US-China Education Review A, January 2016, Vol. 6, No. 1, 1-19
doi:10.17265/2161-623X/2016.01.001

The Effectiveness of Tying Teacher Evaluation Policy to Student Achievement in South Korea

Sung Tae Jang
University of Minnesota, Minneapolis, USA

Experts in the field of educational policy have identified accountability as an important educational issue. The most critical debate related to educational accountability has examined the various standards used to evaluate teachers' accountability in the United States (U.S.), specifically in terms of whether student achievement should be included when evaluating teacher accountability. Similar to the U.S., in South Korea, various governmental efforts have sought to diversify and solidify standards for teacher evaluation. Debate might emerge as to whether or not student performance should be included in teacher evaluation in South Korea in order to emphasize stronger accountability. Thus, this paper examines how the tie between teacher evaluation and student achievement in South Korea relates to growth rates in individual student achievement. To this end, this study conducts a micro-level analysis focusing on the effects of individual schools' teacher evaluation policy on individual students' achievement. The analysis uses 2,655 teachers in 150 middle schools as well as data on 5,677 middle school students' math and reading test scores from the Korean Educational Longitudinal Study (KELS), implemented by the Korean Educational Development Institute (KEDI). This study provides insights into the implications of teacher evaluation policies in South Korea by highlighting the effectiveness of tying such a policy to changes in students' achievement. This examination speaks to the timeliness of research in terms of including students' achievement in the teacher evaluation policy, which is also closely related to other policy changes, such as performance-pay systems in school organizations.

Keywords: teacher evaluation policy, teacher accountability, policy effects, longitudinal analysis

Introduction

Freedom and accountability are intertwined values in a democratic society. Based on this concept, in recent educational reform initiatives, governments have been eager to provide more freedom to each school to develop their own curricula that reflect the needs of school communities. For example, School Autonomy Policy, a critical movement in educational reform in South Korea, has emerged to improve schools' organizational effectiveness by providing K-12 schools with more autonomy. At the same time, the government has implemented an accountability policy requiring that each school and teacher demonstrate that their students achieve observable educational outcomes. One way in which the government is carrying out its policy is by implementing incentive systems and a nationwide test.

Between the values of freedom and accountability, the latter has been discussed as an educational issue throughout the educational administration field. The most critical debate related to teacher accountability has

Sung Tae Jang, Ph.D. candidate, graduate teaching assistant, Department of Organizational Leadership, Policy, and Development, University of Minnesota.

examined the various standards used to evaluate teachers' accountability throughout the United States (U.S.), especially in terms of whether student achievement should be included when evaluating teacher accountability. The National Council on Teacher Quality (2012) found that 30 states include student achievement in their teacher accountability policy while more than 40 states have sought to change their laws related to education policy.

Similar to the U.S., in South Korea, various governmental efforts have sought to diversify and solidify standards for teacher accountability. South Korean President Lee's government announced 100 major projects, including establishing a support system to increase students' achievement and close the achievement gap by 2009. The projects also included stabilizing a national standardized evaluation system of student accomplishment. Since 2008, this test has targeted every student in 3rd and 6th grades in primary school, 3rd grade in middle school, and 1st grade of high school. In addition, to address the special enforcement related to accessing information from an educational organization, each school has to disclose the test results in order to improve accountability. Similar to teachers' strikes against nationwide testing, the teachers, teachers' union, and some parents have come out against this test, and some teachers have even refused to administer the test (Jeong, Shin, & Lee, 2009).

Although South Korea does not require schools to include student achievement in their teacher evaluation policy, some schools are currently using student achievement as a standard for evaluating teacher performance by providing each school more autonomy from the government. Thus, debate might emerge as to whether or not student performance should be included in teacher evaluation policy in South Korea in order to emphasize stronger accountability. In addition, many disagreements have emerged in terms of how to evaluate teacher accountability as well as the validity and reliability of adopting a teacher accountability policy using student achievement in South Korea.

Scholars are concerned about this issue because changing the teacher evaluation policy is closely related to other policy changes, such as performance-pay systems in school organizations. In addition, the policy might affect the degree of teacher motivation or satisfaction, which can affect educational outcomes. Scholars have carefully examined this issue because a change in the evaluation policy can also influence teachers' collective actions, which might have negative educational impacts on school organizations (e.g., the loss of classes).

Although some schools use student achievement to evaluate teachers' capability in South Korea, a sufficient discussion about the effectiveness of teacher evaluation tied to student achievement is lacking in South Korea. Some scholars might criticize that this examination is too early to conduct nowadays as tying teacher evaluation to student achievement is not a step in South Korea's policy formulation or implementation. However, as DiMaggio and Powell's (1983) isomorphism theory implies, educational organizations sometimes use benchmark models after a successful case of leading an organization in the same nation as well as overseas. Thus, a test-based evaluation system might be used as a policy in South Korea following the American evaluation policy (Jeong et al., 2009). In addition, as the Korean Educational Longitudinal Study's (KELS) data indicate, some schools are already using student achievement in evaluating teachers' ability; therefore, thoroughly examining the effectiveness of test-based accountability is necessary to provide further related policy insights.

The purpose of this study is to examine the effectiveness of a teacher evaluation policy tied to student achievement in terms of student test scores in the South Korean context. This paper first briefly explores

evaluation policy-related literature in both the U.S. and South Korea from various theoretical perspectives. It then summarizes the relevant literature focusing on teacher evaluation policy tied to student achievement. The method section will describe the three-year longitudinal data and linear mixed-effects model (LMM). Then, the following section will discuss the association between teacher evaluation policy tied to student achievement and the growth patterns in students' performance.

Review of the Related Literature

Teacher Evaluation Policy

Educational accountability refers to the school or teacher's responsibility to achieve educational outcomes. Although schools can have different educational outcomes, such as direct service, human capital, or development of civic engagement (see Mitchell & Mitchell, 2003), the current teacher evaluation system in the U.S. is based on measuring students' test performance. Unlike the high-stakes evaluation policy in the U.S., South Korea's evaluation policy uses low-stakes testing with a less powerful rewards and sanctions system than the U.S..

U.S.. In the U.S., the No Child Left Behind Act's (NCLB) policy requires strong accountability while providing autonomy in each school. The policy focuses on increasing accountability, providing more freedom, utilizing proven methods, and providing more school choices for parents (U.S. Department of Education, 2004). In particular, the NCLB includes seven principles (U.S. Department of Education, 2006):

1. Ensure that all students are proficient by 2014 and set annual goals to ensure that the achievement gap is closing for all groups of students;

2. Set expectations for annual achievement based on meeting grade-level proficiency, not on students' background or the school's characteristics;

3. Hold schools accountable for students' achievement in reading/language arts and mathematics;

4. Ensure that all students in tested grades are included in the assessment and accountability system, hold schools and districts accountable for the performance of each student subgroup, and include all schools and districts;

5. Include assessments in each grade, 3rd through 8th, and in high school for both reading/language arts and mathematics and ensure that they have been operational for more than one year and receive approval through the NCLB peer-review process for the 2005-2006 school year; the assessment system must also produce comparable results from grade to grade and year to year;

6. Track student progress as part of the state data system;

7. Include student participation rates and student achievement on a separate academic indicator in the state accountability system.

