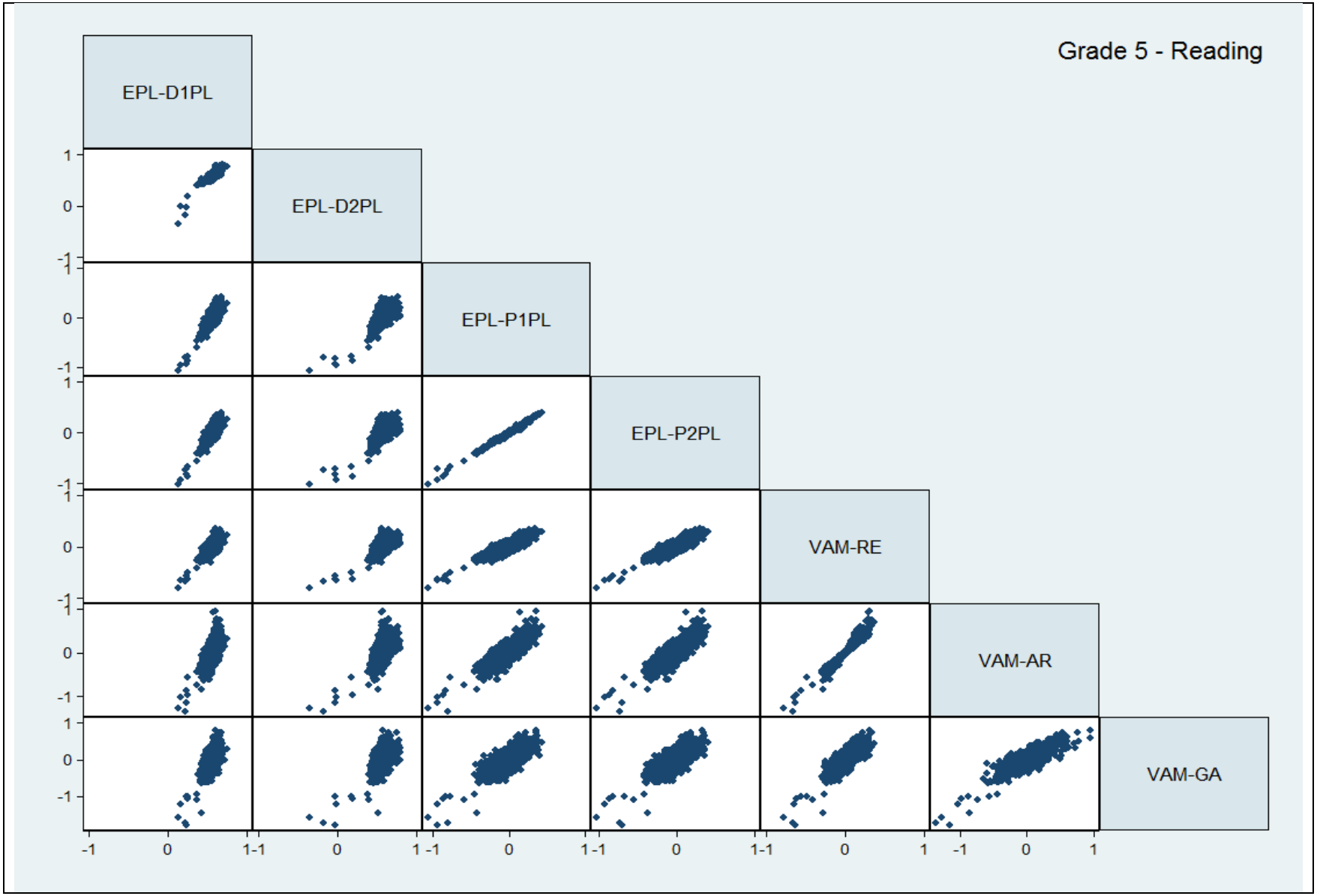


Figure 4-2. Scatter plots of the EPL and VAM estimates (Grade 5, Reading)

their slope estimates, more parsimonious EPERF-1PL models may be preferred. Model fit and relevant issues in model selection are discussed further in Chapter 5.

4.2. Relationship to Student and Teacher Characteristics

How the teacher effect estimates were associated with some key student and teacher background variables was observed focusing on mathematics. For the properties of quality teacher effectiveness measures, the estimates are expected not to be dependent on their students' characteristics, but are rather dependent on reasonable teacher quality indicators, which may provide criterion validity evidence. Because the teacher effect estimates are highly correlated with each other, as shown in the previous section, it was expected that their relationships to student or teacher characteristics were not considerably different from each other.

4.2.1. Relationship to Student Characteristics

For the relationship to student characteristics, several student background variables were averaged across students by teacher, and the aggregated data of the teacher-level were used for relevant analyses. Bivariate correlations between each teacher effect estimate and their average student backgrounds are shown by grade-level in Table 4-3. Overall, the correlations tended to be higher in Grades 6 and 7 (secondary schools in the bottom rows) than in Grades 4 and 5 (elementary schools in the top rows), and to be lower in the EPERF-based teacher effects using dichotomous outcomes (EPL-D1PL and EPL-D2PL) than in those using polytomous outcomes (EPL-P1PL and EPL-P2PL), or than in the EPROF-based teacher effects (VAM-RE and VAM-AR).

First, the teacher effect estimates were positively associated with the average prior and outcome achievement of students, as well as the proportion of students classified at the proficient level, and the degree of associations was slightly weaker in the EPL than in the VAM. For example, while the correlations with prior achievement ranged from .11 to .37 for the VAM, those fell to between .04 and .36 for the EPL. The correlations with the outcome achievement in both EPL and VAM were significantly higher, ranging from .56 and .69 for the VAM, and from .43 to .63 for the EPL.

Second, the teacher effect estimates were negatively associated with the average challenge index (CI) of students, which implies that the higher teachers' effect estimates were, the lower the average CI of their students were. The correlations to the average CI of students were -.05 to -.37 in the EPLs, which tended to be weaker than those in the VAM, which ranged from -.12 to -.38. The proportion of the economically-disadvantaged showed similar patterns. Note that the economically-disadvantaged group indicator was one of the CI indicators. Last, associations with the number of students and with the proportion of female students were small.

In order to investigate whether the composition of students who worked with each teacher can predict their teachers' estimated effects, the teacher effect estimates were regressed by a set of student background variables. Those variables were aggregated at the teacher-level, so that each variable represented classroom composition rather than individual student characteristics. The results are displayed in Tables 4-4 for elementary schools and 4-5 for secondary schools. Because the results were very similar between EPL-D1PL and EPL- D2PL, between EPL-P1PL and EPL-P2PL, and between VAM-RE and VAM-AR, only those of EPL-D1PL, EPL-P1PL, and VAM-RE are displayed.

Results show that it is not easy to find any consistent or common pattern of the coefficients across different grades and estimates. A distinct observation is that the proportion of economically disadvantaged students was a strong and negative predictor of all three teacher effect estimates at every grade-level, after controlling for other student background variables. In contrast, the proportion of limited English proficient students was positively associated with the teacher effect estimates, except Grade 7. This may be because Asian students who show high performance in mathematics are likely to be classified as LEP students at the beginning of their immigration. The effects of the proportion of Asian students on the teacher effect estimates were significant in Grade 4 and Grade 7. Prior test score was a significant positive predictor of the VAM-RE estimate in Grade5-7, and of the EPL-D1PL estimates in Grade 6 and Grade 7. Prior test scores were associated weakly with the EPL-D1PLestimates, either positively or negatively.

The *t*-values corresponding to the student background variables, such as free/reduced lunch, targeted assistant program, and limited English proficiency, which were listed as student CI indicators, tended to be smaller in the EPL estimates than in the VAM estimates. Looking at the R-squares, the cases of the EPL-D1PL tended to have lower R-square values than the EPL-P1PL or the VAM. While the set of student background variables explained the 7-19% variance of the VAM estimates and the 5-21% of the EPL-P1PL estimates, those accounted for only half of the variance in the EPL-D1PL. With respect to the school levels, student characteristics explained better the secondary school teachers' effect estimates by more than twice those of the elementary school teachers'. More dynamic relationships between each teacher effect estimate and student background variables needs to be examined further.

Table 4-3. Correlations between teacher effect estimates and student background variables in mathematics

Student background variables	EPL				VAM	
	D1PL	D2PL	P1PL	P2PL	RE	AR
			Elementary (Grade 4, Grade 5)			
Average prior test-scores	(.08, .16)	(.04, .14)	(.12, .21)	(.12, .20)	(.14, .23)	(.11, .17)
Average outcome test-scores	(.46, .54)	(.43, .53)	(.54, .62)	(.54, .61)	(.58, .67)	(.56, .62)
Proportion of proficient students	(.59, .66)	(.56, .64)	(.57, .65)	(.57, .64)	(.56, .63)	(.53, .59)
Average CI of students	(-.08, -.17)	(-.05, -.15)	(-.13, -.22)	(-.13, -.20)	(-.15, -.24)	(-.12, -.18)
Proportion of female	(-.03, .05)	(-.02, .05)	(-.03, .04)	(-.03, .04)	(-.02, .03)	(-.02, .03)
Proportion of economically-disadvantaged group	(.-15, -.20)	(-.13, -.19)	(-.20, -.26)	(-.19, -.24)	(-.20, -.25)	(-.17, -.18)
N. of students	(-.01, .01)	(-.01, -.02)	(-.01, .02)	(-.02, .03)	(-.03, .00)	(-.03, -.01)
			Secondary (Grade 6, Grade 7)			
Average prior test-scores	(.23, .25)	(.22, .23)	(.36, .34)	(.35, .33)	(.37, .34)	(.29, .26)
Average outcome test-scores	(.51, .51)	(.50, .48)	(.64, .62)	(.63, .61)	(.67, .64)	(.60, .57)
Proportion of proficient students	(.41, .59)	(.40, .57)	(.66, .62)	(.66, .62)	(.38, .63)	(.58, 57)
Average CI of students	(-.24, -.26)	(-.23, -.23)	(-.37, -.34)	(-.36, -.33)	(-.38, -.34)	(-.29, -.26)
Proportion of female	(-.01, -.06)	(-.01, -.06)	(-.01, -.06)	(-.00, -.06)	(-.04, -.04)	(-.05, -.05)
Proportion of economically disadvantaged group	(-.17, -.23)	(-.16, -.21)	(-.41, -.37)	(-.40, -.37)	(-.22, -.35)	(-.25, -.25)
N. of students	(.01, .02)	(.01, .01)	(.03, .05)	(.04, .06)	(.00, 01)	(-.02, .00)

Table 4-4. Regression of teacher effect estimates on the average student background variables in mathematics (Elementary)

Grade 4	EPL-D1PL		EPL-P1PL		VAM-RE	
	Beta (se)	t	Beta (se)	t	Beta (se)	t
Economically-disadvantaged	**-.16 (.04)**	**-4.20**	**-.16 (.04)**	**-4.10**	**-.14 (.03)**	**-4.35**
Free/reduced lunch	-.02 (.02)	-.71	-.02 (.02)	-1.03	**-.05 (.02)**	**-2.51**
Targeted assistant program	.04 (.03)	1.58	**.11 (.03)**	**4.19**	**.08 (.02)**	**3.51**
Special education	-.07 (.16)	-.46	.01 (.16)	.03	-.06 (.14)	-.41
Limited English proficiency	**.15 (.06)**	**2.53**	**.13 (.06)**	**2.14**	**.16 (.05)**	**2.98**
Disability	-.04 (.15)	-.24	-.16 (.15)	-1.09	-.08 (.13)	-.62
Prior test score	-.02 (.02)	-1.32	-.01 (.02)	-.91	.00 (.01)	.20
Proportion of female	-.07 (.06)	-1.10	.07 (.06)	-1.22	-.03 (.05)	-.63
Proportion of Asian	**.22 (.11)**	**1.98**	**.38 (.10)**	**3.50**	**.30 (.10)**	**3.12**
Proportion of Black	.03(.02)	1.44	-.01 (.02)	-.39	**.07 (.02)**	**3.18**
Proportion of Hispanic	-.04 (.06)	-.58	-.08 (.06)	-1.31	-.06 (.05)	-1.19
R-square	.033		.062		.071	
Adjusted R-square	.027		.056		.065	
Grade 5	EPL-D1PL		EPL-P1PL		VAM-RE	
	Beta (se)	t	Beta (se)	t	Beta (se)	t
Economically-disadvantaged	**-.11 (.05)**	**-2.10**	**-.16 (.05)**	**-3.39**	**-.12 (.04)**	**-3.60**
Free/reduced lunch	**-.10 (.03)**	**-3.17**	**-.07 (.03)**	**-2.44**	**-.07 (.02)**	**-3.34**
Targeted assistant program	-.00 (.03)	-.00	-.00 (.03)	-.10	.04 (.02)	1.72
Special education	.23 (.21)	1.09	.18 (.21)	.87	.20 (.15)	1.34
Limited English proficiency	**.27 (.08)**	**3.18**	**.24 (.08)**	**2.92**	**.21 (.06)**	**3.57**
Disability	-.19 (.21)	-.91	-.25 (.21)	-1.21	-.17 (.15)	-1.11
Prior test score	.02 (.02)	1.47	.03 (.02)	1.72	**.04 (.01)**	**3.90**
Proportion of female	**.15 (.07)**	**2.14**	.11 (.07)	1.53	.05 (.05)	1.06
Proportion of Asian	-.04 (.14)	-.28	.03 (.14)	.24	-.05 (.10)	-0.48
Proportion of Black	-.05 (.03)	-1.81	**-.07 (.03)**	**-2.27**	**.04 (.02)**	**2.22**
Proportion of Hispanic	.02 (.07)	.30	-.04 (.07)	-.54	-.00 (.05)	-.01
R-square	.060		.083		.089	
Adjusted R-square	.054		.077		.083	

Table 4-5. Regression of teacher effect estimates on the average student background variables in mathematics (Secondary)

Grade 6	EPL-D1PL		EPL-P1PL		VAM-RE	
	Beta (se)	t	Beta (se)	t	Beta (se)	t
Economically-disadvantaged	**-.21 (.04)**	**-5.39**	**-.19 (.04)**	**-4.85**	**-.13 (.03)**	**-3.88**
Free/reduced lunch	-.03 (.02)	-1.27	-.02 (.02)	-.74	**-.07 (.02)**	**-3.98**
Targeted assistant program	.06 (.03)	1.82	**.08 (.03)**	**2.38**	**.10 (.03)**	**3.39**
Special education	-.24 (.22)	-1.13	-.30 (.23)	-1.32	-.25 (.19)	-1.30
Limited English proficiency	**.20 (.07)**	**2.73**	**.20 (.08)**	**2.54**	**.21 (.06)**	**3.19**
Disability	.14 (.21)	.68	.15 (.22)	.69	.19 (.19)	1.02
Prior test score	-.01 (.01)	-.51	**.05 (.01)**	**3.30**	**.07 (.01)**	**6.13**
Proportion of female	.01 (.06)	.30	.03 (.06)	.43	-.08 (.05)	-1.42
Proportion of Asian	.17 (.11)	1.54	.17 (.12)	1.50	**.21 (.10)**	**2.17**
Proportion of Black	**-.09(.02)**	**-3.82**	**-.15 (.02)**	**-6.00**	.01 (.02)	0.62
Proportion of Hispanic	**-.20 (.06)**	**-3.53**	**-.25 (.06)**	**-4.41**	**-.10 (.05)**	**-1.97**
R-square	.130		.216		.187	
Adjusted R-square	.125		.211		.181	
Grade 7	EPL-D1PL		EPL-P1PL		VAM-RE	
	Beta (se)	t	Beta (se)	t	Beta (se)	t
Economically-disadvantaged	**-.14 (.04)**	**-3.56**	**-.20 (.04)**	**-5.09**	**-.14 (.03)**	**-4.63**
Free/reduced lunch	-.04 (.02)	-1.56	-.02 (.02)	-.68	-.03 (.02)	-1.73
Targeted assistant program	.02 (.04)	.49	.02 (.04)	.65	**.08 (.03)**	**3.09**
Special education	-.22 (.21)	-.1.02	-.13 (.21)	-.65	-.04 (.15)	-.27
Limited English proficiency	.05 (.06)	.77	.05 (.06)	.83	.07 (.04)	1.56
Disability	.22 (.21)	1.07	.09 (.21)	.41	.09 (.15)	.58
Prior test score	.03 (.01)	1.90	**.03 (.01)**	**2.43**	**.06 (.01)**	**6.04**
Proportion of female	**-.12 (.06)**	**-2.06**	**-.15 (.06)**	**-2.57**	-.08 (.04)	-1.78
Proportion of Asian	**.57 (.12)**	**4.93**	**.71 (.11)**	**6.29**	**.54 (.08)**	**6.46**
Proportion of Black	-.04 (.03)	-1.55	**-.06 (.03)**	**-2.54**	**.07 (.02)**	**3.97**
Proportion of Hispanic	-.04 (.06)	-.78	-.03 (.05)	-.55	.03 (.04)	.79
R-square	.112		.181		.183	
Adjusted R-square	.106		.176		.177	

4.2.2. Relationship to Teacher Characteristics

Several available teacher background or qualification variables were used to predict the teacher effect estimates; the results are shown in Table 4-6 for elementary schools and Table 4-7 for secondary schools. As displayed, it appears that no variable predicted either the EPL or the VAM estimates with reliability and reasonableness. Looking at the R-squares, the set of variables explained less than 2% of variance in the estimates.

