ELSEVIER

Studies in Educational Evaluation 30 (2004) 205-223

Studies in Educational Evaluation

www.elsevier.com/stueduc

LINKAGE BETWEEN TEACHER QUALITY, STUDENT ACHIEVEMENT, AND COGNITIVE SKILLS: A RULE-SPACE MODEL[1]

Tao Xin*, Zeyu Xu and Kikumi Tatsuoka*****

*Beijing Normal University, China
**American Institutes for Research, Arlington, Virginia, USA
***Teachers College, Columbia University, New York, USA

Abstract

The topic of teacher credentials and student performance is revisited in an international setting using the TIMSS-99 data. The lack of consistent positive link between credentials and performance can be explained via three routes: measurement problem of "teacher quality" input, measurement problem of "student outcomes", and the production function form that is assumed to link the input and the output. Although there is a small literature focusing on student outcome measurement problems, suggesting the use of cognitive achievements rather than test scores, in most cases those cognitive measures are nothing but math and science scores. This study contributes to the literature by borrowing from the measurement and psychometrics theories to decompose single scores into three categories of cognitive abilities. The hypothesis is that teachers may play a crucial role in the development of some student cognitive skills while not in the others. Using a "rule-space" model, this article identifies three cognitive skills: the process skill, the reading skill and the mathematical think skill. The study finds that: (1) in general teacher credentials have no effect on any type of cognitive skill development as well as on the test score, and (2) the within-teacher variance of student performance is much larger than between-teacher variance in Japan and Korea, whereas the reverse is true in the US and the Netherlands. The phenomenon of "private tutoring" is quoted as an explanation of this pattern.

0191-491X/04/$ - see front matter
doi: 10.1016/j.stueduc.2004.09.002

There are extensive studies on the relationship between school resources, student achievement and adult success (e.g., Burtless, 1996; Card & Krueger, 1992; Case & Deaton; 1999, Cawley, Heckman, Lochner & Vytlacil, 2000; Currie & Thomas, 2001). In the economics of education and education and social policy literature, the production function approach is usually adopted to describe education as a production process comparable to factory manufacturing, with school resources and family support as the input and achievement scores and future earnings as the output. Within the production function framework, educators, parents and students seek to maximize the productivity in transforming input into output by achieving two efficiencies: technical efficiency and allocative efficiency. Technical efficiency abstractly refers to the latest technology available for production, and allocative efficiency is defined in strict economic terms that resources should be allocated to each input to such a point that the marginal cost of that input equalizes its marginal product. We often see three types of studies using the production function approach: the impact of school resources on achievement and cognitive skills, the impact of achievement and cognitive skills on adult success (usually measured by earnings), and the impact of school resources on adult success. Yet how strong the links are is far from conclusive. Despite Card and Krueger's (1992) finding that labor market outcome is related with school resources, the lack of relationship with school resources is more generally true for recent studies of earnings than earlier investigations, while more recent studies have tended to find stronger effects of cognitive skills on earnings (Hanushek, 1997). Typically a one-standard-deviation difference in cognitive scores is associated with about a 3 to 4% difference in earnings, after controlling for race, gender, educational level, and experiences (Levin, 2001).

Background

The relationship between school resources and student performance is an issue of even more controversy. The debate started since the Coleman Report (Coleman et al., 1966), which finds predominant importance of family background and little significance of school factors in deciding school performance using education production function analysis. Teacher salary accounting for the largest portion of school budgets (Belfield & Levin, 2002), teacher quality has been the focus of discussion in school resource efficiency studies. Large amounts of research have been done and several important meta-analyses arrive at very different conclusions on the basis of empirical results (Hanushek, 1997, 1986; Hedges, Laine, & Greenwald, 1994a, 1994b). The *Hanushek versus Hedges* debate is the most famous one that still keeps updating (Hanushek, 2003; Hanushek & Luque, 2003). In 2003 there are also summarizing studies that specifically focus on the effectiveness of teacher quality (Rice, 2003; Wayne & Young, 2003). These latest teacher quality reviews still cannot reach an agreement on the existence of a teacher quality – student performance relationship.

Not being able to find any conclusive positive effect of teacher quality on student performance poses questions for policy makers on the wisdom of spending so much money based on a teacher-pay system that is mainly based on observable credential measures such as teacher's highest degree, experience, previous test scores and teaching certificate. Researchers try to explain the puzzle from three perspectives: measurement of input, measurement of output, and the form of link between the two.

