

MULTILEVEL RELATIONS BETWEEN EXTERNAL ACCOUNTABILITY, INTERNAL ACCOUNTABILITY, AND MATH ACHIEVEMENT: A CROSS-COUNTRY ANALYSIS

Pilnam Yi

Hongik University, South Korea
E-mail: pilnamyi@hongik.ac.kr

In-soo Shin

Jeonju University, South Korea
E-mail: s9065031@jj.ac.kr

Abstract

External accountability policies have spread fast across various educational systems over the past decades. This research examines the relations of internal and external accountability with students' math achievement drawing on PISA 2012. With a sample of 44 educational systems, of which external accountability policies were identified, the research conducted three-level hierarchical linear modelling (HLM) analyses. This research found that some internal accountability factors had tighter relations with math achievement, while the relations of external accountability policies with student performance were rather tenuous. However, equity of student math achievement was better ensured under strong accountability systems. The results suggest that policy makers of each country should consider strengths and weaknesses of external accountability in their own educational contexts.

Keywords: *external accountability, educational equity, internal accountability, math achievement, PISA.*

Introduction

The past two decades have seen educational accountability combined with student assessment spread fast across educational systems worldwide. Although there exist substantial cross-national variations in specific policy measures, educational policy makers have increasingly adopted standards-, and performance-based educational reforms accompanied by national assessment, which was motivated by growing international educational testing and the global new public management trend (Kamens & McNeely, 2010; Morris, 2011). England and the U.S. were forerunner countries in performance-based accountability policies, and it did not take long for many countries in distant world regions, for example, South Korea, Australia, Norway and Sweden to follow suit (Elstad, 2009; Lingard, 2010; Lundahl & Waldow, 2009; Sung & Kang, 2012). Scholars suggest that this transnational policy diffusion was accelerated by OECD PISA among others (Meyer, Tröhler, Labaree, & Hutt, 2014; Ozga, 2013).

The global policy convergence toward performance-based accountability in education raises effects of external accountability policies on student achievement as a significant research issue, which inevitably conveys important policy implications. With its growing stature in educational policy reforms across nations, pervasive external accountability movement in education becomes a global educational policy issue that needs rigorous research evidence.

Given the significance of policy concerns about this topic, a growing number of empirical studies have accumulated, but study results remain mixed and inconclusive.

Previous research, conducted predominantly in the U.S. and the U.K. context, has not reached a consensus about the effectiveness of external accountability, which features state-wide tests, public reporting of test results, and rewards or sanctions based on the test results. Some scholars found significantly positive effects of external accountability (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2005), while a meta-analysis on the effect of test-driven external accountability reported a modestly positive effect on average achievement (Lee, 2008). Yet others found no significant effect of external accountability policies, particularly with little equity enhancement (Lee & Reeves, 2012; Lee & Wong, 2004). Some cross-country studies suggested that student test scores were significantly higher in nations with external exit exams (Bishop, 1998; Schütz, Lüdemann, Woessmann, & West, 2010; Woessmann, 2005) while another study by Lee & Amo (2017) reported no significant growth of student achievement in nations with high school exit exam policies compared to other nations without high school exit exam policies.

On the other hand, many scholars contended that external accountability alone would not be linked to improved, and long-term student learning outcomes, and that internal accountability and organizational capacity should precede external accountability (Elmore, 2004; Newmann, King, & Rigdon, 1997; O'Day, 2002). However, these critics of externally driven educational accountability policy tended to put forth their arguments based on qualitative case-study results.

This research aims to reveal multilevel relations between external accountability of educational system, internal accountability within schools, and students' math achievement, which was the main subject of PISA 2012. Accordingly, many background questions were asked about math.

Literature Review

Concepts of External vs. Internal Accountability

School accountability, in which school performance is evaluated using student performance measures, is increasingly prevalent around the world (Figlio & Loeb, 2011). Although there are various conceptions of educational accountability, such as political, legal, bureaucratic, professional, moral, and market accountability (Adams & Kirst, 1999), performance-based, or test-based accountability has predominantly driven educational reform policies in many educational systems for the past decades. Under the external accountability system, schools are externally or outwardly accountable for student academic performance, which is sometimes published for public information, and based on which schools are rewarded or penalized. Generally, school accountability systems include three elements: state-wide student tests, public reporting of school performance, and rewards or sanctions based on some measures of school performance or improvement (Kane & Staiger, 2002). Examples of external accountability policies include No Child Left Behind (NCLB) Act in the U.S. and National Assessment Program – Literacy and Numeracy (NAPLAN) in Australia.

In contrast to external accountability, many scholars have emphasized internal, school-oriented accountability, in which system professionally negotiated standards for school performance development (Adams & Kirst, 1999; Newmann, King, & Rigdon, 1997). While an external accountability system imposes standards from outside and mostly lacks a capacity building component, an internal accountability model establishes performance standards through professional negotiation and knows site needs for capacity building. Most importantly, alignment between site and system standards is possible and impact of teaching is direct with internal accountability models.