Since the 1980s, standard-based reform has required more accountability among schools. Students' achievement tests have been utilized as a critical factor for evaluating educational accountability (Elmore, Abelmann, & Fuhrman, 1996; Hanushek & Raymond, 2005). Recently, the value-added model (VAM), which factors students' improvement into educational accountability, has also been discussed in the U.S.. The VAM seeks to isolate the contribution of individual teachers to student learning in a particular subject in a particular year by using statistical methods. However, both supporters and detractors of this approach recognize that the method might not be very reliable. Teachers ranked at the top in any one year might find themselves at the bottom in the subsequent year (e.g., Loeb & Candelaria, 2013). In addition, Goldhaber and Theobald (2012)

found that the biggest difference between value-added estimates comes between models that ignore possible school effects and models that explicitly recognize them. The importance of school contexts speaks to the value of building capacity within an organization, not simply focusing on individual teachers as single entities. In other words, inducements alone might not be sufficient for creating the infrastructure and capacity necessary to improve student achievement. The VAM is still contested in the American context, and the model's reliability is continuously being examined.

South Korea. Unlike the teacher evaluation policy in the U.S., South Korean teacher evaluation policy does not yet fully include student performance. Since the 1990s, South Korea has emphasized accountability related to outputs and outcomes throughout society. Schools have to be diversified and have more autonomy, and they should continually increase students and parents' choice of schools while formulating educational policy. On March 20, 2008, the Ministry of Education, Science and Technology (MEST) announced School Autonomy Policy, which aimed to improve school organizations' effectiveness by providing K-12 schools with more autonomy and simultaneously requiring more accountability (Jang, 2011). The MEST further expanded the nationwide student standardized achievement test and reinforced the commitment to providing open information about the test results. Thus, the South Korean government has attempted to promote increases in school autonomy and accountability through such efforts (Jeong et al., 2009).

As a second effort to increase accountability of teachers, a teacher evaluation system to promote professional development (TESPD) was developed in South Korea in 2010. The object of the TESPD is all teachers who teach primary, middle, and high school, including national, public, and private schools. The evaluation items include teaching and teachers' guidance of students as well as overall school management of principals and vice principals. The detailed items for evaluating accountability include class preparation, teaching, evaluation and feedback, guidance of individual students both inside and outside the school, social life guidance, school supervision, and educational planning (Lee, 2010).

To evaluate these items, students' evaluation, parents' evaluation, and peer evaluation are utilized. Students and parents evaluate their degree of satisfaction with the teacher and express their opinion to give feedback to each teacher. In terms of peer evaluations, colleges evaluate teachers through the usual observations and classroom performance. To manage the TESPD, each school and educational district office has a committee to implement the TESPD objectively and appropriately. Teachers submit a self-capability development plan after receiving the results from colleges, parents, and students. Schools, districts, and the government also establish and provide plans for supporting the TESPD. Although teacher accountability in South Korea includes capability building components, teacher accountability is still driven by external approaches in that the school assessment policy and teacher evaluation policy are promoted by the government officials and standardized indicators of student achievement (Song, 2013).

Several studies have examined the impact of evaluation policy on teachers and school organizations from different perspectives. These studies illuminate the impact (e.g., teacher motivation) and issues (e.g., power structure and achievement gap) related to teacher evaluation policy.

Finnigan and Gross (2007) conducted an interpretivist study related to teacher evaluation policy, focusing on whether motivation levels changed as a result of the evaluation policy and policy mechanisms. The authors drew on expectancy and incentive theories, which define "a person's belief about the likelihood that his or her efforts will result in the desired outcome" (p. 596) and "the external policy mechanisms that establish a penalty or reward for desired behavior or results" (p. 596), respectively. In particular, the authors explored three types

of incentives to which teachers might respond: solidary, purposive, and material. Based on these theories, the authors analyzed teacher motivation in low-performing schools in Chicago.

This study concluded that teachers were conflicted in their feelings toward the policy itself. Both the quantitative and qualitative data indicated that, although the teachers understood the need for accountability and standards, they felt that the means by which the policy was carried out did not respect the challenges they faced in many cases. The interview data also revealed a measure of frustration as teachers expressed doubt about their students' ability to meet the school's academic goals. In addition, the analysis of teachers' time spent on instructional activities, professional development, and preparation for instructional activities highlighted that teachers increased efforts across these areas after their schools were placed on probation. The case study ultimately revealed that teachers in these schools responded to all three types of incentives; the fear of job loss was an exceptionally serious concern for just over one third of teachers.

Herr and Arms (2004) examined the impact of teacher evaluation policy on a single-sex academy (SSA) by exploring how accountability measures affected the implementation of what was widely touted as being primarily a gender-based reform for students at one California middle school. The impact of high-stakes testing on curriculum and the roles of teachers was addressed as well as how the multiple, simultaneously implemented reforms derailed the possibilities inherent in the SSA.

The study revealed that teachers expressed concern with the intrusion of testing on curriculum and the teaching of larger concepts; they felt that the emphasis had become skill-and-drill oriented in terms of the impact of high-stakes testing. In addition, as teachers tried to implement the mandated test preparation and packaged reading program, authentic teaching got lost along the way. Thus, the authors concluded that it "was a school on the move with the rising test scores to prove it, while the lived experience in classrooms was one of a narrowed, reduced curriculum, less meaningful lessons and learning, and less authentic teaching" (Herr & Arms, 2004, p. 550).

Teacher Evaluation Policy Tied to Student Performance (TEPSP)

The second strand of relevant literature focuses on TEPSP. Some studies identified the positive influence of TEPSP on average student achievement (Carnoy & Lobe, 2002; Hanushek & Raymond, 2005; Jacob, 2005).

Specifically, Carnoy and Lobe (2002) examined factors influencing a strong evaluation policy and whether stronger statewide evaluation policy improves student outcomes. The authors analyzed each state's evaluation policy using a 0-5 index of strength based on the use of high-stakes testing to sanction and incentivize schools. As dependent variables, the authors measured average student performance at the state level based on tests, such as the National Association of Educational Progress (NAEP) math test scores, 9th grade retention rates, and high school survival tests. The authors employed recursive models to examine the impact of evaluation policy on student outcomes. Their regression models included student achievement as a dependent variable, testing whether the percentage of 8th graders or 4th graders achieving at the basic skills level or better increased more in states with stronger accountability policy between 1996 and 2000. In addition, the authors included 9th grade retention rates in states and 10th-12th grade survival rates as dependent variables, running these models to verify the number of specifications. They used data sets from the NAEP and National Center of Educational Statistics (NCES) for math results, data from state departments of education webpages for accountability policy details, and NCES data for retention rates and high school completion rates. To measure the strength of accountability policies in each state, data from the Consortium for Policy Research in Education

(CPRE) were used. The results of their study indicated that a positive and significant relationship exists between the strength accountability policy and math achievement gains. This positive relationship exists for all Black, White, and Hispanic students. In addition, the higher the rate of minority students, the stronger the accountability policy is. However, the data did not indicate a relationship between accountability and 9th grade retention rates and high school completion rates. Thus, the authors provided evidence that stronger pressure on schools or districts during high-stakes testing leads to larger gains in student performance in NAEP math scores.