Measurement error biases production function estimates towards zero in regression analysis. Teacher quality is very difficult to quantify. Are qualified teachers, as judged by observable credentials, also likely to be quality teachers? Teacher credentials are the most directly available measures that do not account for teachers' personality, within classroom activity and many other aspects of teaching. In fact, teacher quality is only weakly linked with teacher pay, which in turn is based on observable and quantifiable teacher attributes (Ballou & Podgursky, 2000). A direct examination of principal's teacher rating, which is mainly based on teacher credentials, and students' rating of teachers finds very low correlations (Murnane & Cohen, 1986).[2] The measurement of output is also problematic. Most individual level studies use test scores as the only outcome. Scores are easily available. They are an important output of schools, and they can somewhat predict students' future labor market success. Therefore score is an important and very realistic approximation of the quality of education. But score is not the whole story: Score is too narrow a measure of outcome because there are other abilities that cannot be measured by academic testing. Score is also too broad because it embodies many latent cognitive abilities that work together to produce the score. Finally, there are studies trying to explain the non-existence of resource effect on student performance by exploring the production function form that relates input to output. The format of education production function is unknown. In most empirical analysis production function is assumed to take linear additive format. Some researchers suggested more flexible functional forms like the transcendental logarithm function and found significant school resource effect (Figlio, 1999).

Existing studies traditionally focus on cognitive development as a measure of output (e.g., Hanushek, 1997; Murnane, 1975; Summers & Wolfe, 1977). Cognitive development, in turn, is usually measured by test scores. Using scores as the dependent variable, current studies reached mixed conclusions on the effect of teacher's degree level and major field, experience, and certificate status. Degree levels are found to have both significant positive effect (e.g., Ferguson & Ladd, 1996) and significant negative effect (e.g., Ehrenberg & Brewer, 1994) on student scores. But in most cases, without controlling for the degree major field, the relationship is indeterminate (Wayne & Young, 2003). With degree major included, however, it is found that having a math-major teacher positively contributes to high school students' math scores (Goldhaber & Brewer, 2000). On the other hand, according to Wayne and Young (2003), findings from existing studies on teacher experience effect are difficult to interpret for many reasons. For example, experience captures the teacher labor market surplus or shortage at the time of hiring, and therefore its effect is hard to generalize. Finally, it is found that if teachers have certification in teaching mathematics, students' math score is on average better than when teachers are not certified or certified in a different subject (Goldhaber & Brewer, 2000). When teacher certification is examined without controlling for certified field, no determinate effect can be detected.

This study contributes to the existing teacher quality literature by focusing on improving the measurement of student outcomes. It is admitted that measuring teacher quality by credentials is very crude. But credentials are feasible features based on which policy makers and administrators can make large-scale employment decisions and payment plans. By watching more closely at, say, classroom activities to judge teacher quality may have no realistic applications beyond experiments and academic discussions. Governments at different levels still have to rely on teacher's degree, teaching certificate, experience and

other credentials to make administrative and policy decisions. Therefore, this study decides to use these traditional measures of teacher quality. On the other hand, even though many studies claim to use ability as the outcome measure, those ability indices are simply math or science scores (Levin, 2001). TIMSS tests students on different fields of knowledge and various problem-solving capabilities using a variety of question formats. This offers us an opportunity to apply a "Rule-Space Model" to decompose student scores into achievements in several knowledge fields and ability development categories. Using the 1999 TIMSS data, this study looks for international evidence on teacher quality effect for 7th and 8th graders.