Elmore (2004) conceptualized internal accountability as holding teachers accountable for student learning in line with personal responsibility and shared expectations, combined with

certain consequences. In other words, teachers and schools are accountable for student learning, of which standards are shared in the school community. When the shared expectations are not met or internalized by teachers, there should be internal consequences. Elmore and Fuhrman (2001) argued that internal accountability should precede external accountability in school reforms for genuine school improvement.

Relations of External Accountability with Student Achievement

In theory, the effectiveness of external accountability in education are premised on a principal-agent model, in which the parent is a principal who commissions school teachers as agents to educate the child on his or her behalf (Wößmann, Lüdemann, Schütz, & West, 2007). Due to the asymmetric information problem concerning the efforts of teachers, external accountability measures such as high-stakes testing, public reporting of school results, and rewards or penalty based on the results are supposed to incentivize teachers to make efforts.

In practice, empirical investigation on the effects of external accountability on student achievement has been carried out in the U.S. domestic and international context, although there are some other country cases. The U.S. cross-state analyses used the National Assessment of Student Progress (NAEP) results to identify the effects of the NCLB on student achievement (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2005; Lee & Reeves, 2012; Lee & Wong, 2004).

The US-based empirical studies revealed that evidence was mixed: some studies reported significantly positive effects of external accountability, particularly on math achievement (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2005), while others suggested that the NCLB did not generate sustainable and generalizable policy effects, particularly with little equity enhancement (Lee & Reeves, 2012; Lee & Wong, 2004).

Hanushek and Raymond (2005) argued for consequential accountability policies based on their findings that reporting results alone had minimal impact on student performance. By contrast, Burgess, Wilson, & Worth (2013) found that abolition of school performance tables negatively affected school effectiveness in Wales, the U.K. This finding supported by public reporting only could impact school efforts.

In a meta-analysis, Lee (2008) concluded that the high-stakes accountability policy showed a modestly positive effect on average achievement. Figlio and Ladd (2008), in their review of literature, suggested that school accountability policy seemed more effective in math than in reading with a modest effect size in general, and it hardly reduced the achievement gap, particularly between White and Black students.

Recently, Lee and Amo (2017) found that neither student accountability policy (high school exit exam with consequences for students) nor school accountability policy (high-stakes testing with consequences for teachers and schools) affected average achievement growth in grade 8 math of NAEP. In addition, they found either student-targeted or school-targeted accountability policy did not close the achievement gaps in the U.S.

International cross-country studies using PISA data have provided fairly positive evidence on the relations of accountability measures with student achievement. Hanushek and Woessmann (2010) suggested that accountability was an important institutional feature that contributed to higher student performance. For example, students in systems with central exit examinations were more likely to perform better (Jürges, Richter, & Schneider, 2005; Woessmann, 2005) and public posting of school performance was positively related with student achievement (Boarini & Ludemann, 2009). Accountability measures aimed at teachers and schools were positively associated with student achievement (Schütz, Lüdemann, Woessmann, & West, 2010). Particularly, external accountability was effective when combined with autonomy (Hanushek, Link, & Woessmann, 2013). In addition, the relations of various accountability measures with student achievement did not significantly differ for students with different SES (Schuetz, Luedemann, West, & Woessmann, 2013).

Recently, Lee and Amo (2017) indicated that school accountability policy was found to be an ineffective policy from an international comparative perspective. That is, initially low-performing countries were more likely to adopt stronger school accountability policies, which later did not contribute to making more academic progress or closing the achievement gaps within countries. Although the study has a methodological limitation of a small sample size, which is linked to a lack in statistical power to detect policy effects, it is worthwhile to note their consistent findings of ineffectiveness of external accountability policies.

Relations of Internal Accountability with Student Achievement

Critics of external accountability policies suspected that externally imposed accountability might not ensure school improvement, which actually requires internal organizational capacity (Elmore & Fuhrman, 2001; Newmann et al., 1997; O'Day, 2002; Vanhoof & Petegem, 2007). They recognized the importance of internal accountability for bona fide school improvement. Furthermore, some scholars proposed professional accountability based on trust as an alternative to external accountability to realize school reforms (Møller, 2009; O'Neill, 2013; Sahlberg, 2008). However, those who oppose to external accountability emphasizing internal accountability tended to base their proposition on a conceptual ground rather than on empirical evidence.

An empirical study that combines external and internal accountability was conducted by Lee and his colleagues (2014), who attempted to disentangle the relations between external standards, internal standards and student achievement using longitudinal data of the U.S. In this study, internal standards are translated as teacher expectation considering students' prior achievement and background. They found that the linkage between state standards and student achievement was tenuous, whereas the linkage between teacher standards and student achievement was solid and reciprocal.

Elmore (2004), based on school case studies, concluded that schools with strong internal accountability functioned more effectively under external accountability pressures. He writes, "Strong internal accountability is a condition that precedes and determines a school's response to external accountability (Elmore, 2004, p. 134)." There are few empirical analyses on the relations of internal accountability as a comprehensive concept with student achievement. Rather, researchers explored how each component of internal accountability, such as principal leadership, and teacher morale is related to student performance. For example, Hallinger and Heck (1998) reviewed literature on principal leadership and student achievement and reported that school principals have indirect, yet statistically significant influence on student achievement. Leithwood and Day (2008) claimed that principal leadership affects student outcomes indirectly and most powerfully through staff motivation. The lack of empirical studies comparing the relations of external versus internal accountability with student achievement using international data justifies the significance of this study.