Hanushek and Raymond (2005) conducted an additional study related to evaluation policy, focusing not only on student achievement, but also on achievement gaps among Black, Hispanic, and White students. The authors sought to isolate the effects of teacher evaluation policy on performance, such as differences in circumstances and policies of each state and time-varying inputs (e.g., parental education and school spending). They also categorized the state results for Black, Hispanic, and White students to identify gaps among ethnicities. The authors measured student performance in 37 states according to the average score on NAEP. In addition, evaluation policies in each state were classified by "whether or not they both report results and attach consequences to school performance (i.e., consequential states) or simply stop at providing a public report (i.e., report card states)" (Hanushek & Raymond, 2005, p. 306). Similar to Carnoy and Lobe's (2002) study, Hanushek and Raymond (2005) found that the introduction of a consequential evaluation policy had a positive impact on NAEP performance. However, the report card states' evaluation policy did not have a significant influence on performance. In addition, both Blacks and Hispanics achieved smaller gains relative to Whites. Thus, the evaluation policy did not uniformly close the achievement gaps among ethnicities.

In summary, several studies examining the effectiveness of TEPSP have shown improvements in students' achievement scores on high-stakes tests. However, these previous studies are limited in terms of using state-level data, which might cause aggregation bias and cannot reveal the effect of teacher accountability on individual students' achievement. In addition, this limitation might lead to methodological issues in terms of missing values, thereby, causing unreliable results, although this was not mentioned in any of the studies. In order to address these limitations, this study analyzes individual students' data to highlight the effectiveness of TEPSP at the student level and uses a more accurate statistical model to address the methodological issue. In particular, this study examines the effectiveness of teacher evaluation tied to student achievement in terms of the individual students' development or change on standardized test scores.

To achieve a better understanding of the effectiveness of teacher evaluation tied to student achievement, this paper establishes the following four research questions:

1. To what extent does variability exist in individual students' achievement scores in terms of intercept (starting point) and slope (change rate)?

2. Which student-level variables can explain the variability in the intercept and slope of individual students' achievement score?

3. Which school-level variables can explain the variability in the intercept and slope of individual students' achievement score?

4. What is the relationship between TEPSP and the changes in individual achievement scores? (i.e., To what extent can TEPSP explain variability in the intercept and slope of individual students' achievement scores?)

To examine these research questions, the next part describes the data and the model-selection procedure to establish a base model for capturing the variability of individual scores. A covariance test is subsequently employed to examine the variables and explain the variability in the intercept and slope of student achievement

scores, which will suggest the best model for explaining and identifying the effectiveness of teacher evaluation tied to student achievement in the South Korean context. Finally, based on the identified best model, this paper examines the longitudinal effect of test-based evaluation policy.

Method

Data and Research Design

This study used data from the KELS implemented by the Korean Educational Development Institute (KEDI). KELS started in 2005 with 6,908 1st grade student samples from 150 middle schools[1] in South Korea. As an ongoing study, it has tracked these students at one-year intervals to investigate learning and educational activities experienced in their families, schools, and social lives as well as their cognitive and non-cognitive development.

Because the data have a hierarchical structure with students grouped in schools, this paper deals with two levels of grouping in the data. With the addition of time, the data expand to a three-level data hierarchy: (a) within students varying across time; (b) between students; and (c) between schools. The dependent variable, y_{sij}, denotes the value of Korean scores[2] for i student (i = 1, …, 5,677) in s school (s = 1, …, 124) at time j (j = 0, 1, 2, corresponding to values of first, second, or third year of middle school, respectively). As middle school students in South Korea have different teachers each year, which prevents an interdependency issue, this study uses two levels of grouping (student- and school- level). The students' Korean language performance was analyzed using a LMM with three time points. LMM for longitudinal data is a widely used method in social sciences, biostatistics, economics, and education. Although analysis of variance (ANOVA) has also been widely used in the education field, the strengths of LMM make it more attractive to advanced researchers when analyzing data. In other words, LMM allows for missing data and various options for the variance-covariance matrix of random effects (Ryoo, 2011).

This study used a random effect because individual test scores have a unique variability within themselves. TEPSP is considered to be a static binary predictor, where 0 indicates teacher evaluation not including student performance and 1 indicates teacher evaluation tied to student performance. As this is a non-experimental design, this study cannot report casual effects. However, it can investigate the effectiveness of TEPSP on students' performance through model comparison (Ryoo & Hong, 2010). The reasons for using LMM will be specifically revisited later in this paper (see Model Selection).

Furthermore, the variance components including variances-covariances of random effects that contain potentially useful information regarding individual differences are estimated using maximum likelihood (ML). ML is the method by which parameter estimation is tied to a particular distribution in order to find the distribution that best matches the data. This study used ML estimation instead because the ML method provides the most satisfactory approach for obtaining estimates. In addition, it enables us to obtain an unbiased estimation. Finally, this study has a large enough sample; the restricted ML is used with small samples.

In this study, the model was assumed to have an independent error structure. As the measurement points of testing are far apart (i.e., a year), students' scores will change, meaning that residuals will be uncorrelated with

[1] School systems in South Korea are composed of six years in primary school, three years in middle school, and three years in high school.

[2] The Korean test examines language capability, including reading, grammar, listening, speaking, and writing in South Korea, similar to a language test in the U.S..

each other. This situation would be hard to explain using other error structures, such as auto-regressive or compound symmetry, in the field of social sciences, such as education.

The decision to select a model within a nested model should be based on the likelihood ratio test (LRT) between the full model that includes the random effect(s) and the reduced model excluding the random effect(s) (Anderson, 2008). This study used a step-up approach to examine models to best capture the effect of TEPSP among the model-building approach (e.g., step-up, top-down, subset, and inside out) as it is "a common practice that applied researchers seek for the best fitting model starting from the simplest model, such as random-intercepts model, proceeding to more complex models until the selected model is not significantly different from the more complex model" (Ryoo, 2011, p. 28). The Linear Mixed-Effects Models using "Eigen" and S4 (LME4) package (Bates, Maechler, & Bolker, 2012) in R 2.15.1 (R Core Team, 2012) was used to conduct analyses in this study.

Sampling

The sample for KELS was gathered from 703,914 1st grade students in 2,929 middle schools throughout South Korea with the exception of physical education focused middle schools and branch schools. The sampled students' parents constituted the sampled parents, and their teachers and schools constituted the sampled teachers and schools. A stratified cluster random sampling method was used to acquire the sample. First, the nation was classified with strata according to the regional scale, and sample schools as clusters were extracted from the strata. Finally, the sampled students were extracted from those sample schools. The regional scales include four categories: Seoul (capital city), five major metropolitan areas, urban areas, and rural areas. The number of schools in each regional category was determined using a proportionate stratified sampling method based on the proportion of the number of students in each category. Fifty students were sampled from each of the middle schools. A random sampling method was used to sample the schools from each stratum and students from each school. Table 1 shows the population and sampling size of 1st grade middle school students according to the regional scale in 2005.