In mathematics in elementary schools, teachers possessing advanced degrees such as masters or doctorates appeared to achieve higher scores in their VAM and EPL than those who did not. This did not hold in secondary schools. Studies on teacher quality or teaching quality has demonstrated that while most teacher credentials or qualifications are not related to student achievement (e.g., Buddin, Zamarro, 2009; Kane et al, 2008), teaching experience is often considered to matter (e.g., Clotfelter, Ladd & Vigdor, 2007; Rockoff, 2004). Surprisingly, number of years of teaching associated negatively with the teacher effect estimates, which was consistent in elementary schools, except for the EPL-P1PL estimates for Grade 5 teachers. Only in the EPL-P1PL estimates for Grade 7 teachers did number of years of teaching predicted the teacher effect.

Table 4-6. Regression of the teacher effect estimates on the teacher background variables in mathematics (Elementary)

Grade 4 (N=1,234)	EPL-D1PL		EPL-P1PL		VAM-RE	
	Beta (se)	t	Beta (se)	t	Beta (se)	t
Female	-.010 (.018)	-0.56	-.008 (.017)	-.91	-.005 (.015)	-.30
Advanced degree	**.032 (.014)**	**2.29**	.008 (.017)	1.73	**.019 (.015)**	**1.28**
Hrs. professional development	.000 (.000)	.044	.000 (.000)	.54	.000 (.000)	-.64
Credentials						
Professional	-.121 (.086)	-1.41	-.107 (.091)	-1.17	-.116 (.078)	-1.48
Provisional	-.126 (.086)	-1.47	-.073 (.087)	-1.07	-.115 (.078)	-1.47
Years of teaching	**-.003 (.001)**	**-3.41**	**-.000 (.000)**	**-.31**	**-.001 (.001)**	**-1.23**
Adjusted R-square	.003		.003		.005	
Adjusted R-square	.008		.000		.012	
Grade 5 (N=1,308)	EPL-D1PL		EPL-P1PL		VAM-RE	
	Beta (se)	t	Beta (se)	t	Beta (se)	t
Female	.009 (.017)	0.52	.002 (.017)	.12	.006 (.013)	.41
Advanced degree	**.047 (.016)**	**2.91**	**.016 (.016)**	**2.48**	**.037 (.012)**	**3.13**
Hrs. professional development	-.000 (.000)	-.66	-.000 (.000)	-.19	.000 (.000)	-.11
Credentials						
Professional	.004 (.079)	.06	-.026 (.079)	-.33	-.000 (.057)	-.02
Provisional	.006 (.079)	.08	-.019 (.079)	-.24	-.014 (.057)	-.25
Years of teaching	**-.002 (.001)**	**-2.52**	-.002 (.001)	1.84	**-.002 (.001)**	**-2.66**
R-square	.011		.007		.010	
Adjusted R-square	.006		.003		.006	

Table 4-7. Regression of the teacher effect estimates on the teacher background variables in mathematics (Secondary)

Grade 6 (N=1,156)	EPL-D1PL		EPL-P1PL		VAM-RE	
	Beta (se)	t	Beta (se)	t	Beta (se)	t
Female	-.005 (.016)	-.034	-.001 (.018)	-.04	.003 (.016)	.22
Advanced degree	.003 (.016)	.17	.017 (.018)	.97	.025 (.015)	1.64
Hrs. professional development	-.000 (.000)	-.71	.000 (.000)	-1.31	-.000 (.000)	-1.39
Credentials						
Professional	-.100 (.085)	-1.18	-.163 (.128)	-1.28	-.112 (.108)	-1.04
Provisional	-.175 (.084)	-1.10	-.165 (.128)	-1.29	-.122 (.108)	-1.13
Years of teaching	-.001 (.001)	-.55	-.001 (.001)	-1.23	-.002 (.001)	-1.77
R-square	.006		.006		.008	
Adjusted R-square	.000		.000		.002	
Grade 7 (N=1,034)	EPL-D1PL		EPL-P1PL		VAM-RE	
	Beta (se)	t	Beta (se)	t	Beta (se)	t
Female	-.006 (.018)	-0.34	.001 (.018)	.04	-.001 (.014)	-.09
Advanced degree	.023 (.019)	1.22	.031 (.019)	1.61	.026 (.014)	1.81
Hrs. professional development	-.000 (.000)	-.01	-.000 (.000)	-.37	.000 (.000)	-.68
Credentials						
Professional	.092 (.111)	.81	.160 (.111)	1.42	.151 (.085)	1.79
Provisional	.066 (.111)	.58	.130 (.111)	1.15	.120 (.086)	1.39
Years of teaching	.001 (.001)	1.04	.001 (.001)	1.16	.001 (.001)	.91
R-square	.008		.013		.016	
Adjusted R-square	.002		.006		.010	

4.3. Consistency of the Teacher Effect Estimates

Intra-person (teacher) rank correlations were monitored within each model to evaluate the extent to which individual teachers' estimates were consistent across different subjects – mathematics and reading – and grades. The intra-teacher rank correlation between mathematics and reading or between different grade-levels within the same subject were expected to be moderate or high rather than very low, under the assumption that teacher capability of teaching is stable so that legitimate teacher effect estimates do not change dramatically depending on context.

Consistency of the estimated teacher effects between mathematics and reading is shown in Table 4-8. Rank correlations of the teacher effect estimates between mathematics and reading ranged from .19 to .31 across grades. The VAM-RE and VAM-AR tended to be more consistent than the EPL estimates, to a marginal degree. For Grade 6, the EPL-P1PL estimates were more stable across subjects: the correlations were above .3 except VAM-GA. The number of teachers who taught both reading and mathematics decreased as grade-level increased. In Grade 7, only 170 teachers taught both subjects, so that the correlations of the teacher effect estimates between mathematics and reading may be of questionable.

Consistency of the teacher effect estimates between different grade-levels within each subject was examined also. As Table 4-9 displays, correlations of the estimates across different grade-levels within the same subject, in general, were smaller than those between different subjects within each grade-level. It needs to be noted, however, that the number of teachers teaching multiple grade-levels was small. Rank correlations of the teacher effect estimates

between Grades 5 and 6 in the same subject were highest, above .15. In particular, consistency of the EPL-P1PL and EPL-P2PL appeared better.

Table 4-8. Intra-teacher rank correlations between mathematics and reading

	Across Grades (N=4,157)	Grade 4 (N=1,647)	Grade 5 (N=1,506)	Grade 6 (N=657)	Grade 7 (N=174)
EPL-D1PL	.209*	.187*	.215*	.307*	.017
EPL-D2PL	.191*	.152*	.168*	.303*	-.002
EPL-P1PL	.273*	.263*	.253*	.401*	.097
EPL-P2PL	.268*	.261*	.243*	.398*	.075
VAM-RE	.312*	.313*	.285*	.394*	.089
VAM-AR	.290*	.293*	.271*	.351*	.085
VAM-GA	.194*	.216*	.181*	.258*	-.056

Table 4-9. Intra-teacher rank correlations between different grades

	Mathematics			Reading		
	G4-G5 (N=60)	G5-G6 (N=66)	G6-G7 (N=354)	G4-G5 (N=91)	G5-G6 (N=72)	G6-G7 (N=363)
EPL-D1PL	.110	. 172	-.073	-.020	-.044	.165*
EPL-D2PL	.126	. 152	-.088	.015	-.007	.162*
EPL-P1PL	.110	.278*	.044	.070	.140	.197*
EPL-P2PL	.114	.249*	.042	.104	.129	.186*
VAM-RE	.184	.210	.006	.096	.173	.221*
VAM-AR	.154	.206	-.044	.120	.175	.198*
VAM-GA	.036	-.238	-.188*	-.250*	-.048	-.123*

4.4. Additional Information of Teachers' Performance that the EPERF Produces

The EPERF-based method offers useful information on either the average or the individual teacher's performance estimates as well as his/her student characteristics. This section

articulates what the fitted EPERF models tell us about the average or the individual teacher's performance working with students assigned to a certain challenge index.

In the dichotomous EPERF, as shown in Figures 3-9 to 3-11, as teacher EPL increases, so too does the probability of student success in reaching the proficient level. As the lines of expected probability of success delineate, the 0 EPL is fixed at the point corresponding to .5 probability of student success when student CI is equal to 0. The observed probability of success depends on students' CI: the higher the average student challenge level, the higher the probability of success.

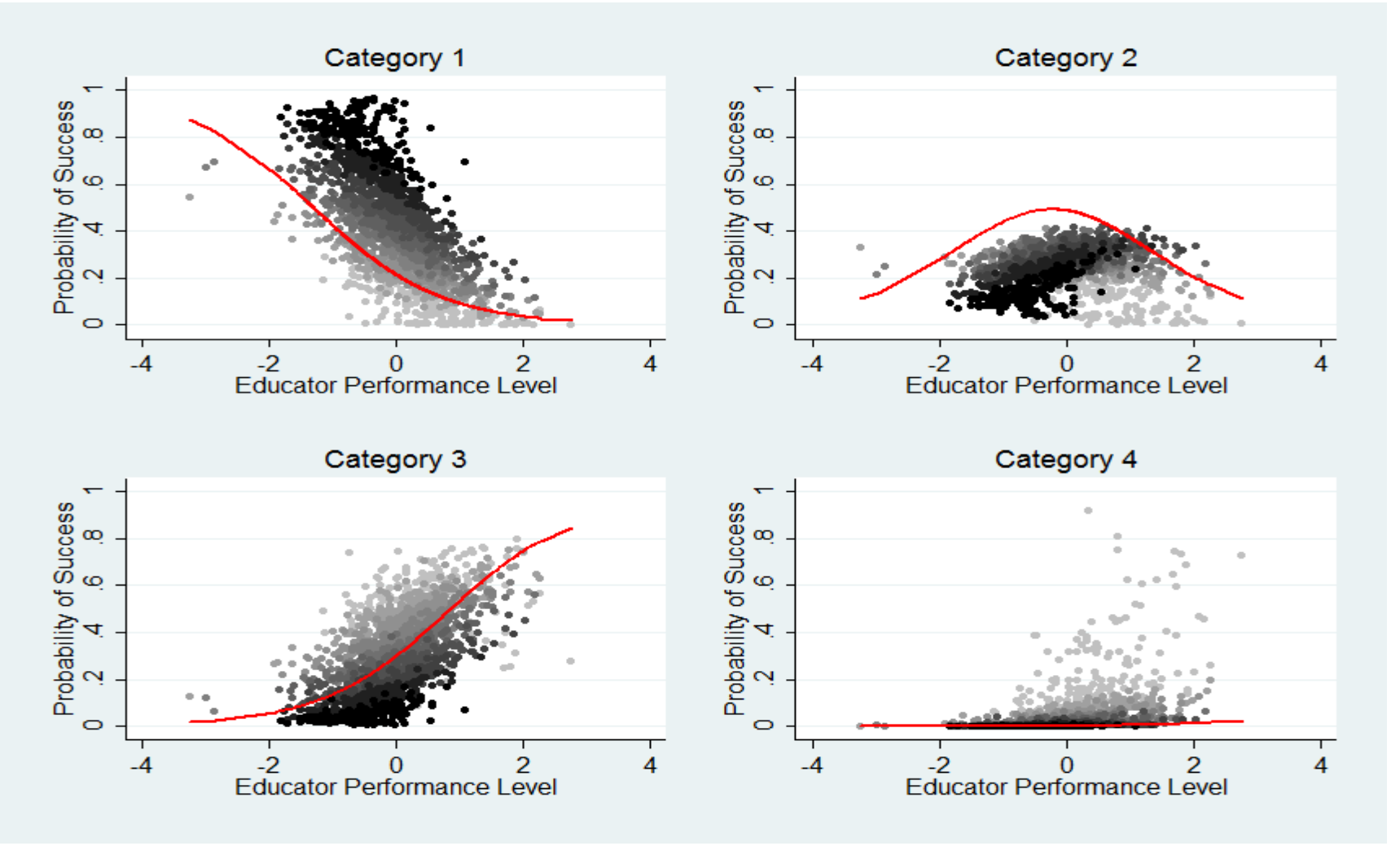
Two noticeable differences between mathematics (Figure 3-9) and reading (Figure 3-11) teacher EPL were found: (1) the mean of EPL is lower in mathematics than in reading; and (2) the variance in EPL is considerably larger in mathematics than in reading. At first glance, it appears on average to be harder for teachers to help students attain the proficient level in mathematics than in reading. On the other hand, supposing students are on the same levels of the challenge index, the probability of their success in mathematics is lower than in reading. Larger variance in teacher effects in mathematics than in reading has been commonly observed in previous studies of value-added measures (Condie, et al., 2011; Nye & Konstantopoulos, 2004; Rockoff, 2004).

In the EPERF use of polytomous performance categories, as shown in Figure 3-12, the 0 EPL is fixed at the point where teachers' expected scores of their students is equal to 2, corresponding to the second of the four performance levels, when student CI is equal to 0. According to the fitted functions, as the estimated teacher EPL increases, their students' average expected scores also increase up to 3, which indicates the "proficient level," not 4, the "advanced level."

