Pakistan Journal of Education
Vol. 34, No. 2, 2017, 119-140

A Multinational Study of the Determinants of Student Achievement in Mathematics and Science: Policy Options for Pakistan

Gulab Khan[*]

Abstract

Enhancing quality of learning for students has remained a primary target of states and economies around the world. Education systems around the world consider student test scores as objective measures (though with limited explanatory power) to determine the "quality" of student learning. Various strategies are adopted to enhance this matric. One tool that schools use to improve student achievement in the form of test scores has been teacher evaluation. In recent decades, teacher evaluation has come into the spotlight around the world in the current policy debates, reforms, and policy analyses. Therefore, as a significant contribution to the current policy environment, this study explores determinants of student achievement by analyzing data on student background, school traits, teacher evaluation, and country traits. It employs a robust dataset using two surveys i.e., Program for International Student Assessment (PISA) 2009 and Teaching and Learning International Survey (TALIS) 2008. Using Ordinary Least Squares as the analytic model, the study explores relationships between test scores in Mathematics and Science and determinants of student achievement at student, school, and country levels. The study finds mixed results for developmental and high-stakes approaches to teacher evaluation. Powerful associations between determinants of achievement at student, school, and country level suggest that it is important to explore implications of these factors on student achievement in schools.

Keywords: Evaluation, monitoring, accountability, student achievement, classroom observations, administrative tracking.

Introduction

[*] Assistant Professor of Education, National University of Sciences and Technology, Islamabad. gulabkhan.nabi@gmail.com

The primary goal of schools is to improve student achievement for all students. Schools endeavor to achieve this goal by identifying and improving factors that are significant in relation to student achievement. Evidence shows that teacher quality plays a critical role in relation to student achievement in schools (Hanushek, 2003; Hanushek, Kain, Brien, & Rivkin, 2005). Therefore, teacher quality has become a driving theme worldwide in educational policy development and analysis. One way schools can improve quality of their teachers is by evaluating them so as to identify their strengths and weakness, develop them professionally, and holding them accountable for the quality of their practice.

Scholars and policymakers (Taylor & Tyler, 2011; Toch, 2008) agree that teacher evaluation is one of the significant approaches to enhance quality of education for all students. This belief in the efficacy of evaluating teachers coupled with a push from various stakeholders for teacher accountability has thrown teacher evaluation into the spotlight of policy-making and practice in recent decades (Wößmann, Lüdemann, Schütz, & West, 2007). It is in this context that this article presents a study conducted on teacher evaluation approaches and their relationships with student achievement in mathematics and science in 21 countries.

Various studies have explored student achievement using predictors related to individual students, their home and family backgrounds and schools (Fuchs &Wößmann, 2007; Zhang & Lee, 2011). Among the many factors, teacher evaluation with different purposes and approaches has been found to relate to and/or affect student achievement in significant ways (Holtzapple, 2003; Milanowski, 2004; Taylor & Tyler, 2011; Schütz, West, &Wößmann, 2007). Literature shows that countries employ a variety of approaches to and purposes of judging teacher quality and effectiveness. While earlier studies (Wößmann et al., 2007; Schütz et al., 2007) have used previous Program for International Student Assessments (PISA) datasets with focus on accountability aspects of teacher evaluation, this study uses the PISA 2009 dataset to explore teacher evaluation with both high-stakes and developmental approaches with particular attention to the internal teacher evaluations especially by the school principals. It uses PISA in combination with information from the Teaching and Learning International Survey (TALIS) administered by the OECD in 2008. The combination of the two surveys generates a rigorous dataset that takes into account perspectives from principals as well as teachers on teacher evaluation practices in the sample countries. The study specifically aims to answer the question, “How do teacher evaluation practices and purposes associate with student achievement in mathematics and science in lower secondary and secondary schools?"

Teacher Evaluation: Purposes, Approaches and Outcomes

Teacher evaluation, which is synonymous with teacher appraisal, can be construed of as performance reviews conducted by different personnel in schools. "The results of appraisals may be used formatively to identify specific needs for professional development, or summatively for decisions related to promotion, rewards or sanctions" (Looney, 2011, p. 442). In other words, teacher evaluation has two main purposes—formative or developmental purpose and high-stakes or accountability purposes (Danielson & McGreal, 2000).

High-stakes purposes of teacher evaluation have the intended objective of holding teachers answerable for the quality of their professional practice. This focus of evaluation is also concerned with making critical decisions on a person's employability, career advancement or, in extreme cases, relieving someone of his/her services for a lack of needed competencies (Scriven, 1981). In contrast, the developmental purposes of teacher evaluation aim to identify professional training needs of the teachers so as to improve their practice. Such professional development aspects may include "…regular feedbacks by the principal and experienced…to identify priorities for both teacher and school improvement" (Faubert, 2009, p. 29).