Research Questions

This research examines the effectiveness of external accountability policies in terms of the level of math achievement and equity of achievement. It first hypothesizes that internal accountability is strongly associated with student math achievement, while the relations between external accountability and student math achievement are weak. With respect to equity of achievement, previous research provided inconsistent findings. Thus, this research explores how external accountability policies at the system level are related with the effects of SES on math achievement. Specifically, we address the following research questions in this research:

1. Whether and how much external accountability policies at the system level and internal accountability components at the school level are associated with students' math achievement?
2. Is a strong external accountability system effective in narrowing math achievement gaps of students from different SES groups?

Methodology of Research

Data and Sample

This research is a quantitative secondary analysis study, using the PISA 2012 data to examine whether and how much external and internal accountability is related to student achievement across different educational systems. The OECD has conducted large scale tests including mathematical, reading, and scientific literacy for international students among 15-year-olds every three years since 2000. The sampling design used for the PISA assessment was a stratified sample design in all countries (OECD, 2014a). The sampling units consisted of schools having 15-year-old students from a comprehensive national list of all PISA-eligible schools in 65 participating countries in PISA 2012. Compared to the previous cycles, PISA 2012 provides richer information on school accountability. For example, measures of quality assurance, principal leadership behaviors, and consequences of teacher evaluation are included in the school questionnaire. Hence, it provides an opportunity to examine multilevel relations between external accountability at the system level, internal accountability at the school level, and student achievement at the individual level.

The sample selection process was as follows. Among 65 education systems that originally participated in PISA 2012, 46 systems were first identified with national-level accountability policies through extensive literature reviews. Then, the acquired information was confirmed with national project managers of PISA 2012 of each country via email in January and February 2017. With a slight revision of information based on the email correspondences, 44 systems were finally selected for analyses of this research. Two systems, the U.K. and Shanghai-China, were excluded in the final analysis due to a severe missing data problem. The sample restrictions resulted in final sample sizes of 314,327 students and 12,183 schools from 44 educational systems. For comparisons of high- and low-external accountability systems as of 2012, 6 countries that take all of three external accountability measures including Australia, Chile, Hungary, Korea, Mexico, and the USA and 8 countries with no external accountability policies in their education system such as Switzerland, Spain, Finland, Greece, Croatia, Liechtenstein, Lithuania, and Macao-China were selected.

Variables

First, the dependent variable of this study is individual students' math achievement in PISA 2012, using 5 plausible values.

Second, external accountability measures 1) whether national testing exists (NATION-TE), 2) whether school performance are reported publicly, for example, on the website (REPORTIN), and 3) whether there are sanctions and rewards based on student performance (SAN-REW). In reference to national testing, implementation of either standards-based tests or exit exams was considered. That is, national testing was coded 1 if either standards-based tests or central exit exam was in place and implemented in the educational system as of 2012. External accountability measures may include strong pressures such as threat of reconstitution, principal transfer and loss of students (Carnoy & Loeb, 2002), but this research confined external accountability measures to only three simplified variables according to Kane and Staiger (2002).

The school questionnaire of PISA 2012 included a few questions indicating the system level accountability such as whether student assessments are used to compare the school to district or national performance, whether achievement data are posted publicly. However, it is uncertain that the state-wide tests and public reporting of school performance were practically institutionalized in an educational system since there existed substantial variation across schools that answered the questionnaire. Moreover, the PISA background questionnaire did not provide information about whether there are sanctions and rewards based on test results. Thus, extensive reviews of documents from OECD, UNESCO, World Bank, EU, national govern-

ments, and official websites in addition to scholarly articles were conducted to identify the status of external accountability systems.

Third, we constructed internal accountability measures based on Elmore (2004)'s concept of internal accountability, which consists of teachers' individual responsibility aligned with collective expectations and internal consequences when misalignment is detected. First, individual responsibility from the perspective of teachers can be equivalent to teacher morale factor in PISA (TCMORALE). The questionnaire asks the respondent (principal) how much they agree with the following questions: 1) the morale of teachers is high; 2) teachers work with enthusiasm; 3) teachers take pride in this school; and 4) teachers value academic achievement. Second, collective expectations can be measured by parental expectation and their participation in school activities, teacher monitoring, and principal leadership. The PISA 2012 includes a question about parental expectations towards the school by asking how much the school receives parental pressures about high academic standards (PAREXPEC). Also, the proportions of parents who participate in a variety of school-related activities are asked (PARPART). For teacher monitoring, a question about methods of monitoring the practice of math teachers was used (TMONITOR). Regarding principal leadership, we used four factor variables provided by the OECD. The school questionnaire for PISA 2012 contained 21 items about school leadership activities and 4 factor variables are presented: LEADCOM, LEADINST, LEADPD, and LEADTCH (OECD, 2014a, pp.345-346). LEADCOM represents 4 items for framing and communicating the school's goals and curricular development; LEADINST does 3 items for instructional leadership; LEADPD does 3 items for promoting instructional improvements and professional development; and LEADTCH does 3 items for teacher participation in leadership. Third, for the measure of internal consequences (CONSEQ), we used the question concerning teacher appraisals. The question reads, "To what extent have appraisals of and/or feedback to teachers directly led to a change in salary, a financial bonus or another kind of monetary reward, opportunities for professional development activities, a change in the likelihood of career advancement, public recognition, changes in work responsibilities that make the job more attractive, and a role in school development initiatives?"