Figure 4-3 represents the relationship between the estimated teacher EPL and their students' average probabilities in each performance category. As teacher EPL elevates, their students' average probability of achieving only the basic level (Category 1) tends to be lower. Students with higher CI, represented as dark spots, have higher probability of reaching this level. On the contrary, the higher the teacher's EPL, the higher his/her students' average probability of achieving the proficient level (Category 3), and students with lower CI show a higher probability of getting this level. The case of partially proficient level (Category 2) shows the uni-modal function with less than 0.6 probability across all EPL. Again, the probability that a teacher helps his/her students achieve the highest performance level, the advanced level (Category 4), does not noticeably change depending on the estimated teacher EPL. That is, teacher EPL appears to have no significant influence on the probability of student success in achieving the advanced level. Rather, that success depends on the average students' challenge levels. Also noteworthy is that only a small number of students achieved this level, and the proportion of students in this level was less than 10% in every grade.

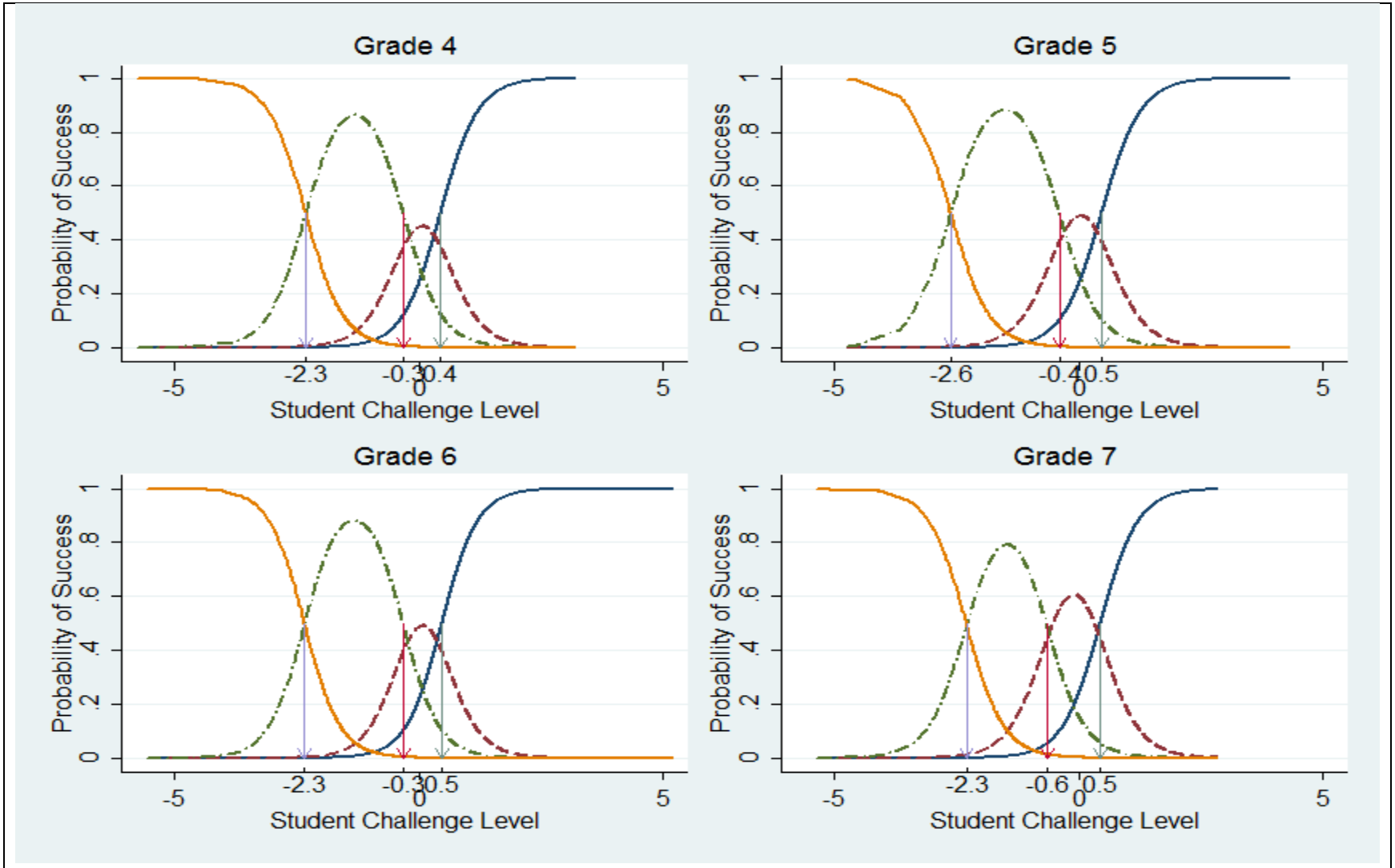
Table 4-10. Slope and threshold parameters (and standard errors) from the two polytomous EPERF-based models

	EPERF-P1PL				EPERF-P2PL			
	G4	G5	G6	G7	G4	G5	G6	G7
Mathematics								
Slope	-2.49 (.02)	-2.59 (.02)	-2.74 (.01)	-2.59 (.01)	-2.53 (.02)	-2.62 (.02)	-2.78 (.02)	-2.60 (.02)
Threshold 1 (B-PP)	-1.19 (.02)	-1.16 (.02)	-1.30 (.02)	-1.17 (.02)	-1.21 (.02)	-1.17 (.02)	-1.32 (.02)	-1.17 (.02)
Threshold 2 (PP-P)	.96 (.02)	.79 (.02)	.85 (.02)	1.64 (.02)	.94 (.02)	.78 (.02)	.84 (.02)	1.65 (.02)
Threshold 3 (P-M)	6.52 (.05)	5.99 (.05)	6.30 (.04)	5.94 (.03)	6.73 (.05)	6.06 (.05)	6.46 (.04)	6.09 (.03)
Reading								
Slope	-1.99 (.01)	-2.13 (.01)	-2.20 (.01)	-1.99 (.01)	-2.01 (.02)	-2.16 (.02)	-2.21 (.01)	-2.01 (.01)
Threshold 1 (B-PP)	-3.39 (.02)	-3.04 (.02)	-2.95 (.02)	-3.43 (.02)	-3.43 (.02)	-3.06 (.02)	-2.97 (.02)	-3.45 (.02)
Threshold 2 (PP-P)	-1.38 (.02)	-1.26 (.02)	-.73 (.01)	-.77 (.01)	-1.39 (.02)	-1.27 (.02)	-.73 (.01)	-.78 (.01)
Threshold 3 (P-M)	3.07 (.02)	2.21 (.02)	2.99 (.02)	3.33 (.02)	3.07 (.02)	2.21 (.02)	3.00 (.02)	3.32 (.02)



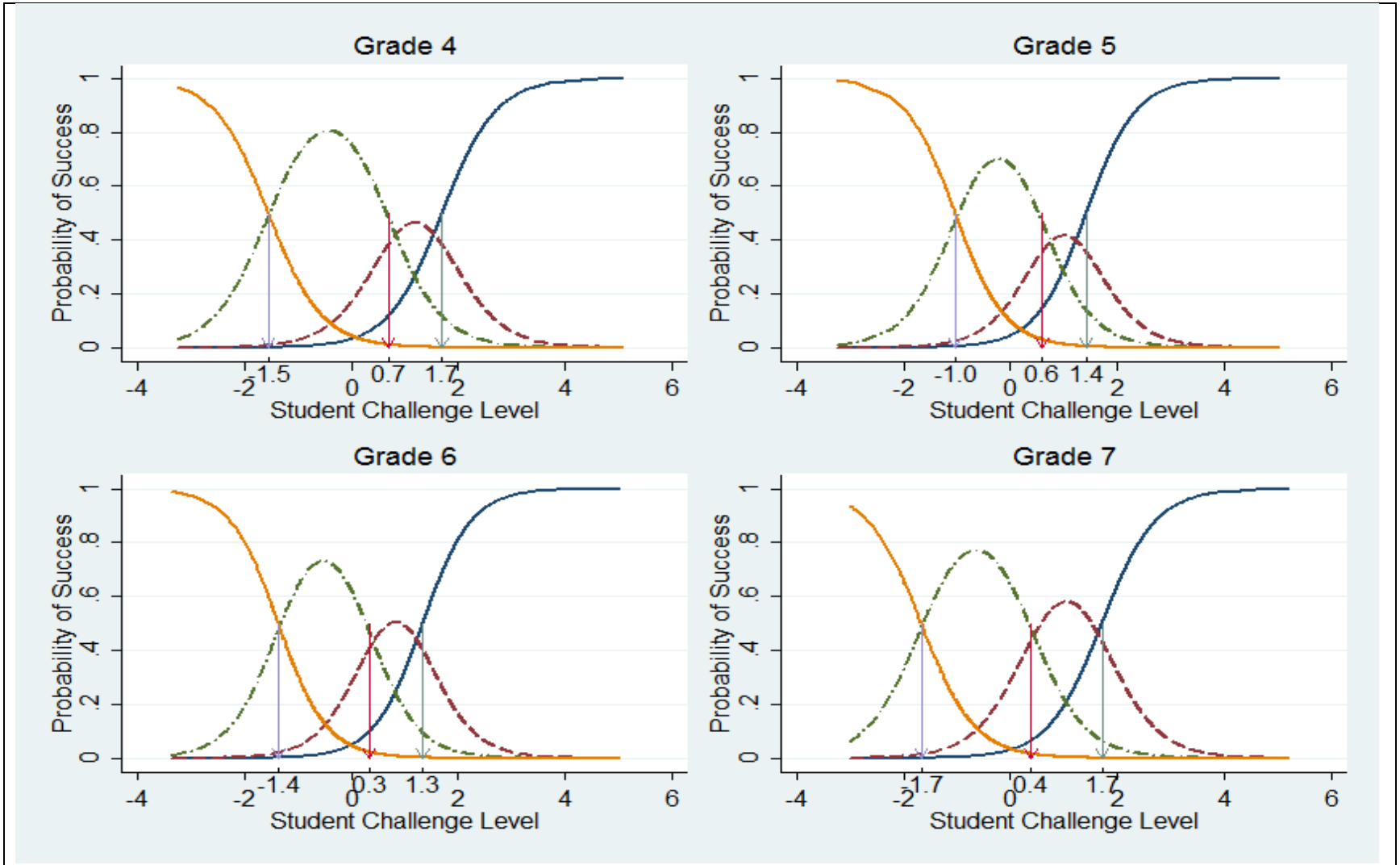
- Each mark represents individual teacher
- Brightness of the marks represents average students' challenge index. The more saturated the gray, the higher the average challenge index.

Figure 4-3. Teacher characteristic curves of the EPERF-P1PL by category (Grade 6, Mathematics)



- From left to right, the orange line represents the probability of attaining the advanced level; the green dotted line the probability of attaining the proficient level; the red dotted line the probability of attaining the partially proficient level; the blue bold line the probability of attaining the basic level.

Figure 4-4. Category characteristic curves from EPERF-P1PL by grade (Mathematics)



- From left to right, the orange line represents the probability of attaining the advanced level; the green dotted line the probability of attaining the proficient level; the red dotted line the probability of attaining the partially proficient level; the blue bold line the probability of attaining the basic level.

Figure 4-5. Category characteristic curves from EPERF-P1PL by grades (Reading)

Table 4-9 displays estimated slopes and thresholds from the EPERF-P1PL and the EPERF-P2PL. Each threshold can be converted into the scale of student CI using the calculation τ_k/γ_{10}, and those thresholds are marked in Figures 4-4 (mathematics) and 4-5 (reading). As the EPERF-D1PL demonstrated, the threshold in each step estimated from the EPERF-P1PL was lower in mathematics than in reading. In mathematics, the second threshold which is the borderline between the proficient and partially-proficient levels, tended to be lower in Grade 7 (-.6) than in other grades (-.3 or -.4); when student CI is fixed, the probability of achieving proficient level is lower in Grade 7 than in other grades. Results of EPERF-P1PL and those of EPER-P2PL are very similar. In reading, locations of the thresholds were more fairly varied across grades than in mathematics.

Noteworthy is that the thresholds are also on the scale of student CI, which means that the target performance level for students is determined by student CI. In other words, which performance level the teachers are expected to achieve with their students in the fitted model depends on individual students' challenge levels; the higher a student's CI, the lower his/her expected performance level (expected scores). Note that, on the contrary, in the VAM, any target test score or gain score is not particularly specified, and higher scores may be always advantageous for teachers. This distinctive characteristic of the polytomous EPERF-based models brings an important policy advantage over using the EPROF-based VAM.

Individual teachers' performance also can be graphed along with their students' challenge index distributions as Figure 4-6 displays. Examples for the three different teachers are presented: from left to right, each column shows the below average, around average, and above average EPL teacher respectively. The first row shows their EPL-D1PL; and the second row shows their EPL-P1PL. The small dots represent the students (actual) observed performance levels via their

CI: for the dichotomous, proficient (1) or non-proficient (0) level; for the polytomous outcome, advanced (4), proficient (3), partially proficient (2), or basic (1) level. Triangles represent the students' expected probabilities of being proficient or expected performance levels given teachers' EPL estimates.

4.5. Summary

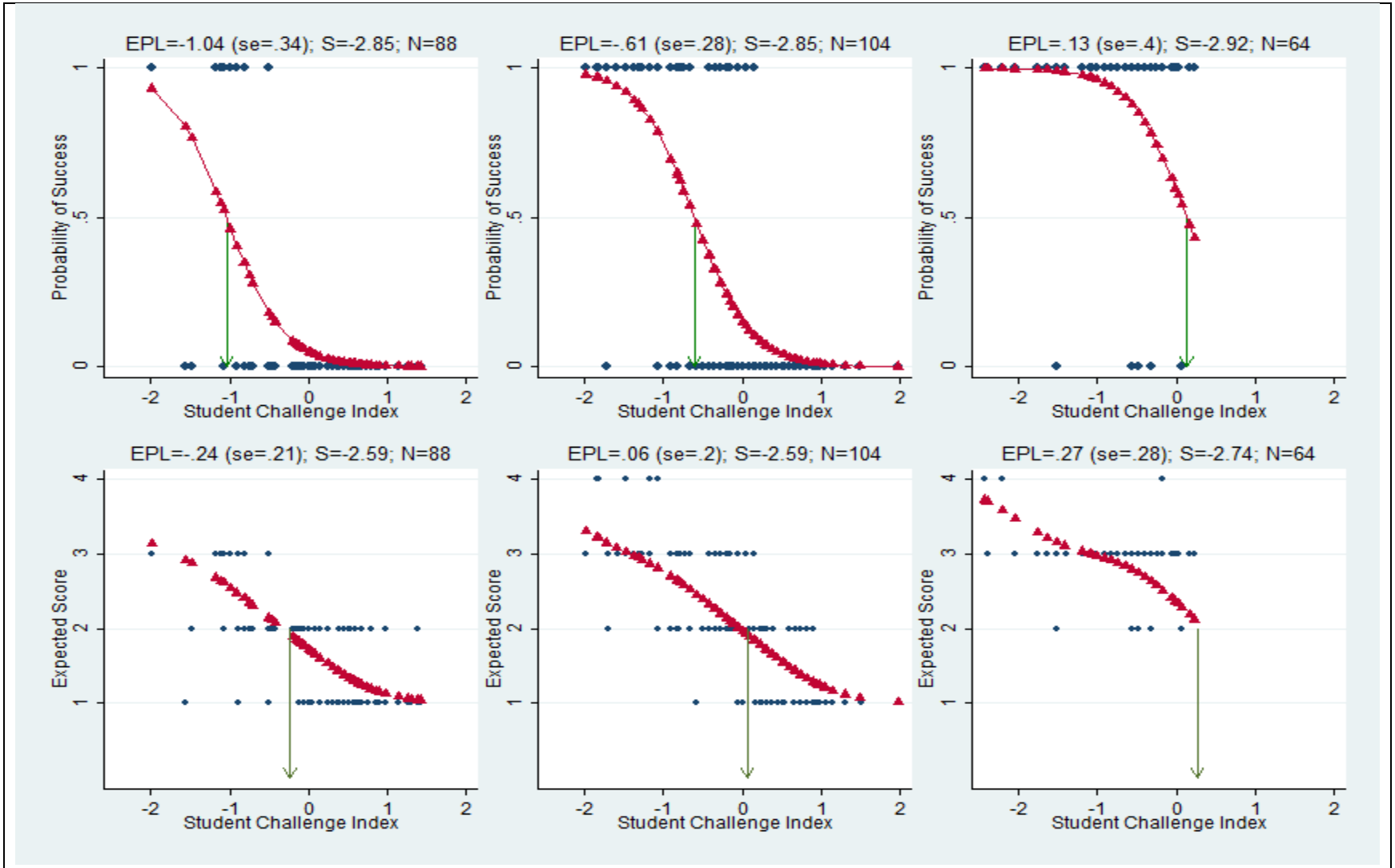
First, rank correlations between individual teachers' EPL and VAM ranged from .63 to .92 for mathematics, and from .37 to .91 for reading. Excepting the gain score-based VAM, the correlations were fairly high, showing above .81 for mathematics, but were moderately high, showing above .60 for reading. In particular, the EPL estimates based on the polytomous performance categories were very close to the VAM estimates in terms of ranking teachers (above .9 rank correlations for mathematics; above .80 rank correlations for reading).