Schools evaluate teachers using a variety of instruments and evaluators. Within schools, principals and peers evaluate teachers using instruments such as classroom observations and student achievement including student test scores. They also give feedback to teachers and arrange for reflective sessions to deliberate on successes or failures of observed lessons and lesson plans. Accordingly, an improvement strategy is prepared. Externally, the external evaluator conduct teacher evaluation using tools and means such as student test scores and classroom observations. This type of evaluation has mostly an "accountability" focus (Looney, 2011).

Teacher Evaluation: Empirical Evidence

Teacher evaluation purposes—developmental or high-stakes—do not always work in isolation. A teacher evaluation system may simultaneously carry both the "developmental" and the "high-stakes" purposes. Therefore, this study has operationalized and categorized empirical evidence on teacher evaluation into two broad streams. The first stream (Sartain et al., 2011; Wenglinsky, 2002) consists of empirical evidence that explores standards-based approaches such as classroom observations and rubrics as well as subjective modes of teacher evaluation. The second stream (Goldhaber & Hansen, 2010; Sanders & Horn, 1994; Stronge & Tucker, 2000) consists of literature on teacher evaluation approaches that often use student test scores as a primary measure of teacher performance.

Developmental Teacher Evaluation and Student Achievement

Many studies (Gallagher, 2004; Holtzapple, 2003; Kimball, White, Milanowski, & Borman, 2004; Rockoff & Speroni, 2010; Sartain et al., 2011; Taylor & Tyler, 2011; Tyler, Taylor, Kane, & Wooten, 2010) explore teacher evaluation practices that focus on within-classroom processes and interactions with the purposes of assessing and developing teachers' practice so as to improve student achievement.

Holtzapple (2003) explored how teacher evaluation scores in Cincinnati's Teacher Evaluation System (TES) linked with student achievement. The TES drew upon Danielson's (1996) framework consisting of the domains such as planning and preparation, the classroom environment, instruction, and professional responsibilities. Holtzapple (2003) found that though the evaluation system successfully predicted performance at the extremes (unsatisfactory and distinguished) of performance ratings, it did not effectively predict student achievement at the middle (proficient and basic) level of teacher evaluation ratings. His analyses of student gains and teacher evaluation scores showed that if teachers received "unsatisfactory" and "basic" ratings on "Teaching and Learning Domain," it reflected negatively on student achievement as shown by a lower score relative to predicted score on the basis of prior year's achievement.

Kimball, White, Milanowski, and Borman (2004) studied the relationship between student achievement and standards-based teacher evaluation scores. Their study was similar to Holtzapple's (2003) in the use of Danielson's (1996) framework at their research site. Kimball et al. (2004) found in their multilevel statistical modeling that though there were positive significant relationships between teacher evaluation ratings and student achievement in all subjects and grades that they tested, coefficients were not statistically significant in all cases. In contrast, Milanowski (2004) whose research was also based on teacher evaluations using the Danielson's framework, found small to moderate positive correlations in each of the tested subject. Though the relationships were at best moderately positive, he still considered them significant given that measuring teacher effectiveness using standards-based evaluation rubrics may be noisy due to a number of other confounding factors. Furthermore, a combined analysis of studies conducted at three sites by Milanowski, Kimball, and White (2004) showed that the standards-based teacher evaluations have "…substantial positive relationship with the achievement of the evaluated teachers' students" (p. 19).

Gallagher (2004) explored a teacher evaluation system that had elements of both the developmental and the high-stakes approaches to assessing teacher effectiveness. A predominant focus of the teacher evaluation system at his research site was assessing within-classroom processes followed by feedback. In his study, Gallagher (2004) found strong and statistically significant relationships between teacher evaluation scores and student achievement in reading. The findings for mathematics were positive but statistically insignificant. Similarly, Rockoff and Speroni (2010) in a study of subjective and objective measures of evaluating teachers found these measures to bear significant connection with student performance. They studied teacher evaluations conducted by professional mentors who worked with the new teachers and who made evaluations based on student achievement as a result of first year of teaching of these new teachers.

High-stakes Teacher Evaluation and Student Achievement

High-stakes teacher evaluations have as their main purposes judging teacher effectiveness and making "consequential decisions" (Danielson & McGreal, 2000) relating to, for example, personnel issues of teachers including hiring, firing, salary adjustment and accountability. In high-stakes evaluations, a main source of evidence has been in the form of how well students perform in various assessments.