Finally, in the multilevel model specification, we included a few key control variables that could influence student achievement at the student, school, and system level. On the basis of prior empirical studies, we included gender, SES, and math self-efficacy as control variables at the student level. School control (private or public) and school average SES were controlled for at the school level. Finally, at the system level, per capita GDP (LNGDP) was controlled for.

Data Analysis

A three-level hierarchical linear modeling (HLM) was used to estimate the relations between external and internal accountability and students' achievement, controlling for variables affecting student achievement. This multilevel analytical method was chosen because the data have a hierarchical structure with individual students nested within schools within educational systems (Raudenbush & Bryk, 2002). The level 1, level 2, and level 3 models are as follows:

Unconditional Model

Level 1 model (student level):

$$Y_{ijk} = \pi_{0jk} + e_{ijk}, \quad e_{ijk} \sim N(0, \sigma^2).$$

where Y_{ijk} is the math achievement of student i in school j and country k ; π_{0jk} is the intercept for school j in country k ; and e_{ijk} is a level-1 random effect that represents the deviation of student ijk 's score from the predicted score based on the student-level model.

Level 2 model (school level):

$$\pi_{0jk} = \beta_{00k} + \gamma_{0jk},$$

where β_{00k} is the mean achievement in school k , and γ_{0jk} is a level-2 random effect that represents the deviation of school mean.

Level 3 model (country level):

$\beta_{00k} = \gamma_{000} + u_{00k}$,
where γ_{000} is the grand mean, and u_{00k} is a level-3 random effect

Conditional Model

Level 1 model (student level):

$$Y_{ijk} = \pi_{0jk} + \sum_{p=1}^P \pi_{jk}(\alpha_p)_{ijk} + e_{ijk}, e_{ijk} \sim N(0, \sigma^2).$$

where y_{ijk} is the math achievement of student i in school j and country k ; π_{0jk} is the intercept for school j in country k ; α_p are $p = 1, \dots, P$ student characteristics that predict math achievement; are the corresponding level-1 coefficients that indicate the direction and strength of association between each student characteristics, α_p , and the outcome in school jk ; and e_{ijk} is a level-1 random effect that represents the deviation of student ijk 's score from the predicted score based on the student-level model.

Level 2 model (school level):

$$\pi_{0jk} = \beta_{00k} + \sum_{q=1}^{Qp} \beta_{0qk}(\pi_{0jk}) + \gamma_{0jk},$$

where β_{00k} is the intercept for country k in modeling the school effect π_{0jk} ; β_{0qk} is the corresponding coefficient that represents the direction and strength of association between school characteristic π_{0jk} ; and r_{0jk} is a level-2 random effect that represents the deviation of school mean.

Level 3 model (country level):

$$\beta_{00k} = \gamma_{000} + \sum_{s=1}^{Spq} \gamma_{00s}(\beta_{00k}) + u_{00k},$$

where γ_{000} is the intercept term in the country-level model for β_{00k} ; γ_{00s} is the corresponding level-3 coefficient that represents the direction and strength of association between country characteristic β_{00k} ; and u_{00k} is a level-3 random effect that represents the deviation of grand mean.

Results of Research

Descriptive Statistics

Table 1 shows the descriptive statistics for the full sample of data from 314,327 students and 12,183 schools from 44 educational systems. In addition, the descriptive statistics for the sample of low-accountability system, which includes 8 countries with 2,131 schools and 62,689 students, and for the sample of high-accountability system, which contains 6 countries with 2,846 schools and 66,909 students are presented.

Table 1. Descriptive statistics.

Variable	Full Sample		Low-accountability system		High-accountability system	
	Mean	SD	Mean	SD	Mean	SD
Level 1	(n=314,327)		(n=62,689)		(n=66,909)	
Math	475.6	101.2	498.8	92.5	455.3	94.8
Male	0.50	0.50	0.50	0.50	0.50	0.50
SES	-0.31	1.13	-0.10	0.98	-0.49	1.23
Self-efficacy	-.04	.97	.06	.96	-.10	.94
Level 2	(n=12,183)		(n=2,131)		(n=2,846)	
Public	0.81	0.39	0.80	0.40	0.74	0.44
SES_mean	-0.38	0.84	-0.08	0.54	-0.58	1.00
TCMORALE	-0.09	1.00	-0.10	0.97	0.00	1.00
TMONITOR	2.09	0.94	1.48	1.00	2.39	0.79
PAREXPEC	1.82	0.71	1.53	0.65	1.95	0.72
PARPART	11.80	12.5	10.08	8.64	14.34	16.05
LEADCOM	0.11	1.03	-0.41	0.97	0.34	1.00
LEADINST	0.01	1.01	-0.36	1.00	0.08	1.03
LEADPD	0.08	1.01	-0.12	0.95	0.10	1.02
LEADTCH	0.08	1.02	-0.14	0.94	0.13	1.10
CONSEQ	2.02	0.69	1.72	0.66	2.11	0.65
Level 3	(n=44)		(n=8)		(n=6)	
NATIONTE	0.82	0.39	0.00	0.00	1.00	0.00
REPORTIN	0.32	0.47	0.00	0.00	1.00	0.00
SANREW	0.14	0.35	0.00	0.00	1.00	0.00
LNGDP	10.18	0.58	10.58	0.88	10.06	0.78