Second, both the relationship to student characteristics and the relationship to teacher characteristics were not noticeably different between the EPL and VAM estimates. Still, the associations with the student background variables tended to be weaker in the EPL than in the VAM. In particular, the relationship to students' prior test scores tended to be weaker in the EPL based on the dichotomous outcome, than in the VAM. Most of the teacher background variables did not predict both the EPL and VAM estimates; only advanced degrees and years of teaching were significant predictors of the teacher effect estimates in secondary schools.

Third, intra-teacher rank correlations across different subjects and different grade levels also were similar between the EPL and the VAM. Correlations of the estimates between mathematics and reading ranged from .18 to .40, except Grade 7, and the correlation tended to be

marginally larger in the VAM estimates than in the EPL. Consistency of the estimates between different grades was small and not significant, except for the EPLs based on the polytomous outcomes between Grade 5 and Grade 6.

Finally, the fitted EPERF-based models can provide teachers and policy-makers with additional information of how the average or the individual teacher performed with his/her students assigned to certain challenge indices.



- Small dots represent the students' observed (actual) performance levels
- Triangles represent the students' expected probabilities of being proficient or expected performance levels given teachers' EPL estimates

Figure 4-6. Examples of individual teachers' educator performance functions (EPL-D1PL on the top; EPL-P1PL on the bottom)

CHAPTER 5. EXAMINATION OF THE MODEL-FIT

This chapter monitors the model fit and several quantities to evaluate assumptions of the educator performance function (EPERF)-based models. Some model fit indexes of the education production function (EPROF)-based models also are provided for reference. Three subtopics are addressed: 1) model fit of the EPERF- and EPROF-based models; 2) conditional independence of student success; and 3) amount of dependency of student success within each teacher.

5.1. Model fit of the EPERF-based models

First, in order to have general ideas about the model fit of the fitted EPERF models, the log-likelihood value of each fitted EPERF model, and the difference in the model fit between the 1PL and 2PL models were observed in Table 5-1. For all grades and all models, log-likelihood values tended to be smaller in mathematics than in reading, which suggests that overall the models fitted better for mathematics than for reading. Differences between the EPERF-D1PL and EPERF-D2PL models in the log-likelihood ranged from approximately 20 to 80 depending on different grades and subjects, all of which were statistically significant with 2 degrees of freedom but were relatively small, considering the approximate 50,000 or 100,000 sample size in each model. In particular, the difference in the model fit relative to the associated sample size tended to be much smaller in secondary schools. This implies that individual teachers' random slopes were not considerably different from each other in the dichotomous EPERF-based models. Differences between the EPERF-P1PL and EPERF-P2PL models were larger than those between the dichotomous models, especially for mathematics. Note that in mathematics the difference in

the model fit increases as the grade increases. Apparently, taking into account the teacher-unique slope parameter in Grade 7 mathematics somewhat improved the model fit.

Table 5-1. Log-likelihood for the EPERF-based models

	N	Dichotomous outcomes			Polytomous outcomes		
		EPERF-D1PL	EPERF-D2PL	-2Δ	EPERF-P1PL	EPERF-P2PL	-2Δ
Math							
Grade 4	46,461	-17,472.27	-17,431.77	80.99	-36,113.15	-36,083.35	59.60
Grade 5	55,493	-20,496.53	-20,475.40	42.26	-43,274.48	-43,156.76	235.44
Grade 6	97,768	-34,010.06	-33,996.53	27.06	-73,077.79	-72,942.17	271.24
Grade 7	105,888	-32,081.39	-32,051.78	59.22	-80,111.54	-79,897.27	428.54
Reading							
Grade 4	51,438	-19,625.03	-19,605.26	39.54	-42,568.46	-42,510.93	115.06
Grade 5	61,809	-23,624.31	-23,587.99	72.64	-56,302.43	-56,239.44	125.98
Grade 6	115,025	-45,505.10	-45,493.64	22.92	-102,035.79	-101,991.15	89.28
Grade 7	124,473	-53,110.33	-53,090.01	40.64	-105,968.42	-105,892.09	152.66

For reference, overall model fit indexes of the EPROF-based models are provided in Table A1 in the appendix. The log-likelihood values of the VAM-RE were similar to or larger than those of the polytomous EPERF. R-square values of the VAM-AR or of the VAM-FE were around .6 for mathematics and .5 for reading; at best, those of the VAM-GS were .2.

Second, error rates of the EPERFs were computed by the teacher using Equation 3-5 for the cases of dichotomous EPERF-based models and Equation 3-6 for those of polytomous EPERF-based models; their means, standard deviations, and ranges are shown in Table 5-2. Whether the error rates of the models were statistically lower than the null error rate, whichever was smaller between the proportion of observed successes and the proportion of observed failures, were tested, and the *t*-values are displayed in the last column of the same table. Results

of the EPERF-D1PL and the EPERF-D2PL were almost identical, as were those of the EPERF-P1PL and the EPERF-P2PL.

Table 5-2. Error rate of the EPERF-D1PL and EPERF-P1PL

	Null[1]		Model-based		t-value
			M (SD)	Range[2]	
EPERF-D1PL					
Mathematics					
Grade 4	P(S)	.39 (.21)	.15 (.09)	(0, .58)	47.85
Grade 5	P(S)	.37 (.22)	.15 (.08)	(0, .58)	42.52
Grade 6	P(S)	.33 (.24)	.14 (.08)	(0, .47)	37.95
Grade 7	P(S)	.28 (.23)	.12 (.08)	(0, .43)	30.20
Reading					
Grade 4	1-P(S)	.29 (.17)	.17 (.09)	(0, .55)	38.74
Grade 5	1-P(S)	.32 (.17)	.16 (.08)	(0, .61)	45.71
Grade 6	1-P(S)	.43 (.23)	.17 (.08)	(0, .46)	51.48
Grade 7	1-P(S)	.43 (.22)	.19 (.08)	(0, .53)	47.51
EPERF-P1PL					
Mathematics					
Grade 4	1-P(S)	.48 (.11)	.35 (.11)	(0, .86)	44.21
Grade 5	1-P(S)	.49 (.11)	.35 (.11)	(0, .85)	42.03
Grade 6	P(S)	.49 (.13)	.33 (.11)	(0, .78)	38.91
Grade 7	P(S)	.47 (.13)	.32 (.11)	(0, .83)	35.75
Reading					
Grade 4	1-P(S)	.32 (.08)	.36 (.11)	(.05, .74)	-15.10
Grade 5	1-P(S)	.31 (.10)	.41 (.10)	(.08, .75)	-34.53
Grade 6	1-P(S)	.37 (.12)	.40 (.09)	(.00, .74)	-8.48
Grade 7	1-P(S)	.37 (.10)	.38 (.08)	(.06, .82)	-4.20

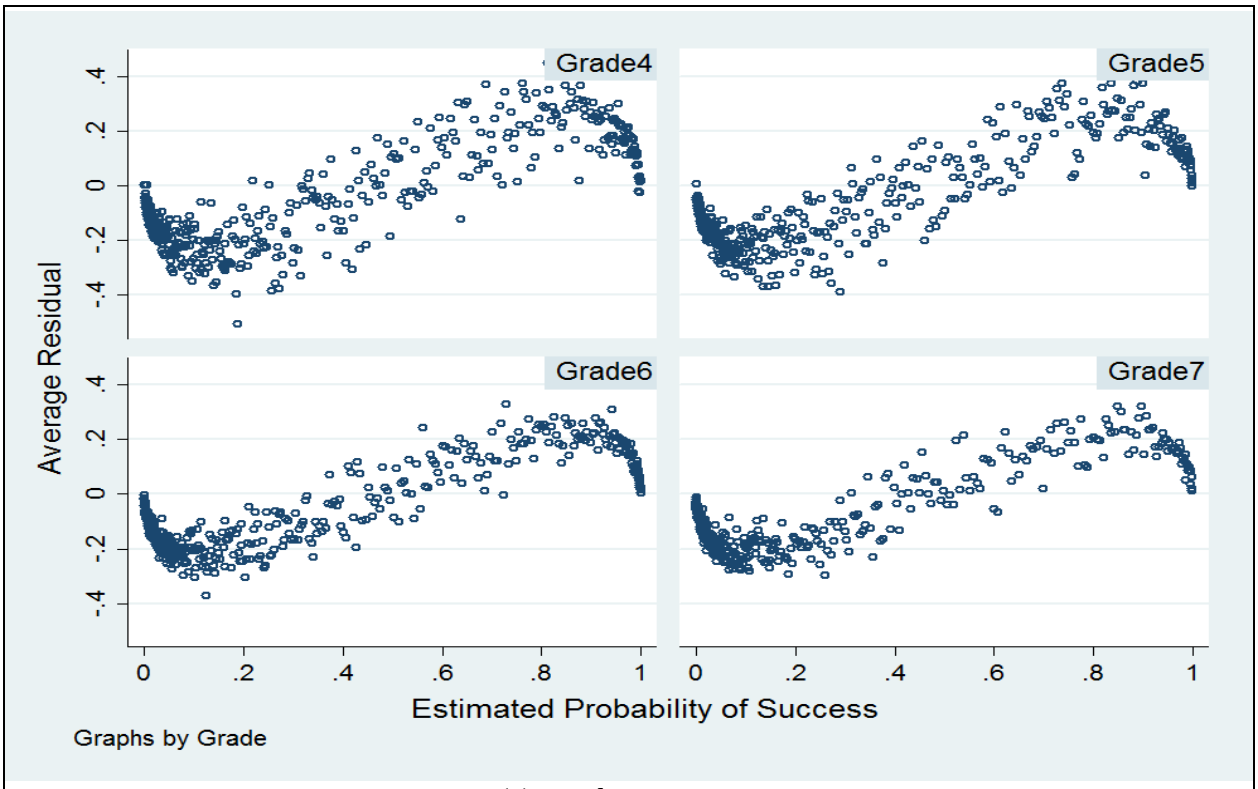
[1.] Null model is the proportion of 1s (success, P(S)), or the proportion of 0s (not success, 1-P(S)), which is simply to assign the same probability to each case without any predictor. The smaller number between P(S) and 1-P(S) was compared to the error rate based on the EPERF.
[2.] Range represents (min, max)

For the dichotomous EPERF-based models, the average error rate in mathematics was up to 15%, and it decreased as the grade-level increased, while in reading it was up to 19%; error rates tended to be smaller in mathematics than in reading. Error rates of the models were significantly smaller compared to the null model in both mathematics and reading (see the

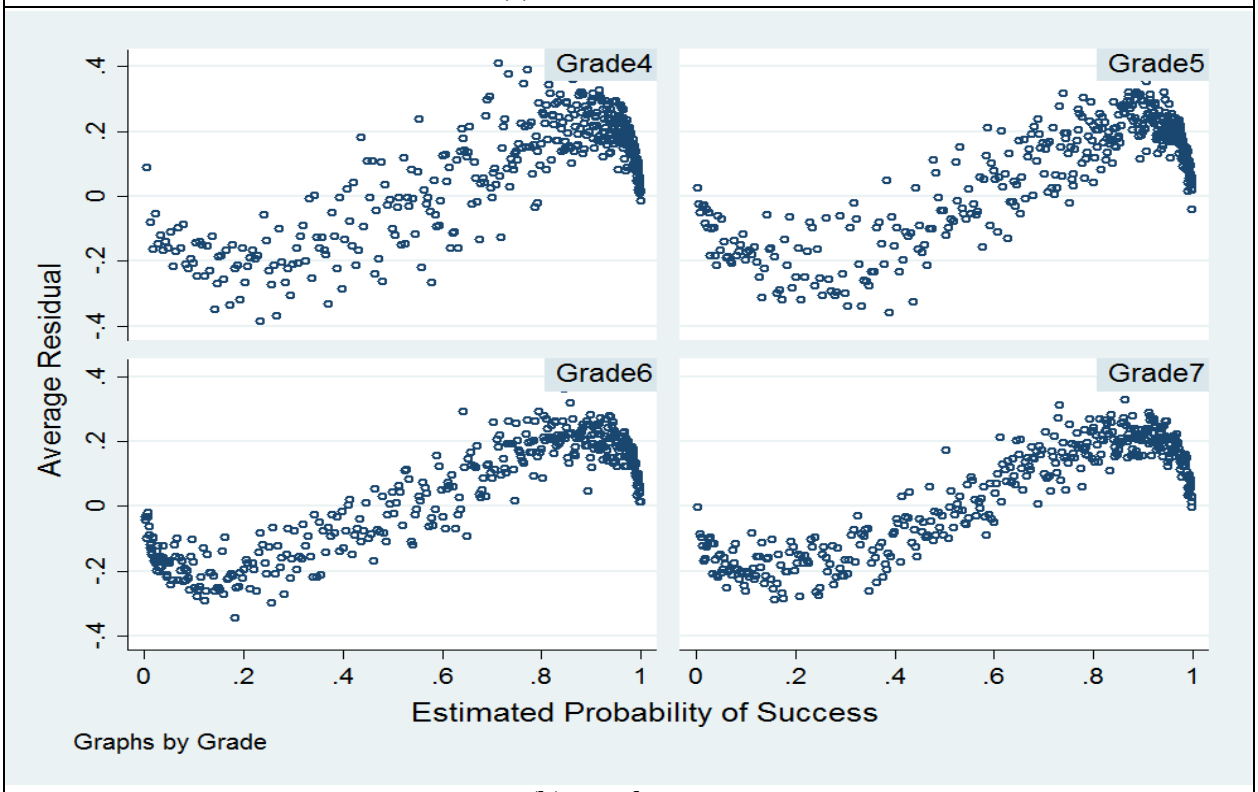
associated *t*-values). For the polytmous EPERF-based models, while the error rate of the model for mathematics was significantly smaller than that of the null, this did not hold true for reading. It does not appear that the polytomous EPERF-based models for reading predict the students' successes better than do the null.