Table 1

Population and Sample Size of Schools, Students, and Teachers

Regional scale	Schools			1st grade middle school students		
	Population	Sample	Percentage (%)	Population	Sample	Percentage (%)
Seoul	362	26	7.18	130,012	1,237	0.95
Metropolitan/major cities	598	38	6.35	197,120	1,939	1.03
Urban areas	1,233	45	3.65	330,000	1,851	0.77
Rural areas	736	41	5.57	46,782	1,881	1.34
Sum	2,929	150	5.12	574,169	6,908	0.98

As Table 1 shows, 150 schools were sampled: 26 schools in Seoul, 38 schools in the five major metropolitan cities, 45 schools in urban areas, and 41 schools in rural areas. In total, 6,908 1st grade middle school students were sampled. In addition, three years of achievement data from the 1st to the 3rd grades of middle school were used to examine the effectiveness of TEPSP. As middle school students in South Korea go to different high schools, each student has teachers held to different teacher evaluation policies, making it challenging to capture the continuous effect of teacher evaluation tied to student achievement in South Korean contexts. Thus, this study focused on analysis in the middle school context.

Variables and Descriptive Statistics

In order to identify the base model for examining the effectiveness of teacher evaluation tied to student performance, this study used student- and school- level covariates. After identifying the base model to capture the variability of the intercept and slope of students' Korean test score, TEPSP and students' Korean scores were added to the base model, with the latter as a dependent variable. The following subsections provide specific explanations about each covariate, focal predictor, and dependent variable.

Student-level covariate. As several educational studies have demonstrated, family socioeconomic status (SES) has a critical effect on a student's achievement (Batool, Naureen, & Kanwa, 2010; Dubow, Boxer, & Huesmann, 2009; Sirin, 2005; White, 1982). Thus, educational research has established SES as a control variable to examine the effect of other focal variables. There are different ways to measure SES, such as family income and parents' education levels. Although family income was available to use in the current study, it was not included because it can lead to many missing values as parents sometimes choose not to provide such information. In addition, mother's education level has been shown to be a critical variable for predicting students' achievement in school (Magnuson, 2007; Parveen & Alam, 2008); thus, it was used as a covariate at the student level for the current study. A dummy-coded covariate for mother's education level was used: less than a high school diploma was coded as 0, and served as a reference group; a bachelor's degree was coded as 1; master's and Ph.D. degrees were coded as 2.

School-level covariate. The region where a school is located is included as a school-level covariate. In the South Korean context, severe achievement gaps exist between cities and rural areas (Shin, 2006; Woo, 2011) as well as within the Seoul metropolitan area (Ha, 2005). Thus, a variable for indicating whether the school is located in an urban or rural area was included as a covariate. Schools in metropolitan areas or small and medium-sized cities were coded as 1 in the urban variable. Schools located in rural areas (except cities) were coded as 0 and served as a reference group. Other school-level covariates were also tested in the initial analysis, but they did not statistically affect student achievement (e.g., school size and per-pupil expenditure), and thus, were not included in this analysis.

Key predictor. TEPSP was the key predictor used. To determine whether a school uses student performance to evaluate teacher capability, the school questionnaire included question 16-2: Does your school use the results of students' performances to evaluate teachers' capabilities? Schools that responded "1 = Yes" were dummy-coded as "Including students' performance in teacher accountability (INCL) (= 1)"; schools that marked "2 = No" were coded as "Not including students' performance in teacher accountability (NOINCL) (= 0)." Students attending schools that did not respond or marked "3 = We do not know" were not included in this analysis. The 110 NOINCL schools (80.3%) included 4,988 students (72.2%); the 14 INCL schools (10.2%) included 689 students (17.8%). There were also 26 (9.5%) missing values, including schools marked as "We do not know"; thus, 1,231 (10.0%) students were treated as missing values. The treatment of missing values will be discussed later in this paper (see Missing Values). The sample sizes in terms of including student performance in teacher evaluation are shown in Table 2. As the main purpose of this study was to examine the effect of TEPSP, the examination included the TEPSP variable. The Korean scores of 5,677 students, excluding those with missing values, were used to examine the effect of TEPSP.

In order to identify the initial differences between students in schools with TEPSP and without TEPSP, this study compared the means of students' performance in the previous year (i.e., 6th grade in elementary

school) between two groups. An independent sample *t*-test showed that the difference in previous student performance between schools with TEPSP (N = 688; M = 5.83; SD = 2.31) and without TEPSP (N = 4,989; M = 5.67; SD = 2.27) was not statistically significant ($t_{(5,675)}$ = -1.783; p = 0.08, two-tailed). This result indicates that the initial conditions (i.e., students' performance) between the two groups were similar, which allows for analyses of the effectiveness of TEPSP.

Table 2

Number of INCL Schools vs. NOINCL Schools

	Schools		Students	
	Number	Percentage (%)	Number	Percentage (%)
Schools with TEPSP	14	10.2	689	17.8
Schools without TEPSP	110	80.3	4,988	72.2
Missing	26	9.5	1,231	10.0
Sum	150	100.0	6,908	100.0

Dependent variable. Students' Korean scores were the dependent variables in this study and were obtained for 1st through 3rd grade middle school students in KELS. KELS administered standardized tests to the sampled students. To develop common criteria and examine the relationship among the test scores in each grade, vertical scaling was used in the KELS test instead of raw scores. In addition, KELS developed vertical scaling using Item Response Theory (IRT), which has been shown to be more reliable and stable than classical test theory (CTT) (KEDI, 2009). Specifically, the test scores are vertical scaling scores with a mean of 300 points in the 1st grade and a standard deviation (*SD*) of 50; they increase 100 points on average annually. In terms of using vertical scaling scores in KELS, Lee, Im, Park, and Kim (2010) concluded that the factors used would not threaten the validity of the vertical scale of KELS. These include the relatively small sample size for applying item response models and the related instability issue of item parameter estimates, the problem of adding manipulated growth information (e.g., 100-point growth per grade) onto the vertical scale, the impact of the use of differentially functioning items as common items for vertical scaling, and the implementation of a number correct-to-scale score conversion table instead of pattern scoring. Thus, the authors recommended maintaining the level of score variation within grades studied in 2005 to support the comparison of student achievement growth over several years after 2005.

Students' grades in middle school were coded as a time variable; for example, the first time point was the Korean score in the 1st grade of middle school. Means (with *SD* in parentheses) for Korean scores in 1st through 3rd grades were 299.8 (57.6), 393.21 (65.37), and 499.69 (61.72), respectively, as shown in Table 3.

Table 3

Descriptive Statistics of Korean Scores in Each Grade

	1st grade	2nd grade	3rd grade
N	6,751	6,438	6,283
Mean	299.8	393.21	499.69
SD	57.6	65.37	61.72

The mean difference in Korean scores between 1st and 2nd grades was 93.41—that is, the mean increased by about 100. In addition, the mean difference in Korean scores between 2nd and 3rd grades was 106.48—that is, the mean increased by more than 100.