At the student level, the average math score is 475.6 for the full sample, whereas it is 498.8 for the low-accountability system sample and 455.3 for the high-accountability system sample. It shows that students' math achievement in the system with strong external accountability measures is lower than that of the low-accountability systems on average. The proportion of male students is about half across all samples. The SES index is -0.31 for the full sample while it is -0.1 for the low-accountability system and -0.49 for the high-accountability system. Students' math self-efficacy in the low-accountability system is 0.06 while it is -0.10 in the high-accountability system.

At the school level, 81 percent of schools are public for the full sample and 80 percent and 74 percent for the low- and high-accountability system respectively. The school mean SES is the lowest for the high-accountability system samples and the highest for the low-accountability system samples. We included 9 internal accountability variables. Teacher morale is not so much different across three sample groups. For teacher monitoring, the high-accountability system countries execute more teacher monitoring (2.39) than the low-accountability system countries (1.48). Parental expectation towards schools concerning academic standards is higher in the high-accountability system (1.95) than in the low-accountability system (1.53). The percentage of parents' participation in school activities was also higher in the high-accountability system (14.34%) than in the low-accountability system (10.08). The high-accountability system shows a higher value in all school leadership factor variables and in the degree of consequences of teacher appraisals.

At the country level, 82% of 44 sample countries (36 systems) implement national testing, and 32% (15 systems) report school performance publicly, and 14% (6 systems) link the test results to rewards and/or sanctions toward schools. Logged GDP of the full sample countries is 10.18 while it is 10.58 and 10.06 for the low- and high-accountability system respectively.

Multilevel analyses of accountability and math achievement

To account for a nested data structure and disentangle the relations between system-level external accountability versus school-level internal accountability and math achievement, 3-level HLM analyses were conducted. The HLM analysis results for the full sample, low-accountability system sample, and high-accountability system sample altogether in Table 2 are presented for comparison.

Table 2. Three-level HLM analysis results.

	Full sample		Low-accountability system		High-accountability system	
	Unconditional model	Conditional model	Unconditional model	Conditional model	Unconditional model	Conditional model
Fixed effects						
Intercept	480.8* (7.05)	475.8* (5.01)	490.5* (9.82)	496.1* (9.20)	473.6* (16.95)	462.9* (14.49)
Level-1 (Student)						
Male		6.0* (1.72)		2.6* (0.76)		8.2* (0.65)
SES		9.4* (1.48)		15.2* (0.49)		6.1* (0.40)
Self-efficacy		32.7* (1.93)		38.0 (0.40)		30.9* (0.37)
Level-2 (School)						
Public		10.0* (3.67)		5.80* (2.30)		5.20* (1.91)
SES_mean		43.1* (5.05)		37.20* (1.78)		33.6* (1.12)
TCMORALE		3.3* (0.61)		1.70* (0.85)		4.70* (0.74)
TMONITOR		-0.55 (0.69)		2.15* (.93)		-2.33* (0.93)
PARPART		0.07 (0.06)		-0.10 (0.09)		0.01 (0.05)
PAREXPEC		2.80* (1.00)		-2.30* (1.12)		4.80* (1.02)
LEADCOM		0.30 (0.85)		-1.50 (1.02)		1.60 (1.06)
LEADINST		-0.23 (0.65)		0.30 (1.07)		-0.90 (1.17)
LEADPD		-4.60* (0.71)		-3.10* (0.92)		-4.80* (0.92)
LEADTCH		1.40* (0.46)		0.6 (1.04)		0.60 (0.95)
CONSEQ		-2.26* (0.68)		-0.2 (1.30)		-1.50 (1.17)
Level-3 (Country)						
LNGDP		8.90 (5.70)		18.0 (13.78)		10.30 (20.26)
NATIONTE		12.2 (12.96)				

REPORTIN	-31.1* (11.70)					
SANREW	17.0 (16.53)					
Variance component						
L1 variance	4,960	3,945	6,104	5,604	4,585	3,694
L2 variance	3,058*	1,080*	2,093*	1,057*	2,775*	937*
L3 variance	2,159*	1,065*	650*	985*	1,939*	1,247*

Notes. Standard errors for regression coefficients are in parenthesis. The analyses with five plausible values are undertaken according to the OECD's guideline (OECD, 2014a, p.147) with the HLM 7 software program.

[^] p < .10 * p < .05

First, in a fully unconditional model for the full sample, there is only one fixed effect of average school mean which was 480.8 and significantly different from zero. In terms of the variance partitioning, the largest percentage (49%) lies between students within schools (i.e., at level 1); a substantial, though smaller, percentage (30%) lies between schools within countries (i.e., at level 2); another portion (21%) lies between countries (i.e., at level 3). The variations between schools and between countries are statistically significant, which justifies the three-level HLM modelling.