Student-level residuals of the dichotomous EPERF-based models were computed based on Equation 3-3; binned plots are displayed in Figures 5-1 and 5-2. Bins represent the 400 quantile groups of students sorted by the expected probability of success (Figure 5-1) or by the average challenge index (Figure 5-2). As shown, average residuals delineated clear common patterns for all grade-levels and both subjects: S-shape along with the estimated probability of success; reverse S-shape along with the challenge index. As per Figure 5-1, residuals were smaller for groups whose expected probability of success was closer to 0, .5, or 1. For groups whose expected probability of success was below .5, the probability of success was underestimated, for those whose expected probability of success was above .5, the probability of success was overestimated. As per Figure 5-2, for student groups whose average challenge level corresponded to average teacher EPL, or was very high or low, residuals were minimal. When the average student CI was smaller than the average teacher EPL, the probability of success tended to be over-estimated; when the average student CI was larger than the average teacher EPL, the probability of success tended to be under-estimated.

Teacher-level residuals computed according to Equation 3-4 also were plotted in Figure A1 in the appendix. Residuals were distributed in ellipses; no legible pattern was found.

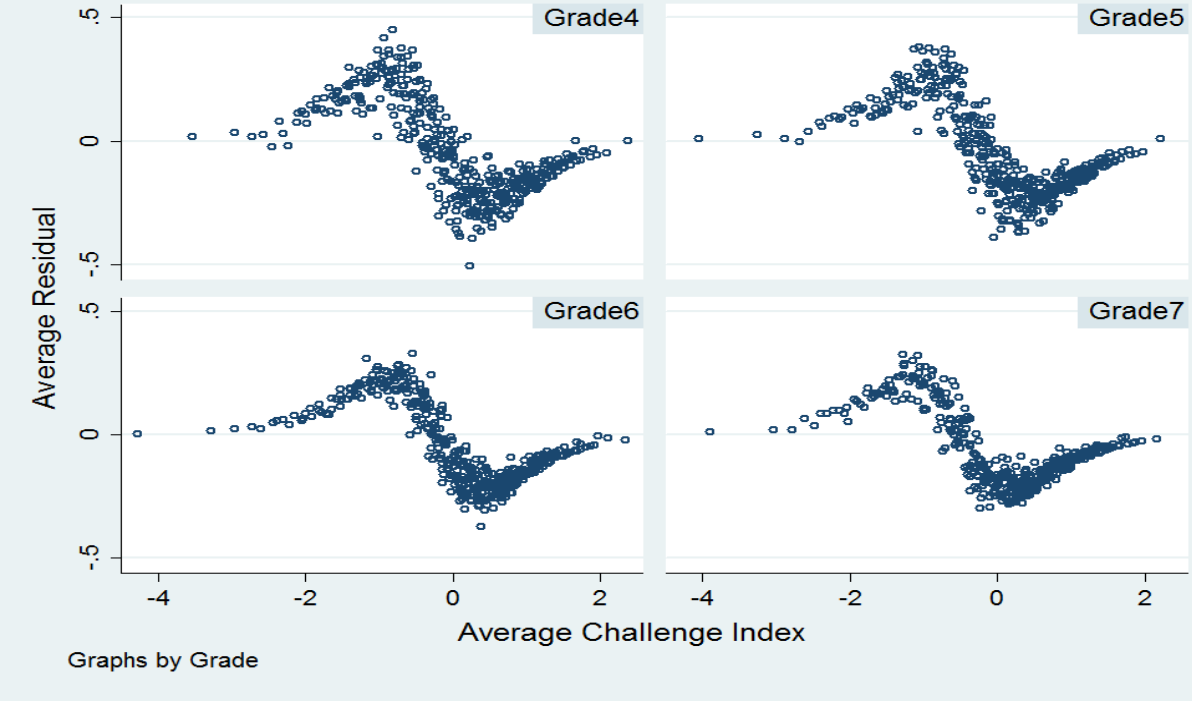


(a) Mathematics

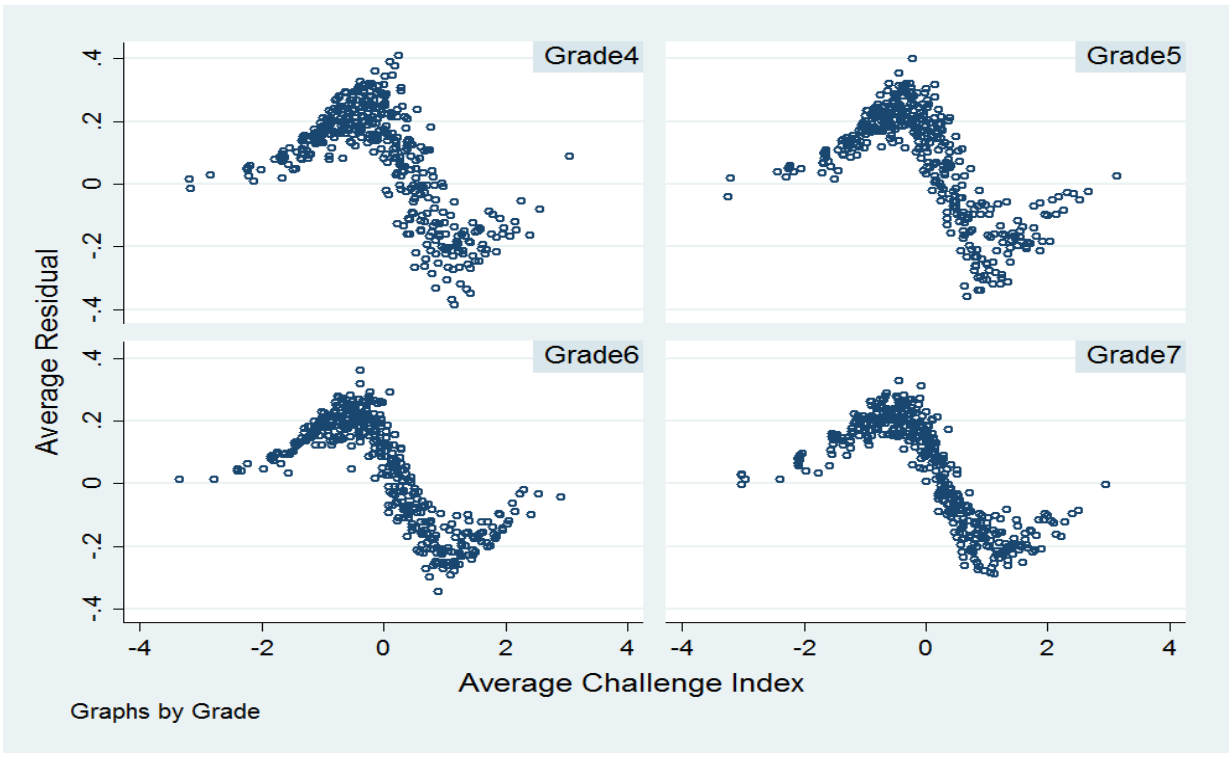


(b) Reading

Figure 5-1. Binned student-level residual plots via their expected probability of success (EPERF-D1PL)



(a) Mathematics



(b) Reading

Figure 5-2 Binned student-level residual plots via their average challenge index (EPERF-D1PL)

5.2. Conditional Independence of Student Success

Conditional independence of student success implies that once teacher's EPL is fixed, the probability of an individual student's success is independent each other. Although several different ways to evaluate the assumption have been suggested in the IRT context, in this study, as a preliminary examination, the distribution of Q_3 values was evaluated. Q_3 is basically a measure of dependency (correlation) of the residuals between heterogeneous groups of students shown in Equation 3-7, and the expected distribution of Q_3 when the conditional independency is tenable, is normal with a mean of 0 and variance of 1/(N-3) (Yen, 1984). Observed distributions of Q_3 based on the EPERF-D1PL are presented in Table 5-3.

Table 5-3. Distribution of correlations of residuals among the different quantile groups of student CI (EPERF-D1PL)

	Mean	SD	Min	Max	t-value[1]
Mathematics (50 groups)					
Grade 4	-.025	.076	-.686	.287	8.93
Grade 5	-.025	.066	-.438	.218	10.36
Grade 6	-.014	.053	-.259	.514	5.90
Grade 7	-.014	.049	-.195	.467	7.33
Reading (25 groups)					
Grade 4	-.016	.041	-.119	.098	5.16
Grade 5	-.010	.060	-.129	.404	2.08
Grade 6	-.018	.038	-.139	.097	6.71
Grade 7	-.012	.039	-.140	.097	4.24

1. Test statistics of the mean difference between the observed Q_3 and the expected Q_3

The mean of correlations of residuals among different groups of students was close to 0 in mathematics and reading. Distributions of the correlation, however, were significantly different from the normal distribution with 0 mean and 1/(N-3) standard deviation in every grade

level, as shown in the associated *t*-value in the last column of Table 5-3. This suggests that the models were likely to violate the assumption of conditional independence; the conditional dependency implies sources other than teacher EPL affect student success, but are not taken into account in the models. For example, student challenge index may have omitted important indicators predicting student success. This can be also a signal of peer effect in classrooms. The degree of associations between the residuals tended to be considerably larger in mathematics than in reading: the correlations in mathematics ranged from -.68 to .47 across all grade levels; those in reading ranged from -.14 to .40.

5.3. Dependency among the Student Success

In order to examine the amount of dependency among student outcomes for the same teacher, conditional intra-class correlations from the EPERF-based model and from the EPROF-based model were observed according to Equation 3-8 and Equation 3-9 (see Table 5-4). Note that dependency among the student success for the same teacher is distinguished from conditional independence among the student success across teachers. Dependency among the student success reflects the amount of variance in students' successes or achievement explained by the estimated teacher random effects. The larger the amount of variance in the student success explained by the teacher random effect, the higher the intra-class correlation. In the meantime, the conditional independence means that no dependency of the variance in the student success is unexplained by the teacher effect among different success.

Intra-class correlations were less than .20 for both the EPERF-D1PL and VAM-RE. The amount of dependency among the student success for the same teacher tended to be larger in the

EPERF-D1PL models than in the VAM-RE models. The dependency was almost thrice larger in mathematics than in reading, which is consistent with the previous studies on VAM. For reference, all estimated random effects from the four different EPERF-based models and the VAM-RE, were displayed in Table A2 in the appendix.

Table 5-4. Intra-class correlations of the EPERF-D1PL and EPROF-RE

	Mathematics				Reading			
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 4	Grade 5	Grade 6	Grade 7
EPEPF-D1PL	.155	.189	.164	.170	.059	.064	.073	.036
VAM-RE	.123	.164	.142	.121	.051	.038	.069	.037

CHAPTER 6. SENSITIVITY ANALYSIS

This chapter explores how locations of cut-scores and the number of student performance categories in a test affect the educator performance function (EPERF)-based teacher effect estimation, through a series of small simulations with real data. It involves an intensive qualitative process, known as standard setting procedure, in practice to determine the number of performance categories, and in sequence to set the cut-scores for each category to sort students based on the level of performance standards of state benchmarks. From a practical viewpoint, therefore, manipulating cut-scores arbitrarily in simulations may not be sensible. Still, to learn more about what to consider in applying this method, it is worthwhile to check how sensitive results from the EPREF-based method are to the locations of cut-scores and the number of performance categories. Of major concern in these simulations is how teacher effect estimates move across different scenarios.

6.1. Sensitivity to Different Locations of Cut-scores

As described in section 3.6, five scenarios for the dichotomous performance category and eight for the polytomous performance category were designed according to score-points dividing students by fixed proportions. For the dichotomous category, students were supposed to be divided into a non-proficient level and a proficient level by the following fixed proportions: (1) 7:3; (2) 6:4; (3) 5:5; (4) 4:6; and (5) 3:7. For the polytomous category, the students were separated into basic level, partially proficient level, proficient level, and advanced level by the following fixed proportions: (1) 4:3:2:1; (2) 3:3:2:2; (3) 3:3:3:1; (4) 2:3:4:1 (5) 1:4:4:1; (6)

1:3:5:1; (7) 2:2:4:2; (8) 1:2:5:2. Cut-points were determined by maximum values of the IRT scale score in the lower performance level in each scenario, and are displayed in Table 6-1 for dichotomous categories, and in Table 6-2 for polytomous categories. Each performance level's means and standard deviations of the scale scores and challenge index (CI) are shown in the same tables.

Looking at the scenarios for dichotomous performance categories (see Table 6-1), as the proportion of students at the proficient level increases from Scenario 1 to Scenario 5, the cut-score for being at the proficient level, the maximum value of the non-proficient level decreases from 1.64 to .58. And the average scale score of the students at the proficient level decreases, while the standard deviation increases. The mean of the non-proficient level decreases as the variance decreases. By contrast, average CI of students at the proficient level increases, as the proportion of students at the proficient level increases. While the range of students' CI within each performance level is constant across scenarios, the deviation of CI in the non-proficient group fluctuates depending on scenario.

Table 6-1. Distributions of student test scores and challenge index by the simulated cut-score for the dichotomous performance category

		Non-proficient		Proficient		Cut-score[4]	CI (Min, Max)
	Ratios[1]	Score[2]	CI[3]	Score	CI		
Original	6.1:3.9	.56 (.53)	.56 (.67)	2.30 (.74)	-.83 (.87)	1.40	NP (-2.62, 5.17)
Scenario 1	7:3	.67 (.59)	.48 (.70)	2.49 (.72)	-.97 (.87)	1.63	
Scenario 2	6:4	.53 (.51)	.59 (.66)	2.24 (.75)	-.78 (.87)	1.33	
Scenario 3	5:5	.38 (.45)	.71 (.43)	2.02 (.79)	-.60 (.87)	1.03	P (-5.57, 2.40)
Scenario 4	4:6	.26 (.40)	.80 (.61)	1.86 (.82)	-.47 (.89)	.81	
Scenario 5	3:7	.13 (.35)	.91 (.59)	1.71 (.86)	-.35 (.90)	.58	

1. Proportion of students at the non-proficient level to those at the proficient level
2. Average IRT scale score; standard deviation in parentheses
3. Average challenge index; standard deviation in parentheses
4. Maximum value at the non-proficient level

Table 6-2. Distribution of student test scores and challenge index by the simulated cut-score for the polytomous performance category

Polytomous	Scenario 1 4:3:2:1[1]		Scenario 2 3:3:2:2		Scenario 3 3:3:3:1		Scenario 4 2:3:4:1	
	Score[2]	CI[3]	Score	CI	Score	CI	Score	CI
Basic	.28 (.40)	.78 (.61)	.14 (.35)	.89 (.59)	.14 (.35)	.89 (.59)	-.06 (.29)	1.03 (.57)
Partially Proficient	1.24 (.24)	.02 (.56)	.98 (.21)	.23 (.54)	.98 (.21)	.23 (.54)	.70 (.21)	.45 (.55)
Proficient	2.11 (.29)	-.69 (.66)	1.70 (.21)	-.35 (.59)	1.92 (.37)	-.53 (.66)	1.74 (.45)	-.39 (.68)
Advanced	3.39 (.66)	-1.74 (.92)	2.86 (.69)	-1.30 (.89)	3.39 (.66)	-1.74 (.92)	3.39 (.66)	-1.74 (.92)
Cut-score 1[4]	.81	5.17	.58	5.17	.58	5.17	.27	5.17
Cut-score 2[5]	1.63	2.80	1.33	2.80	1.33	2.80	1.03	2.80
Cut-score 3[6]	2.65	2.29	2.05	2.29	2.65	2.29	2.65	2.29
	Scenario 5 2:4:2:2		Scenario 6 2:5:2:1		Scenario 7 1:4:4:1		Scenario 8 1:3:5:1	
	Score	CI	Score	CI	Score	CI	Score	CI
Basic	-.06 (.29)	1.04 (.57)	-.06 (.29)	1.04 (.57)	-.24 (.25)	1.15 (.56)	-.24 (.25)	1.15 (.56)
Partially Proficient	.83 (.30)	.34 (.57)	.97 (.38)	.23 (.61)	.59 (.28)	.54 (.58)	.73 (.36)	.43 (.60)
Proficient	1.70 (.21)	-.35 (.59)	2.11 (.29)	-.69 (.66)	1.74 (.45)	-.39 (.68)	1.92 (.37)	-.53 (.66)
Advanced	2.85 (.69)	-1.30 (.89)	3.39 (.66)	-1.74 (.92)	3.39 (.66)	-1.74 (.92)	3.39 (.66)	-1.74 (.92)
Cut-score 1	.27	5.17	.27	5.17	.02	5.17	.02	5.17
Cut-score 2	1.33	2.80	1.63	2.80	1.03	3.02	1.33	3.02
Cut-score 3	2.65	2.29	2.65	2.29	2.65	2.29	2.65	2.29

1. Basic level: Partially proficient level: Proficient level: Advanced level
2. Average IRT scale score; standard deviation in the parentheses
3. Average challenge index; standard deviation in the parentheses
4. Maximum value at the basic level
5. Maximum value at the partially proficient level
6. Maximum value at the proficient level

Table 6-2 shows the scenarios of the polytomous performance levels. As the proportion of students at either the proficient or advanced level enlarged, from Scenario 1 to Scenario 8, cut-scores of the basic level dropped from .81 to .02. Depending on scenario, cut-scores of the partially proficient level ranged from 1.03 to 1.63, while those of the proficient level ranged from 2.05 to 2.65. Average scores of students at each performance level also changed across scenarios, as cut-scores moved.