Student assessment and performance may come in a variety of forms such as school-based tests and external standardized examinations. Proponents (Sanders & Horn, 1994; Stronge & Tucker, 2000) contend that student assessments as an evidence of teacher effectiveness offer good tradeoffs in terms of their objectivity. These proponents suggest using student test scores in valued-added models (VAMs) that apply a pretest-posttest design to statistically isolate teacher effects on student achievement from other confounding factors that emanate at student, school, and family levels (Sanders & Horn, 1994).

To explore the efficacy of student test scores as measures of teacher effectiveness, Bingham, Heywood, and White (1991) studied student performance in a large school system with around 100,000 students. They explored student performance of fifth graders to see if it could be used as a measure to evaluate teachers in high-stakes evaluations. Through a residual and step-wise regression analysis they identified schools wherein teachers had added value to the students whom they taught. They state, however, that their approach could identify only the best and the worst teachers. Following Bingham et al. (1991), Wright, Horn, and Sanders (1997) explored teacher effects on student performance. They applied a mixed-model analysis of variance to study the teacher effects on student achievement. In 20 of the 30 analyses that they conducted, they found teacher effects to be larger than any other effects. Based on their findings, they recommended using student achievement data to assess teachers.

Similarly, Goldhaber and Hansen (2010), using administrative data on teachers and students (grades 4 or 5) showed that employing student test scores as evidence of teacher performance in decisions relating to awarding tenure to teachers (a high-stakes approach to teacher evaluation) had significantly positive effects on student achievement. Restricting their analyses to those teachers whose performance was observed before and after the tenure, teachers who were not selected for tenure had student achievement, on average, more than 11% of an SD lower than teachers who were selected for tenure.

Using student achievement data for accountability to the public such as through posting student results in the media, informing parents about children’s progress, or tracking by administrative authorities had mixed effects on student performance. Wößmann et al. (2007), employing multi-level modeling techniques on the PISA 2003 dataset, reconfirmed findings from the earlier studies (Bishop, 1997, 1999) and asserted that external exit exams had positive relationships with student achievement as measured by test scores after controlling for student, family, school, and country level factors. Their study revealed that schools using external exit exams had students performing significantly better than otherwise.

Table 1

Countries and Cases

Country	Country ID	No. of Cases	No. of Schools	Mean Student Weight
Australia	36	14,251	353	16.93
Austria	40	6,590	282	13.28
Belgium	56	8,501	278	14.04
Brazil	76	20,127	947	103.49
Bulgaria	100	4,507	178	12.87
Denmark	208	5,924	285	10.35
Estonia	233	4,727	175	2.75
Hungary	348	4,605	187	22.94
Iceland	352	3,646	131	1.21
Ireland	372	3,937	144	13.41
Italy	380	30,905	1,097	16.40
Korea	410	4,989	157	126.31
Lithuania	440	4,528	196	8.95
Mexico	484	38,250	1,535	34.00
Norway	578	4,660	197	12.31
Poland	616	4,917	185	91.25
Portugal	620	6,298	214	15.34
Slovak Republic	703	4,555	189	15.21
Slovenia	705	6,155	341	3.06
Spain	724	25,887	889	14.99
Turkey	792	4,996	170	151.61
N	--	212,955	8,116	--

Data and Methods

This study uses two sources of data in order to create a robust dataset that includes perspectives from key stakeholders in teacher evaluation—principals and teachers—in addition to student level information. First, it uses part of the PISA survey conducted by the OECD in 2009 in 65 countries. PISA is a cross-national, large scale survey which is conducted every three years and includes a paper-pencil test in the three subject areas of Mathematics, Science, and Reading. However, this study analyzed predictors of achievement only in mathematics and science. The student tests are given to a sample of 15-year olds in the sampled schools in participating countries.

Second, the study uses information from the OECD (2009a) report that gives descriptive statistics such as country percentages on teachers' perspectives on teacher appraisals and feedback as captured in the TALIS 2008. Like the PISA survey, TALIS is a cross-sectional and cross-national survey administered in 2008 by the OECD to teachers and principals in 22 OECD and 2 partner countries. The study uses part of the PISA 2009 sample consisting of 21 countries that are common between the TALIS 2008 and the PISA 2009 surveys. These countries make up the bulk of the sample in the TALIS 2008.

Table 1 gives the number of cases (212,955) in 8,116 schools in the sample with Iceland having the least (3,646) and Mexico the most (38,250) number of cases. The study uses weights at student and school levels. In order to offset any selection biases and other sampling errors, the student level weights are introduced into the data files. This ensures representative samples and produces unbiased estimates of coefficients on continuous and categorical variables (OECD, 2012).