Missing Values

Missing values are a critical issue that can decrease the sample size, meaning the sample cannot properly represent the population. Fitzmaurice, Laird, and Ware (2004) highlighted three issues related to missing observations in longitudinal studies. First, when the data are missing, the data set becomes unbalanced over time. Second, a loss of information and a reduction in precision occur. Third, certain assumptions about the reasons for any missing information, called the missing data mechanism—such as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR)—are required.

The missing data in this study were categorized as MAR, indicating that "responses are missing depending on the set of observed responses, but are unrelated to the specific missing values that, in principle, should have been obtained" (Fitzmaurice et al., 2004, p. 95). In other words, the data can be considered MAR if the missingness does not depend on the value of X_i after controlling for another variable (Howell, 2007). In this study, the missingness of response variables was caused by transferring to another school, making it impossible to measure the test scores and random absences on the test day (KEDI, 2009); thus, the missing data can be categorized as MAR. As Howell (2007) indicated, the situation in which the data are at least MAR is sometimes referred to as ignorable missingness as we can still produce unbiased parameter estimates without needing to provide a model to explain missingness. Thus, this missingness does not affect the significantly different results in further analysis.

Model Selection

To capture the variability among individual students' 1st grade Korean scores and the individual change in scores from 1st to 3rd grades, this study employed a base model with a fixed intercept and random intercept. In this study, the fixed intercept indicates the mean score of the 1st grade whereas the fixed slope indicates the change rate of mean Korean scores from the 1st to 3rd grades. In addition, the random intercept indicates the variability of 1st grade Korean scores from the mean whereas the random slope indicates the variability of the change rate between participants. Finally, the error means residual indicates variability within participants.

As Figure 1 demonstrates, the 50 randomly sampled students' Korean scores among 5,677 sample students highlighted the variability in intercept scores among participants, clearly indicating the characteristics of students' Korean scores. Thus, this study first employed the base model that included fixed intercept, random intercept, and error.

Using the base model, this study selected the best-fitting model to explain the individual score variability by using the LRT. The LRT is the way of building models with null hypothesis significance testing (NHST). NHST requires the statement of a null hypothesis prior to the data analysis, and is used to determine which predictors should be included in the model. Specially, this paper uses the step-up method that begins with a simple model and NHST is used to see if the model can be made more complex by adding predictors. Thus, the best-fitting model was selected by using the LRT after adding the fixed slope and random slope, in order. NHST uses the log-likelihood test because the testing models are nested across the steps in NHST rather than using the Akaike information criterion (AIC) or Bayesian information criterion (BIC).

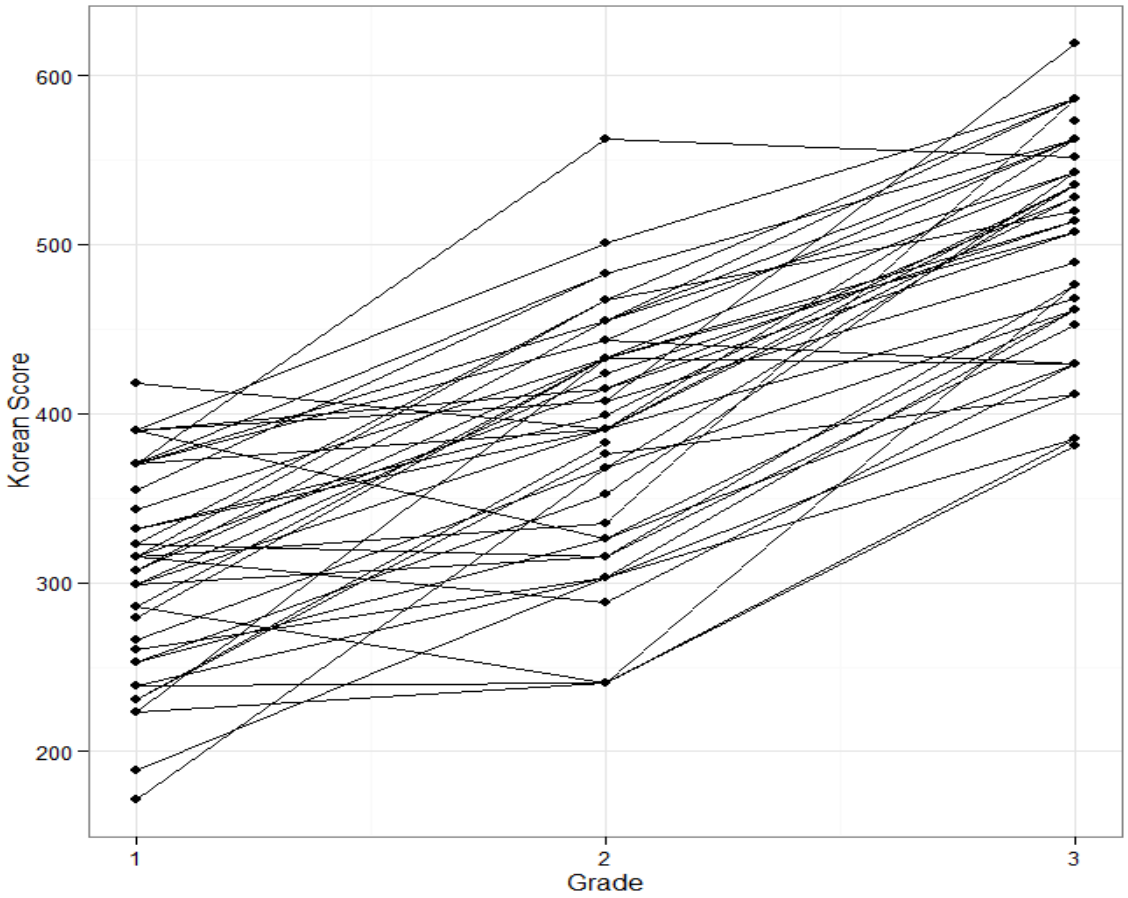

Figure 1. Fifty randomly sampled students' Korean scores.

Results

Table 4 summarizes the model comparison steps and selected base model including Chi-square estimates of the LRT. This result suggests that the final base model including school effects is Model 8, which includes both the fixed and random slopes in both student and school level, mother's education as a student-level covariate, schools' urbanicity as a school-level covariate, and the interaction between schools' urbanicity and grade. Thus, Model 8 was selected to examine the effect of TEPSP.