Distributions of the teacher effect estimates from different scenarios and their correlations are displayed in Table 6-3 for the dichotomous category, and in Table 6-4 for the polytomous category. Since the EPLs resulting from the 1PL models were very similar to those from the 2PL models, only results from the 1PL models – EPL-D1PL and EPL-P1PL – are displayed here. As Table 6-3 shows, as cut-scores for the proficient level dropped, that is, the proportion of students in the proficient level increased from Scenario 1 to Scenario 5, the mean of teacher effect estimates increased from -.18 to .24 (See Figure 6-1(a)). Note that the average EPL increased as the proportion of students achieving proficient level increased, which is one of the characteristics distinguishing the EPERF-based models from the EPROF-based models. Resulting EPLs reflect how successful teachers were in achieving their goals, helping their students to attain a desired performance level, so that the average EPL increases as much as the number of successful students increases. The average VAM, however, set to 0, does not move, and provides no practical information of how teachers performed on average.

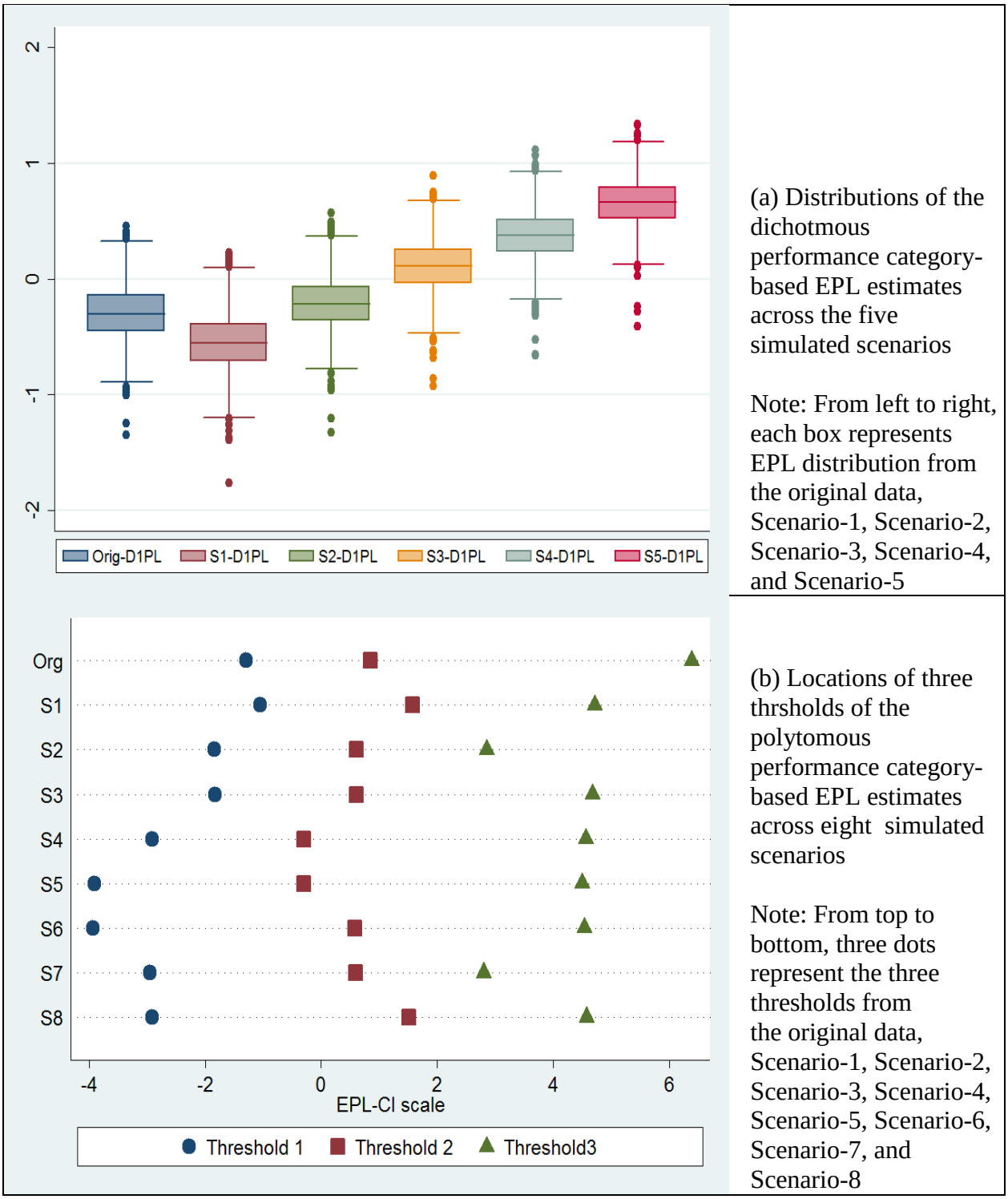
The EPL estimates from Scenario-D2 were closest to the original data, showing a .97 correlation. Notice that the proportion of proficient students in Scenario-D2 was most similar to the original data. By contrast, Scenario-D5, which was most apart from the original proportion of students by performance level, yielded the results most different from the original estimates

Table 6-3. Descriptive statistics of the teacher effect estimates from the five simulated scenarios by different cut-scores of the dichotomous category

N=1,731	EPL-D1PL						Rank correlations					
		M	SD	Min	Max	Slope	0.	1.	2.	3.	4.	VAM
0. Original	6.1:3.9	-.30	.23	-1.34	.45	-2.92						.86
1. Scenario-D1	7:3	-.55	.24	-1.76	.23	-2.92	.91					.86
2. Scenario-D2	6:4	-.21	.22	-1.32	.57	-2.94	.97	.88				.86
3. Scenario-D3	5:5	.11	.22	-.93	.89	-2.85	.85	.79	.88			.85
4. Scenario-D4	4:6	.37	.21	-.66	1.11	-2.80	.78	.72	.80	.89		.83
5. Scenario-D5	3:7	.66	.21	-.41	1.33	-2.68	.69	.64	.72	.81	.89	.78

Table 6-4. Descriptive statistics of the teacher effect estimates from the eight simulated scenarios by different cut-scores of the polytomous category

N=1,731	EPL-P1PL						Rank correlations								
		M	SD	Min	Max	Slope	0.	1.	2.	3.	4.	5.	6.	7.	VAM
0. Original	3.5:2.6:3.5:0.4	.00	.25	-1.18	1.01	-2.74									.92
1. Scenario-P1	4:3:2:1	.00	.26	-1.17	.84	-2.76	.96								.93
2. Scenario-P2	3:3:2:2	.00	.25	-1.16	.80	-2.77	.95	.95							.93
3. Scenario-P3	3:3:3:1	.00	.25	-1.17	.81	-2.73	.96	.95	.98						.93
4. Scenario-P4	2:3:4:1	.00	.25	-1.19	.85	-2.63	.91	.93	.92	.94					.93
5. Scenario-P5	2:4:2:2	.00	.25	-1.23	.81	-2.69	.92	.91	.96	.94	.95				.93
6. Scenario-P6	2:5:2:1	.00	.26	-1.23	.84	-2.64	.90	.93	.92	.93	.95	.92			.93
7. Scenario-P7	1:4:4:1	.00	.25	-1.18	.85	-2.58	.91	.92	.91	.92	.97	.92	.97		.93
8. Scenario-P8	1:3:5:1	.00	.25	-1.25	.82	-2.61	.90	.91	.92	.94	.93	.96	.93	.76	.93



(a) Distributions of the dichotmous performance category-based EPL estimates across the five simulated scenarios

Note: From left to right, each box represents EPL distribution from the original data, Scenario-1, Scenario-2, Scenario-3, Scenario-4, and Scenario-5

(b) Locations of three thrsholds of the polytomous performance category-based EPL estimates across eight simulated scenarios

Note: From top to bottom, three dots represent the three thresholds from the original data, Scenario-1, Scenario-2, Scenario-3, Scenario-4, Scenario-5, Scenario-6, Scenario-7, and Scenario-8

Figure 6-1. Comparison of the teacher effect estimates depending on the different simulated cut-scores

(.67 correlation). The estimated slope-parameter was largest for Scenario2. Slope-parameters became slightly smaller, as the proportion of proficient students grew. The correlation of the EPL estimate from each scenario and the VAM estimates ranged from .74 to .85. Interestingly, it turned out that original VAM estimates were closer to EPL estimates resulting from scenarios that were more similar to the original proportion of students in proficient levels, such as Scenario-D1 to Scenario-D3, than Scenario-D4 and D5.

Looking at the results of the simulated polytomous cases in Table 6-4, no considerable difference in distributions of the estimates among the eight scenarios was found. Correlations with original estimates were above .9 in every scenario, and the estimates were strongly associated with each other, showing more than .9 correlations. Correlations with original VAM estimates were also high and consistent across the scenarios. The slope parameter estimate slightly dropped, in particular in Scenario-P7, when proportions of students in both basic and advanced levels were relatively small. Importantly, the three thresholds in the fitted EPERF-based models moved depending on the simulated scenarios of cut-scores as shown in Figure 6-1(b). For reference, rank correlations of the estimates between all different scenarios – five dichotomous and eight polytomous – are provided in Table A3; their scatterplots are provided in Figure A2 for dichotomous scenarios and in Figure A3 for polytomous scenarios, in the appendix.

The size of the estimated random effect and the model-fit index also were monitored in Table 6-5. For the dichotomous EPERF, while the size of random effect was largest when the proportion of non-proficient students to proficient students was 7:3, it was smallest when the proportion was the opposite. For the polytomous EPERF, the original proportion of students showed the biggest random effect.

Table 6-5. The size of random effect and log-likelihood of the scenarios

	Ratio	Variance of random intercept	Log likelihood
Dichotomous original	6.1:3.9	.644 (.032)	-34010.06
Scenario-D1	7:3	.732 (.038)	-30632.41
Scenario-D2	6:4	.617 (.031)	-34840.89
Scenario-D3	5:5	.564 (.028)	-37526.20
Scenario-D4	4:6	.513 (.026)	-37612.34
Scenario-D5	3:7	.455 (.024)	-36187.82
Polytomous original	3.5:2.6:3.5:0.4	.595 (.026)	-73077.79
Scenario-P1	4:3:2:1	.506 (.022)	-85371.98
Scenario-P2	3:3:2:2	.446 (.020)	-95252.82
Scenario-P3	3:3:3:1	.447 (.020)	-89743.90
Scenario-P4	2:3:4:1	.407 (.018)	-88692.51
Scenario-P5	2:4:2:2	.538 (.023)	-83000.56
Scenario-P6	2:5:2:1	.558 (.024)	-72903.55
Scenario-P7	1:4:4:1	.398 (.018)	-83146.55
Scenario-P8	1:3:5:1	.522 (.023)	-71099.52

6.2. Sensitivity to the Number of Performance Categories

Four different scenarios – six, seven, eight and ten performance categories as described in section 3-6 – were considered, in order to evaluate whether the simulated number of performance categories changed teachers' educator performance level (EPL) estimates. Rank correlations of the resulting teacher EPL estimates among the four different simulation scenarios and the two original categories are displayed in Table 6-6.

Rank correlations between the different scenarios were above .9; those with original EPL estimates based on the four performance categories were above .9; and those with VAM

estimates were above .9 also. Correlations with the original EPL using dichotomous performance category ranged from .87 to .90, which was lower than those among the polytomous category-based EPL estimates. As expected, EPL estimates among the scenarios were closer when the number of categories were similar.

Table 6-6. Rank correlations of the teacher effect estimates among the simulated scenarios by different number of performance categories

N=1,758	Two[1]	Four	Six	Seven	Eight	VAM-RE	Prior test score[3]
Four Performance levels[1]	.90					.92	.21
Six Performance levels[2]	.88	.94				.96	.23
Seven Performance levels[2]	.87	.96	.97			.96	.23
Eight Performance levels[2]	.87	.96	.97	.99		.97	.23
Ten Performance levels[2]	.87	.94	.98	.98	.98	.97	.24

[1.] Original performance categories determined in the original data;
[2.] Simulated performance categories
[3.] Teacher-level

Associations with the original VAM-RE estimates tended to be slightly stronger as the number of categories increased. Those with average student prior test scores also marginally increased, along with increasing number of performance categories. It is concluded that the teacher effect estimates were not much changed according to the number of performance categories. However, the size of random effects reduced and the model-fit slightly worsened as the number of performance categories increased.

CHAPTER 7. CONCLUSION AND DISCUSSION

7.1. Summary of Findings

This study aimed to introduce the educator performance function (EPERF)-based teacher effect estimation model and to evaluate its feasibility, by comparing it with the currently prevailing method, the education production function (EPROF)-based value-added model. It thereby is expected to illuminate research- or policy-relevant issues of its implementation. While the EPROF is a linear model to describe the relationship between students' test scores and backgrounds and their teachers' effects, the EPERF is a non-linear probability model to describe the relationship between teachers' proficiency, and their students' characteristics and successes in reaching a certain performance level.

The EPERF-based models are mainly distinguished from the EPROF-based ones by including the following characteristics: (1) the outcome variable is a dichotomous/categorical variable of performance standard classification; (2) each student is assigned a challenge index value, which is a quantity of the degree of difficulty that teachers face in teaching the student to attain a desired performance standard; (3) the teacher effect is treated as a latent trait, a proficiency in helping students achieve a desired performance standard given their students' challenge levels, and it is scaled to be comparable with the student CI; and (4) the teacher effect estimates are sample-independent, once each student's challenge level is determined.