Table 2 gives descriptive statistics for outcome variables, main predictors, and control variables. The table also gives coding scheme for categorical variables. The outcome variables in this study are student test scores in mathematics and science reported as plausible values (PVs) in the PISA 2009 survey. Main predictors are based on principals' categorical responses to items covering teacher evaluation approaches in the PISA 2009 survey.

These predictors have been grouped into two main categories. The second block in the Table 2 shows variables that include items from the PISA 2009 survey that are based on principals' pedagogical role in relation to their involvement in assessing teachers and their professional development. This category also includes one item that seeks information on use of student assessments for instructional improvement. The third block in Table 2 consists of items on high-stakes approaches to teacher evaluations. These include use of student assessments to evaluate teachers and to judge their effectiveness, if student assessments are tracked by external authority, and if such assessments are posted publicly.

The information taken from the TALIS 2008 to represent country level constructs of teacher evaluation consisted of 14 variables. This study used principal component analysis (PCA) to explore such underlying dimensions as well as to reduce the number variables into viable components. The first component analysis and score generation was run on 8 teacher appraisal criteria. Component analysis with these variables returned two components with Eigen values (EV) greater than 1. Cumulatively, these two components explained 89% of the variance by the 8 variables in teacher appraisal criteria and outcomes. These components were subjected to promax factor rotation. Six of these criteria were loaded onto the first component with component loadings varying between 0.34 and 0.39. This component has been named as "professional outcomes" as evidence of teacher performance in teacher appraisals and feedback. The second component has been named as "other" criteria in teacher appraisal and feedback. The component loadings showed as 0.56 and 0.51 for the two criteria respectively. Component with EV greater than 1 has been retained that explained 74% of the variance by the six variables. Promax factor rotations resulted into component loadings between 0.34 and 0.45. This component has been named as "outcomes and impacts of teacher evaluation."

This study employs Ordinary Least Squares (OLS) as the method of analyzing the data. Regression analyses were run separately on all five plausible values. The coefficients reported in this study are the average of all five coefficients returned by the five separate analyses in each model.

Equations 1 and 2 below represent the models used for the two subjects separately:

$y_{\mathrm{i}} = \alpha_0 + \alpha_1\textit{[Developmental]} + \alpha_3\textit{[High-stakes]} +$
e_i…………………………………....(1)

$y_{\mathrm{i}} = \alpha_0 + \alpha_1\textit{[Developmental]} + \alpha_2\textit{[High-stakes]} + \Sigma\beta_iX_i + \Sigma\delta_iY_i + \Sigma\eta_iZ_i +$
e_i…..........(2)

Equation1 represents models for Mathematics and Science without control variables. In these models, *y* is the predicted score for student *i*in mathematics and science. The parenthetical terms represent the main variables that are labeled as "developmental," and "high-stakes." Equation 2represent models that carry control variables at student, school, and country levels in addition to the main variables. The terms $\Sigma\beta_{\mathrm{i}}X_{\mathrm{i}}$, $\Sigma\delta_{\mathrm{i}}Y_{\mathrm{i}}$, and$\Sigma\eta_iZ_i$give sums of coefficients of the control predictors at the three levels.

Table 2

Descriptive Statistics for Main and Control Variables

Variable	*M*	*SD*	Min	Max
Dependent				
Aggregate plausible value in Math	447.83	98.20	21.00	802.31
Aggregate plausible value in Science	455.70	94.60	37.71	839.74
Developmental				
Classroom observations by school principal (“Quite often” and “very often” coded as 1)	0.58	0.49	0.00	1.00
Principals suggesting teachers for improvement (“Quite often” and “very often” coded as 1)	0.81	0.38	0.00	1.00
Principals informing teachers for updating knowledge and skills (“Quite often” and “very often” coded as 1)	0.91	0.29	0.00	1.00
Assessments used for instructional improvement (“Yes” coded as 1)	0.81	0.39	0.00	1.00
High-Stakes Teacher Evaluation				
Public accountability for student performance (“Yes” coded as 1)	0.34	0.47	0.00	1.00
Student assessments used for evaluating teachers (“Yes” coded as 1)	0.61	0.49	0.00	1.00
Student assessments used for judging teacher effectiveness (“Yes” coded as 1)	0.63	0.48	0.00	1.00
Student assessments tracked by an administrative authority (“Yes” coded as 1)	0.73	0.44	0.00	1.00
Student				
Age	15.78	0.29	15.25	16.33
Girl	0.51	0.50	0.00	1.00
Grade compared to modal grade in the country	-0.17	0.75	-3.00	3.00
Home language other than test language	0.04	0.20	0.00	1.00
Index of socioeconomic and cultural status	-0.73	1.25	-5.71	3.55
School				
Principal’s sex (“Female” coded as 1)	0.40	0.49	0.00	1.00
School type (“Public” coded as 1)	0.84	0.37	0.00	1.00
School size	890.17	756.24	2.00	11268.00
Teacher shortage	0.23	1.17	-1.02	3.34
Proportion of qualified teachers	0.87	0.26	0.00	1.00