In the conditional model for the full sample, only one external accountability policy was statistically significant. That is, public reporting of school performance is negatively associated with student math scores, while national testing and sanctions and/or rewards based on the test results are not related with math achievement. By contrast, many internal accountability measures have statistically significant relations with student math scores. Teacher morale is positively associated with math achievement. Parent's expectation toward schools concerning academic standards is positively related with math achievement. School leadership for promoting instructional improvements and professional development is negatively related with math achievement, while teacher participation in leadership is positively associated with math achievement. Finally, internal consequences of teacher evaluation are negatively related with math achievement. All control variables at the student and school level, i.e., student gender, SES, math self-efficacy, public schools, and school mean SES, have statistically significant relations with math achievement. Only GDP at the country level is not a statistically significant variable. Individual and school mean SES seem to have explained the variances in student achievement.

Second, the results of three-level HLM analyses with a sample of 8 countries with no external accountability policies (low-accountability system), and with a sample of 6 countries with strong accountability policies (high-accountability system) are presented in the 4th and 6th column in Table 2. Among the internal accountability variables, teacher morale is a significant positive predictor of student math achievement consistently across the low- and high-accountability systems. Teacher monitoring is positively related with math achievement in the low-accountability systems, whereas it has a negative relation in the high-accountability systems. Parental expectation toward schools regarding academic standards has a positive association with math achievement in the high-accountability system countries. Just as indicated in the full sample, school leadership for promoting instructional improvements and professional development is negatively related with math achievement in both low- and high-accountability systems. On the other hand, teacher participation in leadership and the degree of consequences of teacher evaluation are not significant predictors of math achievement in the selected low- and high-accountability systems.

The second research question of this study concerns the coefficients of SES. The coefficients of students' SES are 15.2 in the low-accountability systems and 6.1 in the high-accountability systems, which are both statistically significant. In other words, the influence of individual SES on student math achievement is much stronger in the low-accountability systems

than in the high-accountability systems. Also, the influence of schools' mean SES on student achievement is statistically significant in both low- and high-accountability systems, and the coefficients are 37.2 and 33.6, respectively. That is, the influence of schools' mean SES on student math achievement is also stronger in the low-accountability systems.

Discussion

This research examined the relations of system level external accountability and school level internal accountability measures with math achievement drawing on 44 educational systems' PISA 2012 data.

The HLM analysis results indicate that external accountability policies except for public reporting are not significant predictors of student achievement. Public reporting of school performance in nation-wide tests has a negative relationship with math achievement, which is in contrast to the previous finding of Boarini and Ludemann (2009) with the PISA 2006 data of OECD countries. Intuitively, it is hard to explain that public reporting policy has a negative influence on student achievement. Rather, low-performing systems are supposed to adopt public reporting of school performance to stimulate school efforts. The inference is based on the fact that high-accountability systems achieve lower scores and their economic status is lower than low-accountability systems. All in all, external accountability has rather tenuous relations with student achievement.

In contrast to external accountability, internal accountability measures have tighter relations with student achievement. Among others, teacher morale has a significant positive association with student achievement, consistently across all sample groups. This finding confirms the importance of teachers' internal motivation and passion in education (Yi, 2015). Regarding teacher monitoring, the HLM analyses with low- and high-accountability systems present contradictory findings while teacher monitoring has no significant relations with math scores in the full sample. A previous study reported that students in countries with more monitoring of teacher lessons by principals performed better (Woessmann et al., 2007). In the low-accountability systems, higher teacher monitoring is associated with higher student achievement, whereas higher teacher monitoring is linked to lower student achievement in the high-accountability system. In other words, teacher monitoring exercises in the low-accountability systems function well to contribute to student learning. It seems that countries with no external accountability put more importance on school level teacher monitoring that provides direct feedbacks on teachers. This research considered four factor variables of school leadership, among which two variables were found to have a statistically significant relation with math achievement. Oddly enough, school leadership for promoting instructional improvements and professional development (LEADPD) had a consistently negative relation with student achievement across all sample groups. However, when the question items are examined closely, higher values of this variable are likely associated with more student problems and disruptive behaviours, which should be negatively related to student performance. It is notable that teacher participation in school leadership (LEADTCH) has a positive relation with student achievement of the full sample. Recently, distributed leadership has received attention as a determining lever for school reforms (Spillane, 2006). Specifically, an empirical study found significant direct effects of distributed leadership on change in the schools' academic capacity and indirect effects on student growth rates in math (Heck & Hallinger, 2009). Thus, the current study results support the previous study findings. Finally, internal consequences of teacher evaluation have a negative relation with student achievement. It seems that rewarding teachers extrinsically based on teacher appraisals may not be linked to teacher morale, which has a direct positive effect on student performance, and may do harm to educational practices that promote student learning.