For empirical comparisons between the two methods, this study investigated (1) how the two different teacher effect estimates – the educator performance level (EPL) resulting from the EPERF-based method and the value-added measures (VAM) resulting from the EPROF-based

method – are consistent; (2) whether the model fit of EPERF is acceptable; and (3) whether results from the EPERF-based methods are robust to the locations of cut-scores and the number of performance categories. Main findings for each question are described and discussed below.

Regarding the first question, comparing the teacher effect estimates, the rank correlation between the estimates, the relationship to student and teacher characteristics, and the consistency between different subjects or grade levels were examined. First, teacher rankings that the EPERF-based and the EPROF-based methods produced were similar rather than substantially different; they showed above .8 rank correlations for mathematics and mostly above .7 rank correlations for reading. In particular, when the four student performance categories were used for the EPERF, the rank correlation between the EPL and the VAM was higher than when the dichotomous performance categories were used. Second, both the relationship to student characteristics and the relationship to teacher characteristics were not noticeably different between the EPL and VAM estimates. Still, associations with student background variables tended to be slightly weaker in the EPL than in the VAM. Third, consistency of the teacher effect estimates between different subjects or different grade levels was also similar between the EPL and the VAM. However, intra-teacher rank correlations between mathematics and reading tended to be marginally higher in the VAM estimates than in the EPL.

Even though teacher rankings yielded by the two methods were similar in most cases, the EPERF-based methods offered additional interesting information of how average or individual teachers performed with their students toward their mutual goal to enable students to reach a certain performance standard of a state's benchmarks. For example, while variation in teacher effect estimates was larger in mathematics than in reading, it appears that on average teaching mathematics to help students attain the proficient level was more challenging than teaching

reading. The former is what previous studies on VAM have found consistently, but the latter has yet to be stated empirically and explicitly. In addition, from fitting the polytomous EPERF-based models, this study also observed no considerable teacher effect on the probability that students reach the advanced level, the highest level, in mathematics, whereas the teacher effect clearly decreased the probability of being at the basic level, and increased the probability of being at the proficient level.

In answering the second question, the EPERF's model-fit was evaluated via the residual analyses and error-rates. Conditional independence of student success and dependency of student success within each teacher also were examined. While the error rate of the dichotomous EPERF was noticeably smaller in both mathematics and reading, compared to the null models, the error rate of the polytomous EPERF for reading was higher than the null models. Residual analyses suggested that the model-fit of the EPERF was not perfect; what caused the misfit needs further investigation. The assumption of the conditional independence of student success was not completely sustained; this, too, needs scrutiny. In both reading and mathematics, the amount of dependency of student successes tended to be a little larger in the EPL than in the VAM,

For the third question, as a result of real-data simulations, the EPL based on the polytomous performance categories was quite robust to the location of cut-scores; rank correlations between different cut-scores were above .9. The EPL based on the dichotomous performance categories substantially altered teacher ranking when the proportion of the proficient level to the non-proficient level was opposite. The number of performance categories did not substantially change teacher ranking, showing above .87 rank correlations across all scenarios.

7.2. Discussion

Based on these initial examinations of applying the educator performance function (EPERF)-based teacher effect estimations and comparisons to the education production function (EPROF)-based teacher effect estimations, this section offers several suggestions about what should be considered in applying the new method in practice; it also elaborates relevant further research questions yet to be answered.

First, by example this study illustrated how the student challenge index (CI) can be composed. As a result of several experiments with a given set of student background variables available in the state data, the OLS-weighted sum score worked better than did the IRT-calibration, in terms of differentiating students in the degree to which they pose challenges to teachers. A limited variance was observed when applying the IRT-calibration given the number and types of variables. It appears that additional indicator variables, with respect to both the number and variety, were necessary in order for the student CI based on IRT-calibration to ensure a reasonable amount of variance.

Of more critical concern when constructing the student CI, however, is whether the variables sufficiently explain what kinds of students pose more challenges to their teachers. This is because the quality of teacher capability estimates from the EPERF relies substantially on the assumption that the CI considers and includes all possible factors beyond teacher control but that can influence student achievement. As CI indicators, this study used the 7 to 10 student background variables, including prior achievement, which are mostly demographic and dichotomous variables that most states commonly collect and are often used as covariates when fitting the EPROF-based models. This set may have omitted some important indicators, which potentially could have resulted in a violation of conditional independency of student success, and

in a misfit of the EPERF. Students' motivation and school or classroom level characteristics may be critical missing indicators. Therefore, what student characteristics can be additionally taken into account, and what type of the variables are better, are still open questions, given that in the EPERF-based method, student CI is a pivotal concept for measuring teacher effect.

Further, how to construct student CI – that is, which variables to include and how to weight the indicators – is crucial not only for achieving statistically better estimations of teacher EPL, but also as a process for seeking political consensus among educational stake-holders, especially teachers or schools, about defining teacher capability. In that sense, it would be helpful for teachers and/or schools to participate in discussing and determining student CI indicators and weights.

Second, this study examined four different EPERF-based models depending on the number of teacher random effects – one-parameter (random intercept) and two-parameters (random intercept and random slope), and depending on the number of student performance categories – dichotomous and polytomous outcomes. Findings indicate no practical difference between the one-parameter (1PL) and the two-parameter (2PL) models in their model-fit and in their ranking of the teacher effect estimates, despite statistical differences between the two models having been found. Thus, because of parsimoniousness, the 1PL models may be preferred.

Nevertheless, no practical difference between the 1PL and 2PL models in this study cannot abandon the potential usefulness of the 2PL models. Note that using the 2PL models – i.e., allowing a variation of the random slope across individual teachers – is based on the idea that teachers could differ significantly in moderating the relationship between student CI and their success in achieving a performance standard, which is completely feasible. It is still worthwhile

to continue to apply the 2PL models to see whether the restricted amount of variance in the random slope holds when other states' data are used, or even when the way to create the student CI with the same data change. Thereby, what random slope estimates imply about teachers and their performance can be clarified further.

For the polytomous EPERF used in this study, the slope of the student challenge level is fixed as identical for all four performance levels, assuming that the impact of the student challenge level on the probability of student success is identical among different performance levels. This assumption could be released by varying the slope across the four categories. More generalized models can be explored and applied, and it would be interesting to investigate how consistent are teacher effect estimates from the generalized models.

A prominent policy advantage of using polytomous EPERF-based models needs to be underlined: target performance level is varied across students, depending on their challenge levels. Note that in the EPROF-based VAM, there is no specific target test score to inform how well students are expected to perform in the achievement tests. Also, all students are equally expected to achieve higher (gain) scores than the others, regardless where they started. On the contrary, in the polytomous EPERF-based models, the target or expected performance level, which level each student is expected to obtain can be decided according to the student challenge level. In other words, which performance level teachers and students are expected to achieve depends on individual students' challenge levels; the higher a student's CI, the lower his/her expected performance level (expected scores). In addition, in Chapter 6 it was demonstrated that the polytomous EPERF-based models were not very sensitive to the number of performance categories and the locations of cut-scores.

Third, this study attempted to evaluate the basic model fit of the EPERF, and some evidence of misfit were found. However, the misfit does not invalidate usefulness of the EPERF, because the EPROF-based value-added models (VAM) also have the same issue, and the perfect fit of the model for estimating teacher effect is either impractical or unrealistic. Rather, scrutinizing the misfit is potentially beneficial guide to understanding what causes the misfit and exploring how to improve the model. Even though several key assumptions of the EPERF-based method were identified in section 2.2, which conditions in school or education settings might possibly violate the assumptions need thorough exploration. Also it would be meaningful to test some falsifiable assumptions empirically and to provide some guidance in determining when this method is more feasible for different conditions.

For example, distribution of student CI per teacher and its impact on teachers' EPL estimates were not monitored in this study. Whether the distribution of student CI is different or comparable across teachers can be the first question. If any significant teacher-level variance in student CI is found, whether that variation can influence teacher EPL estimates needs investigation. This also can be a way towards understanding student sorting in schools or districts.

Fourth, this study used the two-year student-teacher linked data containing only one-cohort due to restrictions on data availability. It would be worthwhile to use longitudinal and multiple-cohort data, and to check longitudinal consistency of estimates. Also it would be useful to replicate these results using other states' data.

Finally, it should be emphasized that the salient advantages of applying the EPERF-based method is to tighter coordination of student testing models and teacher or school effectiveness models tightly coordinated. Current educator effectiveness evaluation models demand that

teachers or schools to bring their students to higher gains than other students, regardless not only what students are expected to achieve, but also what teachers are expected to achieve with their students. Coordination between the student testing model and the educator effectiveness model can be ensured by circumscribing what teachers are accountable for in students' learning, based on what students are expected to achieve and by informing teachers of where they are in terms of what is to be expected.

APPENDIX

Table A1. Model fit indexes for the EPROF-based models

	VAM-RE	VAM-AR/FE	VAM-GS	N
Fit index	Log likelihood	R-square	R-square	
Mathematics				
Grade 4	-43,885.15	.639	.162	45,685
Grade 5	-41,984.19	.627	.181	54,525
Grade 6	-81,181.57	.660	.144	95,739
Grade 7	-80,141.43	.652	.116	103,794
Reading				
Grade 4	-59,465.13	.509	.081	50,590
Grade 5	-70,328.49	.546	.072	60,740
Grade 6	-132,280.98	.568	.066	112,760
Grade 7	-126,527.94	.533	.074	122,078

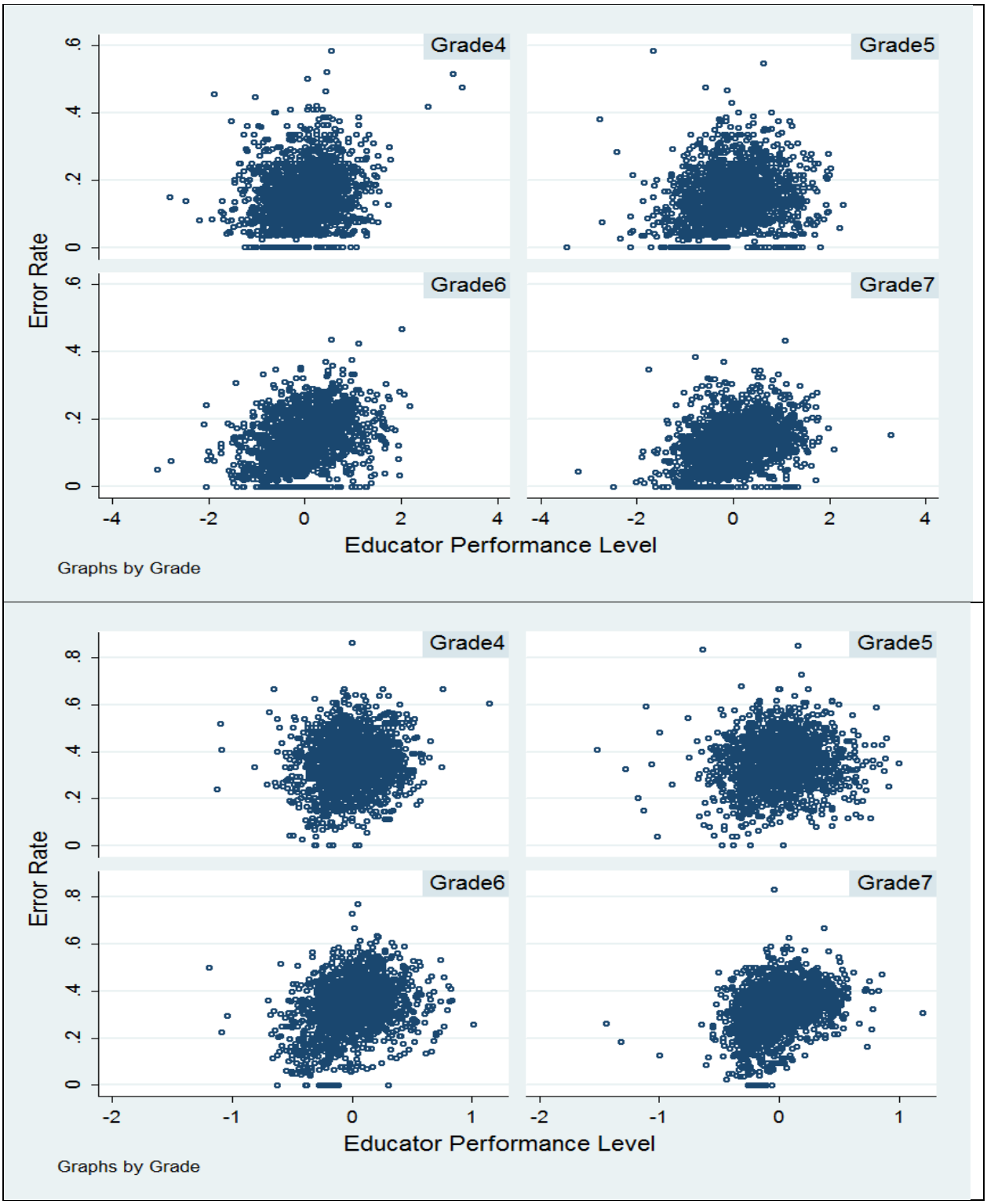


Figure A1. Error rate of student success predicted by the EPERF-based models via estimated educator performance level (EPERF-D1PL on the top; EPERF-P1PL on the bottom; Mathematics)

Table A2. Estimated random effects of the EPERF-based models and VAM-RE model

	EPL-D1PL	EPM-D2PL	EPL-P1PL	EPL-P2PL	VAM-RE
Math					
Random intercept	.604 (.035)	.565 (.036)	.501 (.025)	.481 (.025)	.052 (.002)
	.767 (.040)	.721 (.040)	.617 (.028)	.592 (.027)	.048 (.002)
	.644 (.032)	.623 (.032)	.595 (.026)	.581 (.026)	.050 (.002)
	.663 (.035)	.615 (.035)	.482 (.022)	.458 (.021)	.037 (.002)
Random slope		.310 (.044)		.092 (.015)	
		.112 (.030)		.119 (.014)	
		.092 (.025)		.101 (.012)	
		.135 (.025)		.099 (.009)	
Covariance		.016 (.027)		-.017 (.013)	
between intercept		-.095 (.025)		-.138 (.015)	
and slope		-.048 (.020)		-.056 (.012)	
		-.042 (.021)		-.114 (.011)	
Reading					
Random intercept	.208 (.018)	.193 (.021)	.183 (.012)		.032 (.002)
	.228 (.017)	.223 (.018)	.180 (.011)	.172 (.011)	.023 (.001)
	.260 (.015)	.246 (.015)	.278 (.012)	.271 (.012)	.044 (.002)
	.121 (.009)	.131 (.010)	.137 (.007)	.131 (.008)	.018 (.001)
Random slope		.141 (.030)			
		.185 (.028)		.073 (.009)	
		.047 (.015)		.035 (.005)	
		.059 (.012)		.035 (.004)	
Covariance		-.016 (.020)			
between intercept		-.059 (.015)		-.011 (.007)	
and slope		.025 (.011)		-.001 (.006)	
		-.029 (.007)		.003 (.004)	

Table A3. Rank correlations of the teacher EPL estimates among the simulated locations of cut-scores