Percent girls	50.19	17.00	0.00	100
Student teacher ratio	21.56	16.07	0.27	723.00
Proportion of computers connected to web	0.88	0.25	0.00	1.00
Country				
Professional outcomes (e.g.,student test scores, retention and pass rates) as teacher evaluation criteria	$-9.18e^{-09}$	2.43	-8.12	2.75
Others (e.g., parental feedback and relations with colleagues) as teacher evaluation criteria	$-1.55e^{-08}$	1.09	-1.66	2.36
Outcomes and impact of teacher evaluation	$-1.57e^{-09}$	2.10	-4.83	4.63
Dollars spent on education	883.44	560.51	336.40	3912.80

Results

Table 3 gives regression results for the two models in mathematics and science.

Developmental and High-stakes Approaches to Teacher Evaluation.

Developmental and high-stakes approaches to teacher evaluation in mathematics consisted of eight variables. The findings show that in the absence of control variables, principal's pedagogical roles with respect to teacher evaluation and use of student assessments for instructional improvement show largely negative though insignificant associations with student achievement in mathematics. In model 2, after controlling for factors at student, school, and country levels, 2 of the 4 variables returned negative associations with student achievement in mathematics. Only principals informing teachers about possibilities for updating their knowledge and skills showed as a significant negative correlation (b = -8.329, p < .05) with student achievement in mathematics. In science, all variables under this category returned insignificant associations in model 2.

With regard to high-stakes approaches to teacher evaluation, all but one variable related negatively with student achievement in mathematics without controlling for background factors. In science, two variables showed significant associations with student achievement without background controls. However, like the behavior of variables in the developmental category, the high-stakes approaches to teacher evaluation recorded a change in coefficients when control variables were added in model 2. Public accountability persisted as a significant positive predictor of student achievement with coefficients of 9.595 (p < .001) in mathematics, and 8.710 (p < .001) in science. Use of student assessments for evaluating teachers showed a negative though insignificant relation with student achievement in mathematics and science. Tracking of student assessments by an administrative authority and student assessments used for judging teacher effectiveness showed insignificant associations with student achievement in mathematics and science.

Powerful Influence of Background Factors

Factors at student, school and country level showed powerful associations with student achievement. This behavior of background factors was consistent with previous studies (Fuchs & Wößmann, 2007; Zhang & Lee, 2011).

Student age showed a significant negative influence on student achievement both in mathematics (b =11.251, p < 0.001) and Science (b = -11.384, p < 0.001). Being a girl was a disadvantage with regard to student achievement in mathematics and science. Significant negative associations were observed as being a girl associated with about 18.5 (p < .001) decrease in score in mathematics and with about 7.3 (p < .001) decrease in score in science. Being in a higher grade was naturally reflected in higher scores in mathematics and science. As shown in previous studies, one of the most significant determinants of student achievement was students' socioeconomic status. Belonging to a higher socioeconomic class was associated with over 23 (p < .001) point increase in mathematics and over 22 (p < .001) point increase in science.