Table 4

Model Comparison Using the LRT for Baseline Model in School Level

Null hypothesis	Model comparison	Added predictor	Chi-square	*Df*	Selected base model
H_{01}	Model 1 *vs.* Model 2	Fixed slope effects	153.42***	1	Model 2
H_{02}	Model 2 *vs*. Model 3	Random slope effects	13.86***	2	Model 3
H_{03}	Model 3 *vs*. Model 4	Student covariate (mother education)	203.82***	1	Model 4
H_{04}	Model 4 *vs*. Model 5	School-level random intercept effect	285.91***	1	Model 5
H_{05}	Model 5 *vs*. Model 6	School-level random slope effect	178.06***	2	Model 6
H_{06}	Model 6 *vs*. Model 7	School covariate (urbanicity)	5.64*	1	Model 7
H_{07}	Model 7 *vs*. Model 8	Interaction between urbanicity and grade	9.48**	1	Model 8

Notes. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table 5 shows the parameter estimates and variance components of base model (i.e., Model 8). The fixed effect of estimated intercept is 168.97, illustrating the predicted mean Korean score in the 1st grade. The fixed effect of estimated linear slope is 108.24, indicating the mean increase in Korean score for each grade. In addition, the fixed effect of mother's education is 21.24, indicating, on average, a student who has a mother with a bachelor's or above degree achieves 21.24 higher on the Korean score than a student whose mother is without a bachelor's degree. Similary, the fixed effect of schools' urbanicity (28.52) shows, on average, students in urban areas tend to score 28.52 points higher than students in rural areas. Finally, the interaction between schools' urbanicity and grades shows the different score growth between urban schools and rural schools. In particular, the amount by which Korean scores are expected to change per grade for urban schools is 99.65, which is less than 8.59, while score growth for rural schools is 108.24.

Table 5

Parameter Estimates and Variance Components of Model 8

		Parameter	Estimates	Standard error	*t* value
Fixed effects	Intercept	β_0	168.97	6.85	24.68
	Grade	β_1	108.24	2.60	41.61
	Mother's education	β_2	21.24	1.82	11.66
	Schools' urbanicity	β_3	28.52	7.28	3.91
	Schools' urbanicity × Grade	β_4	-8.59	2.77	-3.11
Variance components	Var (b_{0si})	ϕ_{0si}	1,472.43	-	-
	Var (b_{1si})	ϕ_{1si}	2.94	-	-
	Var (b_{0s})	ϕ_{0s}	488.83	-	-
	Var (b_{1s})	ϕ_{1s}	68.60	-	-
	Var (e_{sij})	σ^2	1,496.33	-	-

The estimated variance of the random student-level intercept (1,472.43) shows the estimated variability around fixed intercept across all students. In addition, the estimated variance of random student-level slope (2.94) is the estimated variability around fixed slope across all students. Finally, the estimated school-level variance of random intercept (488.83) and slope (68.60) indicates the variability around fixed intercept and slope across all schools.

Analysis of the Effectiveness of TEPSP

Effectiveness of TEPSP. To examine the effect of TEPSP on individual students' Korean scores, TEPSP was added to the baseline model (Model 8) as a covariate. Thus, Model 9 was extended from Model 8 by adding a TEPSP covariate to hypothesize that individual variability in intercepts and slopes is caused by the TEPSP variable. As previously indicated, the TEPSP variable is a categorical variable: 0 = NOINCL and 1 = INCL. NOINCL was used as a reference group. Comparing Model 8 and Model 9 as well as Model 9 and Model 10, the effect of TEPSP is statistically significant, but TEPSP and grade interaction is not significant. Table 6 shows the results of the LRT.

Table 6

Model Comparison Using the LRT for TEPSP

Null hypothesis	Model comparison	Added predictor	Chi-square	*Df*	Selected model
H_{08}	Model 8 *vs*. Model 9	TEPSP	5.19*	1	Model 9
H_{09}	Model 9 *vs*. Model 10	Interaction between TEPSP and grades	1.37	1	Model 9

Model 9, including TEPSP, is selected to explain the variability in the change of Korean scores. In other words, a significant difference exists between NOINCL and INCL Korean intercept scores, but no interaction between TEPSP and grade.

As Table 7 shows, the slope of Grade (β_1) is 108.25, which indicates the amount by which Korean scores are expected to change per unit of grade for students in rural schools. Urban schools have a different change rate of 95.67 as the interaction (β_4) indicates. In addition, the estimate of mother's education (β_2) shows that for a student with a mother who has a bachelor's degree or above a bachelor's degree, the Korean score in the first grade of middle school is 21.26 points higher than that of a student who has a mother with an educational level below a bachelor's degree. The estimate of β_5 is -8.62, which indicates the difference in Korean scores between non-TEPSP and TEPSP in the 1st grade when other variables are held constant. However, the interaction between TEPSP and grade is not statistically significant, which indicates that there are no differences in score gain per grade between non-TEPSP and TEPSP. The 65.8% of variance in individual students' Korean scores is accounted for by variables in Model 9, which also enables us to assess how effective the model by R^2-type statistic is. Comparing Model 8 and Model 9, the two models differ in that TEPSP does not appear in Model 8, and Model 8 is nested within Model 9. The log-likelihood test shows that there is a statistically significant intercept effect of TEPSP in Model 9 by rejecting the reduced model (Model 8). Thus, Model 9 is retained as the final model in this study. As Figure 2 shows, there is a statistically significant difference in the predicted Korean scores of 1st grade between TEPSP and non-TEPSP, but there are no differences in the growth rates between TEPSP and non-TEPSP.

Table 7

Parameter Estimates and Variance Components of Model 9

		Parameter	Estimates	Standard error	*t* value
Fixed effects	Intercept	β_0	169.65	6.72	25.26
	Grade	β_1	108.25	2.60	41.62
	Mother education	β_2	21.26	1.82	11.68
	Schools' urbanicity	β_3	29.44	7.15	4.12
	Schools' urbanicity × Grade	β_4	-12.58	5.44	-2.31
	Teacher evaluation policy	β_5	-8.62	2.76	-3.12
Variance components	Var (b_{0si})	ϕ_{0si}	1,472.54	-	-
	Var (b_{1si})	ϕ_{1si}	2.94	-	-
	Var (b_{0s})	ϕ_{0s}	458.02	-	-
	Var (b_{1s})	ϕ_{1s}	68.58	-	-
	Var (e_{sij})	σ^2	1,496.33	-	-

Model assumption test. This subsection examines the assumptions for the mixed-effects model. In this study, Model 9 was based on the general mixed-effects model and assumed that a multivariate normal distribution of the random effects that indicate both random effects (e.g., random intercept, random slope, and error) should have a normal distribution. The assumption was tested by looking at a quantile-quantile (Q-Q) plot of the estimated individual intercepts and slopes. As the second assumption, the errors should be normally distributed with a mean of 0. This assumption was also tested using a histogram of errors, and the normal distribution of error was also plausible. Finally, the errors should be independent of each other, and there was no systematical pattern in the plot residuals against predictor (time variable). Thus, the three assumptions for random effects are plausible.