The findings of the HLM analyses with sub-samples revealed that equity of student achievement was better ensured under strong external accountability. For example, Korea, which belongs to a high-accountability system in the sample, implemented the National As-

assessment of Educational Achievement (NAEA), public reporting of school performance on the website, and financial incentives linked to school performance for the purpose of equity enhancement (Lee, 2017). The proportion of students who do not reach a basic proficiency level in Korean, English and math at the school level was publicly posted on a website in Korea as of 2012. On the contrary, Koretz (2017) criticizes high-stakes test-based accountability, discussing negative effects such as inappropriate test preparation, score inflation, corruption of ideal of teaching, and widespread cheating. According to him, the most substantial positive effect of test-based accountability has improved math performance, although it does not persist long. In spite of accumulated evidence about the overall negative effects of test-based accountability, predominantly found in the U.S., the rationale of many other countries' adopting test-based accountability could be ensuring educational equity, of which evidence is found in this research.

There are still several limitations in the current research. First, the research used cross-sectional data to examine the multilevel relations between system level external accountability, schools' internal accountability, and student math achievement. Therefore, the results do not represent causal relations between explanatory variables and math achievement and are open to reversed explanations. Some countries with a strong accountability system may have adopted accountability policy measures to address the low level of student achievement. As it is impossible to determine the causal impact of external accountability on student achievement with cross-sectional data, use of longitudinal data is a research direction forward.

Second, internal accountability measures are independently included in the HLM analyses. Originally, the concept of internal accountability represents a coherent construct, which is based on a level of agreement among teachers on the norms, values, and expectations that shape their work (Elmore, 2004, p.134). When the degree of internal alignment of teacher responsibility, collective expectations of stakeholders, i.e., teachers, parents, and principals, and accountability is high, the school supposedly has strong internal accountability. Although this study attempted to materialize the concept of internal accountability, it did not measure one united concept of internal accountability that reflects the original idea.

Third, we separated the full sample into low- and high-accountability systems and conducted multilevel analyses separately to examine how differently internal accountability measures operate and whether equity of achievement varies. To check the robustness of results, it is desirable to apply a multilevel multigroup analysis in a subsequent study.

Finally, this research conducted a cross-country analysis without considering each country's contextual differences. A case study with a few representative countries will contribute to better understanding of accountability in education.

Conclusions

The findings of this research suggest that top-down external accountability may not be as much effective as expected, while school-based internal accountability factors are more conducive to student achievement. In particular, it is of much importance to encourage teachers to sustain high teacher morale and participate in school decision-making. When teachers are intrinsically motivated, they commit themselves to education, which eventually benefit students' learning and growth. As the findings show, both teacher morale and distributed leadership have significantly positive relations with student performance. Principals need to demonstrate distributed leadership in school so that teachers can have self-determined ownership in school decision-making, which may in turn increase teacher morale.

Another important finding is that strong external accountability may be able to contribute to educational equity. The effects of both individual and school SES on student performance, indicated as coefficients, were lower in the high-accountability system than in the low-accountability system. In other words, individual and school SES are weakly related with student achievement in the high-accountability educational system. This finding suggests that external accountability measures might be beneficially utilized to narrow achievement gaps

between low- and high-SES student groups. Therefore, policy makers of each country should consider strengths and weaknesses of external accountability in their own educational contexts.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A5A8012910).

References

- Adams, J. E., & Kirst, M. W. (1999). New demands and concepts for educational accountability: Striving for results in an era of excellence. In J. Murphy & K. L. Seashore (Eds.), *Handbook of research on educational administration*, 2, (pp. 463-487). San Francisco, CA: Jossey-Bass, Inc., Publishers.
- Bishop, J. H. (1998). The effect of curriculum-based external exit exam systems on student achievement. *The Journal of Economic Education*, 29 (2), 171-182.
- Boarini, R., & Ludemann, E. (2009). The role of teacher compensation and selected accountability policies for learning outcomes: An empirical analysis for OECD countries. *OECD Journals: Economic Studies*, 1, 1-20.
- Burgess, S., Wilson, D., & Worth, J. (2013). A natural experiment in school accountability: The impact of school performance information on pupil progress. *Journal of Public Economics*, 106, 57-67.
- Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24 (4), 305-331.
- Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30 (3), 418-446.
- Elmore, R. (2004). *School reform from the inside out: Policy, practice, and performance*. Harvard Education Press.
- Elmore, R., & Fuhrman, S. (2001). Holding schools accountable: is it working? *Phi Delta Kappan*, 83(1), 67-72. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1007&context=gse_pubs.
- Elstad, E. (2009). Schools which are named, shamed and blamed by the media: School accountability in Norway. *Educational Assessment, Evaluation and Accountability*, 21 (2), 173-189.
- Figlio, D. N., & Loeb, S. (2011). School accountability. In E. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (pp. 383-421). Amsterdam: Elsevier Inc.
- Hallinger, P., & Heck, R. H. (1998). Exploring the principal's contribution to school effectiveness: 1980-1995. *School Effectiveness and School Improvement*, 9 (2), 157-191.
- Hanushek, E. A., Link, S., & Woessmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics*, 104, 212-232.
- 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.
- Hanushek, E. A., & Woessmann, L. (2010). *The economics of international differences in educational achievement*. Retrieved from <http://medcontent.metapress.com/index/A65RM03P4874243N.pdf>.
- Heck, R. H., & Hallinger, P. (2009). Assessing the contribution of distributed leadership to school improvement and growth in math achievement. *American Educational Research Journal*, 46 (3), 659-689.
- Jürges, H., Richter, W. F., & Schneider, K. (2005). Teacher quality and incentives: Theoretical and empirical effects of standards on teacher quality. *FinanzArchiv*, 61 (3), 298-326.
- Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, 54 (1), 5-25.
- Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, 16 (4), 91-114.
- Koretz, D. (2017). *The testing charade: Pretending to make schools better*. The University of Chicago Press.
- Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research*, 78 (3), 608-644.