Table 3

Determinants of Student Achievement in Mathematics and Science

	Mathematics		Science	
	(1)	(2)	(1)	(2)
Developmental				
Classroom observations by school principal	-1.456	1.345	-1.999	0.769
	(-0.38)	(0.51)	(-0.59)	(0.33)
Principals suggesting teachers for improvement	-21.235***	-3.070	-17.897***	-2.348
	(-4.98)	(-1.26)	(-4.65)	(-1.08)
Principals informing teachers for updating knowledge and skills	-14.179*	-8.329*	-7.287	-5.444
	(-2.20)	(-2.41)	(-1.20)	(-1.82)
Assessments used for instructional improvement	12.393**	2.871	8.910*	-0.176
	(2.81)	(0.81)	(2.37)	(-0.06)
High-Stakes				
Public accountability for student performance	17.474***	9.595***	16.521***	8.710***
	(5.00)	(4.22)	(5.30)	(4.40)
Student assessments used for evaluating teachers	-21.199***	-3.403	-20.531***	-3.903
	(-5.77)	(-1.50)	(-6.37)	(-1.96)
Student assessments tracked by an administrative authority	-7.470*	-1.707	-5.537	0.721
	(-2.17)	(-0.76)	(-1.86)	(0.36)
Student assessments used for judging teacher effectiveness	-8.194*	2.482	-6.471	2.797
	(-2.08)	(0.83)	(-1.81)	(1.00)
Student Level Predictors				
Student age		-11.251***		-11.384***
		(-7.91)		(-8.58)
Girl		-18.465***		-7.257***
		(-24.84)		(-10.34)
Grade		31.580***		31.963***
		(31.67)		(35.82)
Index of social, cultural and economic status		23.483***		22.198***
		(36.60)		(41.36)
Home language other than test language		-6.683***		-13.620***
		(-3.78)		(-6.71)
School Level Predictors				
Principal's sex (female)		-12.379***		-7.747***
		(-5.37)		(-4.03)
Public school		-15.906***		-15.754***
		(-5.97)		(-7.07)
School size		0.000		0.001
		(0.31)		(0.62)
Teacher shortage		-4.014***		-4.721***
		(-3.88)		(-5.23)
Proportion of qualified teachers		1.784		10.431**
		(0.54)		(3.10)
Proportion of girls		0.166**		0.190***
		(2.82)		(4.42)
Student teacher ratio		-0.498***		-0.527***
		(-5.66)		(-5.75)
				(N = 210,307)

Table 3

Determinants of Student Achievement in Mathematics and Science (Continued)

	Mathematics		Science	
	(1)	(2)	(1)	(2)
Proportions of computers connected to Web		17.840*** (4.25)		23.850*** (6.26)
Country Level Predictors				
Professional outcomes (e.g., student test scores, retention and pass rates) as teacher evaluation criteria		-13.499*** (-17.62)		-10.290*** (-15.33)
Others (Feedback from parents, relations with colleagues) as teacher evaluation criteria		-5.624*** (-6.00)		-0.553 (-0.60)
Outcomes and impact of teacher evaluation		3.891*** (5.53)		1.613* (2.53)
Dollars spent on education		-0.366 (-1.74)		-0.351* (-2.01)
_cons	491.487*** (81.87)	674.023*** (28.03)	490.640*** (89.54)	661.368*** (29.81)
Average R^2	0.079	0.422	0.072	0.405

t statistics in parentheses; $^{*}p< 0.05$, $^{**}p< 0.01$, $^{***}p< 0.001$(N = 210,307)

Speaking a different language at home was found to have negative association with student achievement in both subjects.

At the school level, having a female principal showed to have a negative influence on test scores in mathematics (b = 12.379, p < .001) and science (b = -7.747, p < .001). A student enrolled in a public (government) school suffered a disadvantage of about 16 points (p < .001) in mathematics and about the same points at the same level of significance in science. Having a shortage of teacher also negatively influenced student achievement by about 4 points in mathematics and about 5 points in science. A school having qualified teachers was a good omen for increasing student achievement in science but this variable returned insignificant positive associations in mathematics. Having a greater proportion of girls in schools marginally influenced positively on student test scores in both subjects. Schools having a larger proportion of computers connected to the web showed significant increase in student achievement in both subjects.

The three components derived through PCA also showed significant associations with student achieving. The first component of teacher evaluation, “professional outcomes,” was negatively associated with student achievement with coefficients of -14.003 (p < .001) in mathematics, and -10.524 (p < .001) in science. The second component, "others," showed significant negative coefficient of -5.855 (p < .001) in mathematics and an insignificant negative coefficient of -0.853 (p = .356) in science. The third component, "outcomes and impact of teacher evaluation," showed a significant positive association with student achievement in mathematics with a coefficient of 3.255 (p < .001). It remained positive but an insignificant association in science (b = .840, p = .235).

The analysis showed that main predictors explained about 8% variation in student achievement whereas a large majority of variation (over 40% in both subjects) was explained by background factors at student, school and country levels. This suggests that while teacher evaluation practices do have a significant bearing on student achievement, the background factors have a prominent role to play regarding increase in student achievement in mathematics and science.

Discussion and Conclusion

This study analyzed teacher evaluation practices with "developmental" and "high-stakes" purposes attached with the process. Developmental purposes included principals' evaluative focus as a pedagogical leader, and use of student assessments for instructional improvement. High-stakes purposes included public accountability, use of student assessments for judging and evaluating teachers, and administrative tracking. Findings in this study have significant implications with regard to how teachers are evaluated and by using what matrices.