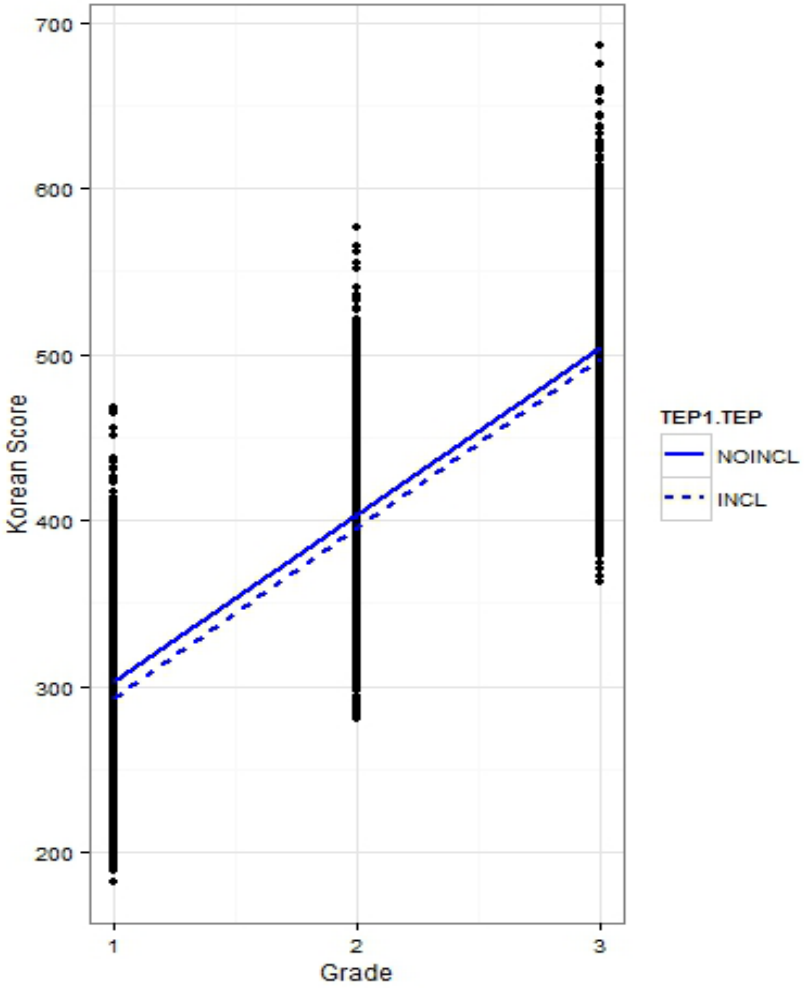

Figure 2. Predicted lines for Korean scores according to TEPSP.

Discussion

This study explores the effectiveness of tying teacher evaluation to student achievement at the individual student level in South Korea. Specifically, it examines the relationship between teacher evaluation policy and growth rates in students' individual achievement scores across time in middle school. This study identified the statistically significant difference in the 1st grade students' achievement score between TEPSP and non-TEPSP, but there is no significant effect on the growth of students' achievement score in the presented model. However, the results should be interpreted with caution, because this study does not use an experimental design and the lack of variation (i.e., 110 schools do not include student achievement in the teacher evaluation while only 14 schools include it) could make it difficult to examine the patterns of relationships between the evaluation policy and student achievement. With this in mind, there are several discussion points and recommendations for further research.

The first research question related to the variability of individual scores' intercept; a change was identified from the model selection process. As previous literature illustrates (Magnuson, 2007; Parveen & Alam, 2008), mother's education level (i.e., whether a mother has a bachelor's degree or above or not) positively influences the individual students' initial score in the 1st grade, not surprisingly. However, there is no effect of mother's

education on the growth rate of students' achievement score. That is, the effect of mother's education on the student achievement score sustains across time in the middle school in the model. Furthermore, the achievement gap between rural schools and urban schools exists in the initial test scores: as the data illuminate, urban schools have 29.44 points higher scores than rural schools. However, urban schools have a different growth rate of 95.67 points while rural schools have a growth rate of 108.25 points. This might be caused by the differences in educational inputs, such as difference of teacher quality between urban and rural areas. In addition, educational process based on the policy effect might cause the different student achievement growth between rural and urban schools. For example, high schools in South Korea are grouped into equalization high schools or non-equalization high schools. The latter one requires an entrance exam and selects students for their school, while in the former one, the educational office randomly chooses students among applicants. Thus, the growth rate in equalization or non-equalization high schools might vary between rural and urban areas, which could cause overall different growth rate.

Second, there is a significant difference in the intercepts of individual Korean scores between TEPSP and non-TEPSP. Specifically, the students in schools that do not include student achievement in the teacher evaluation are higher than students in schools that include student achievement in the teacher evaluation in the model. The initial difference between TEPSP and non-TEPSP might be due to other school-level factors, such as the principal's leadership and educational revenue or expenditures, which were not considered in the final model. However, more importantly, the impact of teacher evaluation including student achievement on the growth rates of individual scores, which is the main interest in this study, was not statistically significant, as Figure 2 shows. This result casts a doubt on the assumption that strong accountability through including student achievement will lead to improve student achievement in South Korea. Based on the result that evaluating teachers based on student achievement does not necessarily increase the student growth, a merit pay system incentivizing by the TEPSP might be misaligned.

Third, the relationship between TEPSP and other variables related to students and teachers (e.g., teacher motivation, student satisfaction, and organizational effectiveness) should be explored from a comprehensive standpoint, given that students' educational outcomes cannot be explained using only test scores. This is because unexpected educational outcomes can also emerge in the school organization. Specifically, TEPSP schools might cause stress for teachers. In addition, the teachers might not conduct a study of teaching materials as they focus only on increasing students' test scores. As a result, the degree of students' satisfaction with teaching and overall educational quality might decrease despite increased scores.

Finally, several limitations of this study should be considered. This study used a quantitative method, incorporating a longitudinal LMM. It only explored the linear relationship between TEPSP and students' Korean test scores. As a result, it could not reveal macro policy insights beyond the quantifiable relationship illuminated by using other research methods, such as interviews and class observations. Thus, TEPSP-related issues need to be explored from diverse perspectives, such as interpretivism and poststructuralism, for epistemological flexibility. In addition, this study used only Korean scores as the dependent variable and only sought to examine the effect of TEPSP within middle schools. Thus, the relationship between TEPSP and other subjects' scores, such as math scores, social studies scores, and science scores, from elementary school to middle school as well as exclusively in middle school, should be examined. Finally, other extraneous variables related to teacher and school effects, such as school climate and teacher professionalism, might exist. Further research also needs to use a top-down approach to start with a fuller model that includes at least all plausible

blocking variables. TEPSP itself might not be judged as having succeeded or failed; thus, efforts should be made to develop a more delicate model related to TEPSP in future research.

Conclusion

This study sought to identify the variability of individual students' Korean test scores as well as examine the relationship between TEPSP and the change in individual Korean test scores. The results of this study led to several conclusions.

First, the results of analyzing the TEPSP model indicated variability in individual intercepts and slopes in Korean scores. Second, the results indicated that TEPSP's influence on the growth rate in individual students' Korean scores within middle school is not statistically significant. This study also emphasized the need for cautious interpretation of the results, as this study did not use an experimental design and TEPSP has not yet been implemented as specific policy in South Korea.

To provide policy insights based on the results, further research should be conducted to examine whether the test skills cause the increase in the students' scores under TEPSP; such research could prevent TEPSP from making education in South Korea more entrance exam oriented. In addition, further research should examine the relationship between TEPSP and other variables related to students' satisfaction and teacher motivation. Scholars have made it clear that TEPSP could affect students' performance, but there is no well-defined strategy to determine on how it does so.