		Scenario-D[1]1 7:3	Scenario-D2 6:4	Scenario-D3 5:5	Scenario-D4 4:6	Scenario-D5 3:7	Original-P[2]
Scenario-P[2]1	4:3:2:1	.88	.87	.89	.91	.81	.96
Scenario-P2	3:3:2:2	.85	.90	.87	.87	.87	.95
Scenario-P3	3:3:3:1	.83	.89	.87	.87	.87	.96
Scenario-P4	2:3:4:1	.77	.82	.90	.87	.83	.91
Scenario-P5	2:4:2:2	.78	.83	.90	.86	.80	.91
Scenario-P6	2:5:2:1	.83	.89	.83	.80	.77	.90
Scenario-P7	1:4:4:1	.84	.89	.85	.83	.80	.92
Scenario-P8	1:3:5:1	.86	.83	.80	.80	.79	.90
Original-D[1]	6.1:3.9	.91	.98	.85	.78	.69	.90

[1.] D indicates the dichotomous models; [2.] P indicates the polytomous models

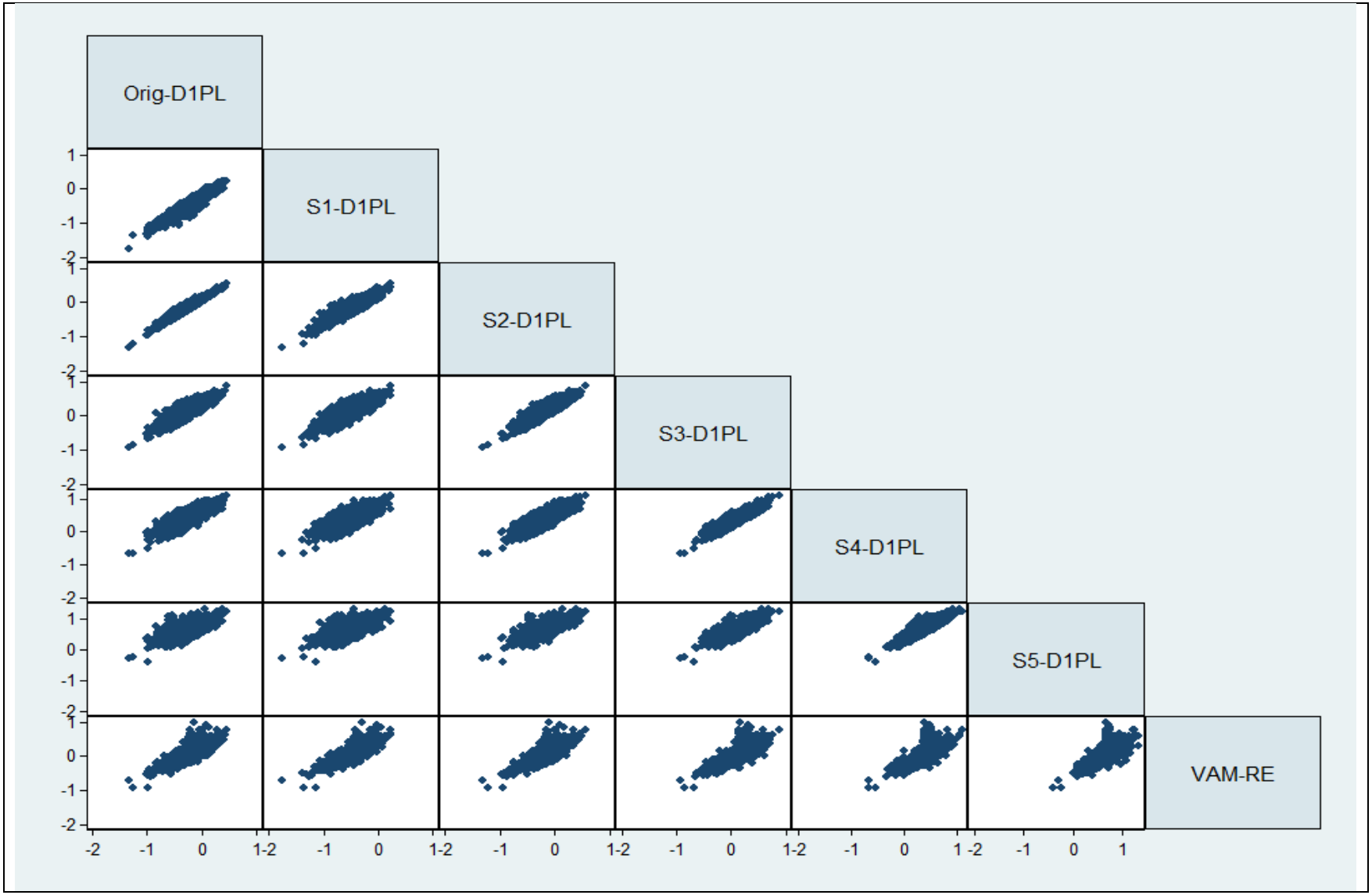


Figure A2. Scatterplots of the teacher effect estimates from different simulated cut-scores (Dichotomous category)

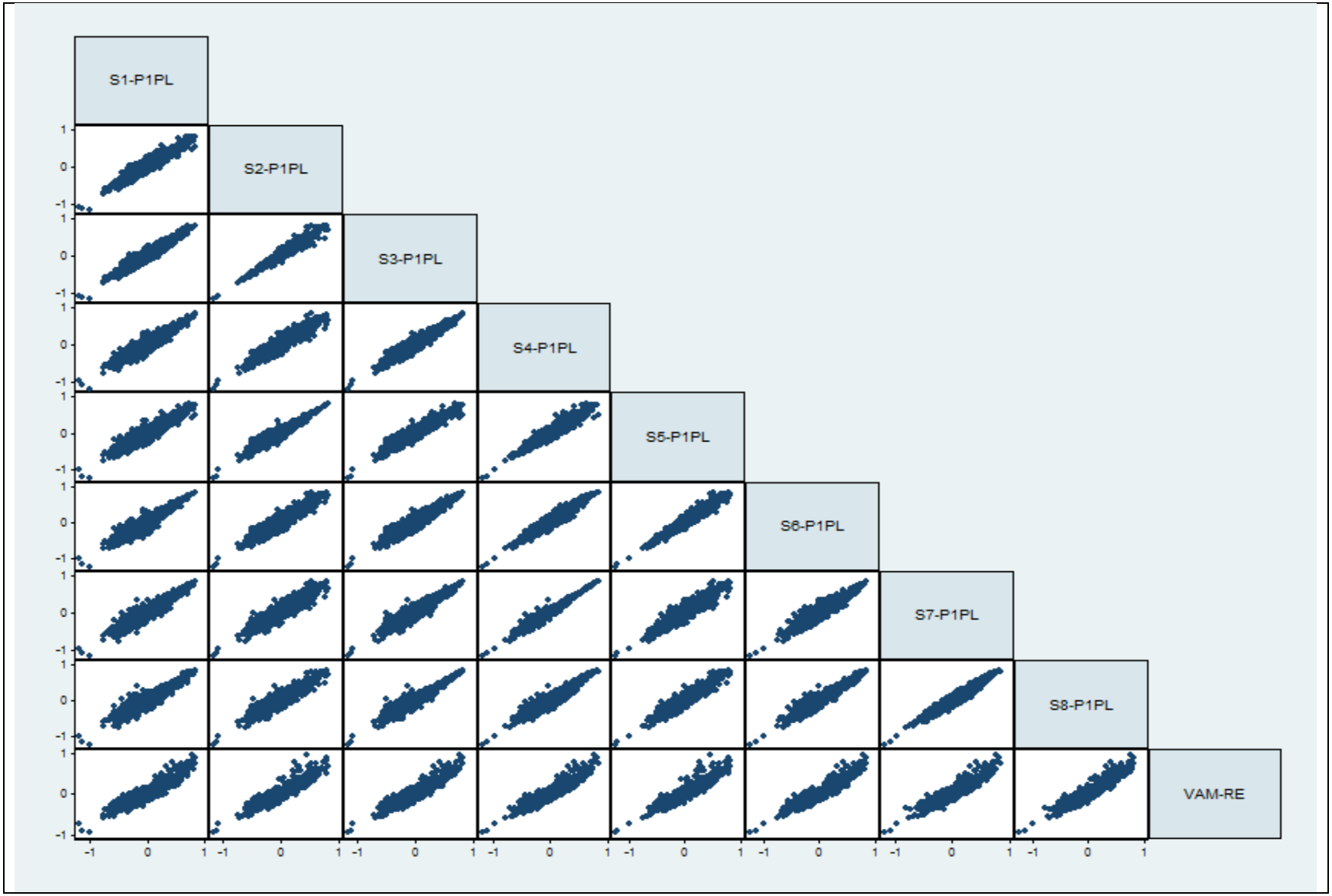


Figure A3. Scatterplots of the teacher effect estimates from different simulated cut-scores (Polytomous category)

BIBLIOGRAPHY

BIBLIOGRAPHY

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95–135.

Baker, E. L., Barton, P. E., Darling-Hammond, L. Haertel, E., Ladd, H. F., Linn, L. R., Ravitch, D., Rothstein, R., Shavelson, R. J., Shepard L.A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute.

Baker, D. P., Goesling, B., Letendre, G. K. (2002). Socioeconomic status, school quality, and national economic development : A cross-national analysis of the "Heyneman-Loxley Effect" on mathematics and science achievement. *Comparative Education Review*, *46*(3), 291–312.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Ballou, D. (2009) Test scaling and value-added measurement. *Educational Finance and Policy*, *4*(4), 351–384.

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology : student growth percentiles and percentile growth projections/trajectories. Paper presented at *The National Center for the Improvement of Educational Assessment*. Dover, New Hampshire.

Boardman, A. E., & Murnane, R. J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education*, *52*(2), 113–121.

Bonate, P. L. (2000). Difference scores. In *Analysis of Pretest-Posttest Designs* (Ch.3). FL: Chapman & Hall.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher Preparation and Student Achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440.

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.

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

Brophy, J. (2010). *Motivating students to learn (3rd Ed.)*. New York: Routledge.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics*, *66*(2), 103–115.

Burke, M. A., & Sass, T. R. (2008). *Classroom peer effects and student achievement.* National Center for Analysis of Longitudinal Data in Education Research Working Paper No. 18.

Card, D., Krueger, A., & Card, D. (1998). School resources and student outcomes. *The Annals of the American Academy of Political and Social Science*, *559*(1), 39–53.

Carnoy, M. (2003). *The new accountability: high schools and high-Stakes testing*. New York: Taylor & Francis.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: teacher value-added and student outcomes in adulthood. National Bureau of Economic Research Working Paper No. 17699. Cambridge, MA.

Chudgar, a., & Luschei, T. F. (2009). National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Educational Research Journal*, *46*(3), 626–658.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673–682.

Coleman, J. S. (1969). *Equality and achievement in education*. Boulder, CO: Westview Press.

Condie, S., Lefgren, L., & Sims, D. (2012). *Teacher heterogeneity, value added and education policy*. Rand Corporation Working Paper. Retrieved from http://www.rand.org/content/dam/rand/www/external/labor/seminars/adp/pdfs/2012/lefgren.pdf

Corcoran, S., Jennings, J., & Beveridge, A. (2011). *Teacher effectiveness on high-and low-stakes tests*. New York University Working Paper. Retrieved from https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf

Croninger, R. G., Rice, J. K., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review*, *26*(3), 312–324.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

Dee, T. S. (2003). *Are there civic returns to education?* National Bureau of Economic Research Working Paper No. 9588. Cambridge, MA.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? assessing the impact of unobservables on educational productivity. *The Journal of Human Resources*, *32*(3), 505–523.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school certification status and student achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129–145.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research*, *66*(3), 361-396

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (In press). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*.

Hanushek, E. (1979). Conceptual and empirical issues in the estimation of education production functions. *Journal of human Resources*, *14*(3), 351-388.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, *18*(4), 45–51.

Hanushek, E. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, *100*(1), 84–117.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? an examination of the statistical properties and policy alternatives. *Education Finance and Policy*, *4*(4), 319–350.

Harris, D. N. (2011). *Value-Added Measures in Education*. Cambridge MA: Harvard Education Press.

Harris, D. N., & McCaffrey, D. F. (2010). Value-added: assessing teachers' contributions to student achievement. In M. M. Kennedy (Ed.), *Teacher Assessment and the Quest for Teacher Quality: a Handbook* (1st ed., pp. 251–282). San Francisco: Jossey-Bass.

Harris, D., Sass, T., & Semykina, A. (2010). Value-added models and the measurement of teacher productivity. National Center for Analysis of Longitudinal Data in Education Research Working Paper No. 54.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hulleman, C. S., Durik, A. M., Schweigert, S. A., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology, 100*, 398-416.

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, *16*(9), 965–80.

Jacob, B. A. (2007). Test-based accountability and student achievement: an investigation of differential performance on NAEP and state assessments. National Bureau of Economic Research Working Paper No. 12817. Cambridge MA.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17–64). American Council on Education & Praeger.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631.

Koedel, C., & Betts, J. (2009). Value-Added to what? How a ceiling in the testing instrument influences value-added estimation. National Bureau of Economic Research Working Paper No. 14778. Cambridge MA.

Konstantopoulos, S., & Chung, V. (2011). The persistence of teacher effects in elementary grades. *American Educational Research Journal*, *48*(2), 361-386.

Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teacher College Record*, *116*(1).

Koretz, D., and Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), Educational measurement (4th ed.), 531-578. Westport, CT: American Council on Education/Praeger.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: a validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, *25*(3), 287–298.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*(6), 3–16.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Martineau, J. A. (2006). Distorting value added: the use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35–62.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. a, & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). *The inter-temporal variability of teacher effect estimates* (No. 2009-03). Nashville: National Center on Performance Incentives.

Meyer, J. W. (1970). High school effects on college intentions. *American Journal of Sociology*, *76*(1), 59–70.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*(3), 283–301.

Murnane, R. J., Maynard, R. A., & Ohls, J. C. (1981). Home resources and children's achievement. *The Review of Economics and Statistics*, *63*(3), 369–377.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23).

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237-257.

O'Day, J. A. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, *72*(3), 293–330.

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: the importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis*, *30*(2), 111–140.

Papay, J. P. (2010). Different tests, different answers: the stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163–193.

Rabe-hesketh, S., & Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata Volume II: Categorical Responses, Counts, and Survival* (Third Edit.). Texas: Stata Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and data analysis methods* (Second Eds.). Thousand Oaks: Sage publications.

Raykov, T., & Marcoulides G.A. (2011). *Introduction to Psychometric Theory*. New York, NY: Routledge.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519.

Reckase M.D. (2012). The evaluation of teachers and schools using the educator response function (ERF). Paper presented at the 12th Annual Maryland Assessment Conference. Value Added modeling and Growth Modeling with Particular Application to Teacher and School Effectiveness. October 2012. Maryland, MD.

Reckase, M. D., & Li, T. (2007). Estimating gain in achievement when content specifications change: a multidimensional item response theory approach. In L. R. W. (Ed.), *Assessing and modeling cognitive development in school*. Maple Grove, MN: JAM Press.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2).

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214.

Rubin, D., Stuart, E., & Zanutto, E. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *34*(4).

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*(3), 299–311.

Smith, M. S., O'Day, J. A., & Cohen, D. K. (1991). National curriculum, American style: Can it be done? What might it look like? *American Educator*, *14*(4), 10–17, 40–47.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of Research. *Review of Educational Research*, *75*(3), 417–453.

Spady, W. G. (1994). *Outcome-based education: critical issues and answers*. VA: Arlington: American Association of School Administrators.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*, F3–F33.

Wright, B. D., & Stone, M. H. (2004). *Making measures*. Phaneron Press.

Zheng, X., & Rabe-Hesketh, S. (2007). Estimating parameters of dichotomous and ordinal item response models with gllamm. *The Stata Journal*, *7*(3), 313–333.