In Pakistan, an increasing emphasis on student test scores as the primary matric of quality of learning in schools need to be revisited. This is because, even though high-stakes approaches to teacher evaluation indicate that public accountability related positively and significantly with student achievement, a finding consistent with prior evidence (e.g., Hanushek & Raymond, 2005), attaching high-stakes purposes to the process garners consequences that are unintended and in some instances detrimental to the overall educational goals of schools. In this regard, findings of the study sync with the assertions from scholars who caution about using student assessments as the sole measures of teacher performance (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Mathis, 2012; Rosenkvist, 2010). The unintended consequences may come in the form of dissipation of teacher morale and deterioration of a culture of collaboration among teachers (Farrell & Morris, 2004), a narrowing of focus in content and curriculum (Berliner, 2011), and harmful effects such as dropouts for students particularly from disadvantaged backgrounds (McNeil, Coppola, Radigan, & Vasquez Heilig, 2008). Anecdotal evidence and general observation in Pakistan highlight many of these issues associated with high-stakes uses of student test scores.

While public accountability related positively with student achievement, use of student assessments for evaluating and judging teachers and administrative tracking of student assessments bear overall negative though insignificant relationships with student achievement. The findings of the study also have important policy implications with regard to the use of student assessments for making high-stakes decisions in teacher evaluations. Student assessments used for teacher evaluation and for administrative tracking, which often come with high-stakes consequences, appear to be a strategy that suffers pitfalls as suggested by their largely negative associations with student achievement in this study. All in all, results in this study challenge the proposition wherein student test scores are offered as effective measures of teacher performance in high-stakes teacher evaluation systems (e.g., Goldhaber & Hansen, 2010; Sanders & Horn, 1994; Stronge & Tucker, 2000; Wright et al., 1997). In Pakistan, an increasing emphasis on student test scores as the primary matric of quality of learning in schools need to be revisited. Therefore, in the light of this finding and prior evidence (e.g., Berliner, 2011; Koretz, 2008; Menken, 2006; Suen& Yu, 2006), it would be a relevant policy proposition to cut down on the share of student assessments in teacher evaluations, especially involving high-stakes outcomes for teachers. Also, student assessments as the sole measure of teacher performance will need careful examination for the various issues associated with this practice (Kornhaber, 2004; Mathis, 2012; Rosenkvist, 2010). Attaching high-stakes consequences may show short term gains in student achievement, they may not be effective in the long run and that student learning may suffer from issues of watering-down of curriculum leading to what is generally known as “teaching to the test” effect.

Last but not the least, as can be seen in this study and earlier studies, factors at student and school level are powerful determinants of student achievement in both mathematics and science. A policy reform that is blind to the strong undercurrents of socioeconomic disparities that are play at student and school level will yield little to no value with regard to raising quality of learning for all students in the country. The deep divisions that exist among schools in public and private appear to powerfully thwarting any effort to equalize success for all students in the country. The constitutional provision of free and quality schooling up to secondary level will need to be looked at from the dimension of the immense divisions that have made their ways into the country's education system without which it seems quite improbable that educational will see any significant improvement in the medium to long term in Pakistan.

References

Berliner, D. (2011). Rational responses to high-stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287–302.

Bingham, R. D., Heywood, J. S., & White, S. B. (1991). Evaluating schools and teachers based on student performance: Testing and alternative methodology. *Evaluation Review*, *15*(2), 191–218

.

Bishop, J. H. (1997). The effect of national standards and curriculum-based exams on achievement. *American Economic Review*, *87*(2), 260-264.

Bishop, J. H. (1999). Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review, 6* (2), 349-398.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, B. L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi DaltaKappan, 93*(6), 8-15.

Farrell, C., & Morris, J. (2004). Resigned compliance: Teacher attitudes towards performance-related pay in schools. *Educational Management Administration & Leadership, 32*(1), 81–104.

Faubert, V. (2009). *School evaluation: Current practices in OECD countries and a literature review*, OECD Education Working Papers, No. 42, OECD Publishing. http://dx.doi.org/10.1787/218816547156

Fuchs, T., & Wößmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics*, *32*(02), 433-464. DOI 10.1007/s00181-006-0087-0

Gallagher, H. A. (2004). Vaughn Elementary's Innovative Teacher Evaluation System: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*, *79*(4), 79–107.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review*, *100*(2), 250–255.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *The Economic Journal*, *113*(485), F64–F98.