This study is significant in terms of cultivating TEPSP-related issues as unexplored areas in South Korea. This study can serve as a starting point to lead to more developed models using longitudinal data and diverse studies from different perspectives related to TEPSP in South Korea.

References

Anderson, D. R. (2008). *Model-based inference in the life sciences: A primer on evidence.* Retrieved from http://books.google.com/books?id=DIP_h4aMhi YC

Bates, D., Maechler, M., & Bolker, B. (2012). *LME4: Linear mixed-effects models using S4 classes* (R package version 0.999999-0). Retrieved from http://CRAN.R-project.org/package=lme4

Batool, H., Naureen, S., & Kanwa, S. (2010). Studying the effects of socioeconomic status of parents on student's academic achievement. *Journal of Educational Research, 13*(2), 204.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331. Retrieved from http://www.jstor.org/stable/3594120

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review, 48*(April), 147-160.

Dubow, E. F., Boxer, P., & Huesmann, L. R. (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill Palmer Quarterly, 55*(3), 224-249. doi:10.1353/mpq.0.0030

Elmore, R., Abelmann, C. H., & Fuhrman, S. H. (1996). The new accountability in state education reform: From process to performance. In H. F. Ladd (Ed.), *Holding schools accountability: Performance-based reform in education* (pp. 65-98). Washington, D.C.: The Brookings Institution Press.

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal, 44*(3), 594-629. doi:10.3102/0002831207306767

Fitzmaurice, G., Laird, N., & Ware, J. (2004). *Applied longitudinal analysis*. Hoboken, N.J.: John Wiley & Sons, Inc..

Galecki, A., & Burzykowski, T. (2013). *Linear mixed-effects models using R*. New York, N.Y.: Springer.

Goldhaber, D., & Theobald, R. (2012). Do different value-added models tell us the same things (Carnegie Knowledge Network).

Ha, B. W. (2005). A study on the improvement of the educational differences in a decentralization era—The case of Seoul. *The Journal of Educational Administration, 23*(3), 167-193. Retrieved from http://www.riss.kr/link?id=A76481362

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327. Retrieved from http://www.jstor.org/stable/3326211

Herr, K., & Arms, E. (2004). Accountability and single-sex schooling: A collision of reform agendas. *American Educational Research Journal, 41*(3), 527-555. Retrieved from http://www.jstor.org/stable/3699438

Howell, D. C. (2007). The analysis of missing data. In W. Outhwaite, & S. Turner (Eds.), *Handbook of social science methodology.* London: Sage.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*, 761-796. doi:10.1016/j.jpubeco.2004.08.004

Jang, S. T. (2011). A structural analysis of the effectiveness of school autonomy: Focusing on elementary schools (Master's thesis, Korea University). Retrieved from http://dcollection.korea.ac.kr/jsp/common/DcLoOrgPer.jsp?sItemId=000000032322

Jeong, J. Y., Shin, I. S., & Lee, H. S. (2009). A study on the test-based educational accountability system. *The Journal of Korean Teacher Education, 26*(1), 241-261. Retrieved from http://www.riss.kr/link?id=A76381587

Jeong, S. H. (2008). Evaluation of the validity of the teacher evaluation system based on the concept of teaching staff's expertise and accountability. *The Journal of Elementary Education, 21*(2), 409-434.

Korean Educational Development Institute (KEDI). (2009). *Korean educational longitudinal study 2005* (V). Seoul: KEDI.

Koschoreck, J. W. (2001). Accountability and educational equity in the transformation of an urban district. *Education and Urban Society, 33*, 284-304. doi:10.1177/0013124501333004

Lee, G. M., Im, H. J., Park, I. Y., & Kim, Y. J. (2010). A validity study for vertical scale of Korean Educational Longitudinal Study (KELS) 2005. *The Educational Evaluation Research, 23*(3), 617-640. Retrieved from http://www.riss.kr/link?id=A82404496

Lee, K. H. (2010). A study on the improvement of teacher evaluation system for professional development. *The Journal of Korean Teacher Education, 27*(3), 43-68. Retrieved from http://www.riss.kr/link?id=A82403887

Loeb, S., & Candelaria, C. (2013). How stable are value-added estimates across years, subjects, and student groups? (Carnegie Knowledge Network).

Magnuson, K. A. (2007). Maternal education and children's academic achievement during middle childhood. *Developmental Psychology, 43*(6), 497-512.

Mitchell, D. E., & Mitchell, R. E. (2003). The political economy of education policy: The case of class size reduction. *Peabody Journal of Education, 78*(4), 120-152.

National Council on Teacher Quality. (2012). *2012 state teacher policy yearbook*. Retrieved from http://www.nctq.org/stpy11/reports/stpy12_national_report.pdf

Parveen, A., & Alam, M. T. (2008). Does mothers' education influence children's personality factors and academic achievement? *Bulletin of Education and Research, 30*(2), 1-6.

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ryoo, J. H. (2011). Model selection with the linear mixed effects model for longitudinal data. *Multivariate Behavior Research, 46*(4), 598-624. doi:10.1080/00273171.2011.589264

Ryoo, J., & Hong, S. (2010). The effect of performance pay on special education student group achievement. *Special Education Research, 9*(1), 5-19. Retrieved from http://www.riss.kr/link?id=A77010604

Shin, H. R. (2006). A study on determinants of educational disparity among middle schools in educational equalized areas (Doctoral dissertation, Gyeongsang National University). Retrieved from http://www.riss.kr/link?id=T10787681

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*, 415-453.

Song, K. O. (2013). Critical analysis on the primary and secondary education accountability policy in Korea. *The Journal of Yeolin Education, 21*(3), 207-235. Retrieved fromhttp://www.riss.kr/link?id=A99765061

Springer, M. G., Pane, J. F., Le, V. N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team pay for performance: Experimental evidence from the Round Rock Pilot project on team incentives. *Educational Evaluation and Policy Analysis, 34*(4), 367-390. doi:10.3102/0162373712439094

U.S. Department of Education. (2004). *More local freedom.* Retrieved from http://www.ed.gov/nclb/freedom/index.html?src=ov

U.S. Department of Education. (2006). *Growth models: Ensuring grade-level proficiency for all students by 2014*. Retrieved from http://www.ed/gov/admins/lead/account/growthmodel/proficiency.html

White, K. R. (1982). The relation between socioeconomic status and educational achievement. *Psychological Bulletin, 91*, 461-481.

Woessmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review, 30*(3), 404-418. doi:10.1016/j.econedurev.2010.12.008

Woo, H. J. (2011). A study on the real state of the education gap between urban and rural high schools and on school innovation (Master's thesis, Daejin University). Retrieved from http://www.riss.kr/link?id=T12889504

Yuan, K., Le, V. N., McCaffrey, D. F., March, J., Hamilton, L., Stecher, B., & Springer, M. G. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. *Educational Evaluation and Policy Analysis, 35*(1), 3-22.