- Lee, J., & Amo, L. C. (2016). International and interstate analyses of student- and school-targeted accountability policy effects. In D. K. Sharpes (Ed.), *Handbook on comparative and international studies in education* (pp.3-21). Charlotte, NC: Information Age Publishing, Inc.
- Lee, J., Liu, X., Amo, L. C., & Wang, W. L. (2014). Multilevel linkages between state standards, teacher standards, and student achievement: Testing external versus internal standards-based education models. *Educational Policy*, 28 (6), 780–811.
- Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990-2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, 34 (2), 209–231.
- Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal*, 41 (4), 797–832.
- Lee, J. H. (2017, August 31). The victims of educational policy changes. *The Munwhailbo*. Retrieved from <http://www.munwha.com>.
- Leithwood, K., & Day, C. (2008). The impact of school leadership on pupil outcomes. *School Leadership & Management*, 28 (1), 1–4.
- Lingard, B. (2010). Policy borrowing, policy learning: testing times in Australian schooling. *Critical Studies in Education*, 51 (2), 129–147.
- Lundahl, C., & Waldow, F. (2009). Standardisation and “quick languages”: The shape-shifting of standardised measurement of pupil achievement in Sweden and Germany. *Comparative Education*, 45 (3), 365–385.
- Meyer, H., Tröhler, D., Labaree, D. F., & Hutt, E. L. (2014). Accountability: Antecedents, power, and processes. *Teachers College Record*, 116 (9), 1–12.
- Møller, J. (2009). School leadership in an age of accountability: Tensions between managerial and professional accountability. *Journal of Educational Change*, 10 (1), 37–46.
- Morris, A. (2011). Student standardized testing. *OECD Education Working Papers*, 65, 1–51.
- Newmann, F., King, M., & Rigdon, M. (1997). Accountability and school performance: Implications from restructuring schools. *Harvard Educational Review*, 67 (1), 41–69.
- O’Day, J. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, 72 (3), 293–329.
- O’Neill, O. (2013). Intelligent accountability in education. *Oxford Review of Education*, 39 (1), 4–16.
- OECD. (2014a). *PISA 2012 Technical Report*. Retrieved from <http://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm>.
- OECD. (2014b). *Databases of PISA 2012* [Data files and Codebooks]. Retrieved from <http://www.oecd.org/pisa/pisaproducts/>.
- Ozga, J. (2013). Accountability as a policy technology: Accounting for education performance in Europe. *International Review of Administrative Sciences*, 79, 292–309.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.
- Sahlberg, P. (2008). Rethinking accountability in a knowledge society. *Journal of Educational Change*, 11 (1), 45–61.
- Schuetz, G., Luedemann, E., West, M. R., & Woessmann, L. (2013). School accountability, autonomy, choice, and the equality of educational opportunities. In M. Windzio (Ed.), *Integration and inequality in educational institutions* (pp. 12–152). Dordrecht: Springer Netherlands.
- Schütz, G., Luedemann, E., Woessmann, L. and West, M. R. (2010). *School accountability, autonomy and choice around the world*. Edward Elgar Publishing.
- Spillane, J. (2006). *Distributed leadership*. San Francisco: Jossey-Bass.
- Sung, Y.-K., & Kang, M. O. (2012). The cultural politics of national testing and test result release policy in South Korea: a critical discourse analysis. *Asia Pacific Journal of Education*, 32 (1), 53–73.
- Vanhoof, J., & Petegem, P. Van. (2007). Matching internal and external evaluation in an era of accountability and school development: Lessons from a Flemish perspective. *Studies in Educational Evaluation*, 33 (2), 101–119.
- Woessmann, L. (2005). The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics*, 13 (2), 143-169.

- Woessmann, L., Lüdemann, E., Schütz, G., & West, M. R. (2007). School accountability, autonomy, choice, and the level of student achievement: International evidence from PISA 2003. *OECD Education Working Papers*, No. 13, OECD Publishing.
- Yi, P. (2015). Do school accountability and autonomy affect PISA achievement? Evidence from South Korea. *KEDI Journal of Educational Policy*, 12 (2), 197–223.

Received: February 21, 2018

Accepted: May 25, 2018

Pilnam Yi

Ph.D., Associate Professor, Department of Education, Hongik University, 94 Wausan-ro, Mapo-gu, Seoul, 04066, South Korea.

E-mail: pilnamyi@hongik.ac.kr

Website: <https://sites.google.com/site/pilnamyisacademichomepage/>

In-soo Shin

Ph.D., Associate Professor, Department of Education, Jeonju University, 303, Cheonjam-ro, Wansan-gu, Jeonju-si, Jeollabuk-do, South Korea.

E-mail: s9065031@jj.ac.kr