Hanushek, E. A., Kain, J. F., Brien, D. M. O., &Rivkin, S. G. (2005). *The market for teacher quality* (Working Paper No. 11154). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w11154

Hanushek, E.A., & M.E. Raymond (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297-328.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, *17*(3), 207–219.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54–78.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press

.

Kornhaber, M. L. (2004). Appropriate and inappropriate forms of testing, assessment, and accountability. *Educational Policy*, *18*(1), 45–70. doi:10.1177/0895904803260024

Looney, J. (2011). Developing high-quality teachers: Teacher evaluation for improvement, *European Journal of Education*, *46*(4), 440–455.

Mathis, W. (2012). *Research-based options for education policy making*. Retrieved from National Education Policy Center website: http://nepc.colorado.edu

McNeil, L. M., Coppola, E., Radigan, J., & Vasquez Heilig, J. (2008). Avoidable losses: High-stakes accountability and the dropout Crisis. *Education Policy Analysis Archives*, *16*(3). Retrieved from Policy Analysis Archives website: http://epaa.asu.edu/epaa/v16n3/

Menken, K. (2006). Teaching to the test: How No Child Left Behind impacts language policy, curriculum, and instruction for English language learners. *Bilingual Research Journal*, *30*(2), 521–546.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-53.

Milanowski, A. T., Kimball, S. M., & White, B. (2004). *The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites*. Retrieved from Consortium for Policy Research in Education website: *www.cpre-wisconsin.org/papers/3site_long_TE_SA_AERA04TE.pdf*

Organization for Economic Cooperation and Development. (2009a). *Creating effective teaching and learning environments: First results from TALIS* Retrieved from Organization for Economic Cooperation and Development website: http://www.oecd.org/edu/school/43023606.pdf

Organization for Economic Cooperation and Development. (2010b). *TALIS 2008 technical report*. Retrieved from Organization for Economic Cooperation and Development website: http://www.oecd-ilibrary.org/education/talis-2008-technical-report_9789264079861-en

Organization for Economic Cooperation and Development. (2012). *PISA 2009 technical report*. Retrieved from Organization for Economic Cooperation and Development website: http://dx.doi.org/10.1787/9789264167872-en

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review*, *100*(2), 261–266.

Rosenkvist, M. A. (2010). *Using student test results for accountability and improvement: A literature review* (Working Paper, No. 54). Retrieved from OECD website: http://dx.doi.org/10.1787/5km4htwzbv30-en

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. doi:10.1007/BF00973726

Sartain, L., Stoelinga, S. R., Brown, E. R., Luppescu, S., Matsko, K. K., Miller, F. K., &Durwood, C. E. (2011). *Rethinking Teacher Evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Retrieved from Consortium on Chicago School Research website: http://ccsr.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf

Schütz, G., West, M. R., & Wößmann, L. (2007). *Autonomy, choice, and the equity of student achievement: International evidence from PISA 2003*. Retrieved from OECD website: http://dx.doi.org/10.1787/246374511832

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 244-271). Beverly Hills: Sage Publications.

Stronge, J. H., & Tucker, P. D. (2000). *Teacher evaluation and student achievement*. Washington, DC: National Education Association.

Suen, H. K., & Yu, L. (2006). Chronic consequences of high-stakes testing? Lessons from the Chinese civil service exam. *Comparative Education Review*, *50*(1), 46–65. doi:10.1086/498328

Taylor, E. S., & Tyler, J. H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers* (Working Paper No. 16877). Retrieved from

National Bureau of Economic Research website: http://www.nber.org/papers/w16877

Toch, T. (2008). Fixing teacher evaluation: Evaluations pay large dividends when they improve teaching practices. *Educational Leadership*, *66*(02), 32–37.

Tyler, B. J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review*, *100*(02), 256–260.doi:10.1257/aer.100.2.256

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, *10*(12), 1–30.

Wößmann, L., Lüdemann, E., Schütz, G., & West, M. R. (2007). *School accountability, autonomy, choice, and the level of student achievement: International evidence from PISA 2003*. doi: http://dx.doi.org/10.1787/19939019

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57-67.

Zhang, L. & Lee, K. A. (2011). Decomposing achievement gaps among OECD countries. *Asia Pacific Education Review*, *12*(3), 463–474. DOI 10.1007/s12564-011-9151-3.

Citation of this Article:

Khan, G. (2017). A multinational study of the determinants of student achievement in mathematics and science: Policy options for Pakistan. *Pakistan Journal of Education*, *34*(2), 119-140.

Received on: April 04, 2017
Revised on: November 03, 2017
Accepted on: November 24, 2017