Rule Space Model

Overview of the Rule Space Method

The present analysis uses the Rule Space Methodology (Tatsuoka, 1983, 1985, 1990, 1995, 1997, in press; K. Tatsuoka & M. Tatsuoka, 1987; M. Tatsuoka & K. Tatsuoka, 1989) to diagnose each student in terms of inferred mastery of specific "attributes" (knowledge and subskill components) assumed to underlie test performance. The present work followed the general outline of any Rule Space analysis, as follows. The first step in a Rule Space analysis of a test involves identifying the specific knowledge and subskill attributes assumed to explain performance on the test items. For the TIMSS-R (1999), this stage involved solution of all TIMSS math items by a team of expert raters, plus the analysis of written student protocols. These assumptions are incorporated in the *Q matrix*, which is an n (items) by k (attributes) binary-valued indicator matrix describing the involvement of the attributes in the items. Given a pattern of correct-incorrect item performance by a student, the student's pattern of mastered and non-mastered attributes can then be inferred. This pattern of mastered and non-mastered attributes is referred to as the student's "knowledge state". This part of the RSM analysis involves several steps. To begin with, Boolean algebra is used to generate all possible combinations of attribute patterns and their corresponding binary item patterns from the Q-matrix. Rule Space is a classification space in which the most plausible knowledge state for a student's item pattern is selected by applying a Bayesian decision rule. Diagnosed knowledge states are vectors summarizing combination of attributes mastered and not mastered by students. Results of the RSM are stored in a dataset consisting of an attribute mastery vector for each student. Note that because the outputs of an RSM analysis convert a dataset of specific test item responses for each student into a vector of attribute mastery probabilities, one can merge the item response patterns of several different tests to a single dataset of students by attributes, as long as tests share the same set of attributes. This property of the rule space method is particularly suitable to the sampling design of TIMSS study because the RSM results from several booklets can be merged into a single dataset of students and attributes.

Description of Cognitive Processing Skills and Knowledge Involved in TIMSS-1999 Mathematical Test

Corter and Tatsuoka (2002) have identified 27 cognitive skills and knowledge attributes involved in solving the 164 mathematics items in the TIMSS 1999 tests for

"Population 2" (8th graders), and validated them statistically. They also have validated them with students' protocols. For each form of the TIMSS math subtest, the involvement of the attributes in solution of each item is summarized in a Q-matrix. The Q-matrix is the representation of researchers' hypothetical cognitive model, in which items are coded by the involvement of attributes. There are no restrictions on coding of a Q-matrix, so that an item can be classified as involving any number of attributes. For example, a single item could involve algebra, geometry, and fractions content knowledge.

Table 1: List of Knowledge, Skill, and Process Attributes Derived to Explain Performance on the TIMSS-R (1999) Math Items, and 1995 General Mathematics in Population 2 (8th graders)

Content Attributes

C1	Basic concepts, properties and operations in whole numbers and integers
C2	Basic concepts, properties and operations in fractions and decimals
C3	Basic concepts, properties and operations in elementary algebra
C4	Basic concepts and properties of two-dimensional Geometry
C5	Data, probability, and basic statistics
C6	Using tools to measure (or estimating) length, time, angle, temperature

Process Attributes

P1	Translate/formulate equations and expressions to solve a problem
P2	Computational applications of knowledge in arithmetic and geometry
P3	Judgmental applications of knowledge in arithmetic and geometry
P4	Applying rules in algebra
P5	Logical reasoning—includes case reasoning, deductive thinking skills, if-then, necessary and sufficient, generalization skills
P6	Problem Search; Analytic Thinking, Problem Restructuring and Inductive Thinking
P7	Generating, visualizing and reading Figures and Graphs
P8	Applying and Evaluating Mathematical Correctness
P9	Management of Data and Procedures
P10	Quantitative and Logical Reading

Skill (item type) Attributes

S1	Unit conversion
S2	Apply number properties and relationships; number sense/number line
S3	Using figures, tables, charts and graphs
S4	Approximation/Estimation
S5	Evaluate/Verify/Check Options
S6	Patterns and relationships (be able to apply inductive thinking skills)
S7	Using proportional reasoning
S8	Solving novel or unfamiliar problems
S9	Comparison of two/or more entities (deleted because of low frequencies in each booklet)
S10	Open-ended item, in which an answer is not given
S11	Using words to communicate questions (word problem)

Corter and Tatsuoka have classified the attributes into three categories: Content Knowledge variables, Cognitive Process variables, and Skill/Item Type variables (shown in Table 1).

The *skill* attributes include certain context-specific and format-specific process skills. This third category of skills is deemed necessary for the following reasons. Many skills in arithmetic and mathematics are associated closely with item types. For example, skill attribute S3 involves reading data or relationships from graphs and figures provided in the actual math item. As another example, proportional reasoning (skill S7) is a process variable, but this type of reasoning is often associated with a particular format of items. In contrast, the *process* attributes are those skills that are used across a wider array of item types. For example, logical reasoning (process attribute P5) is used in many types of items and encompasses several varieties of logical inference.

Data and Variables

Four countries used in the present study are the United States, Korea, Japan, and the Netherlands. The reason for selecting these four countries is that teachers involved the TIMSS 99-R test in these countries have very similar characteristics. Specifically, most of teachers in these four countries have either a bachelor's degree or a master's degree. The detailed information about teacher sampling is listed in Tables 2 and 3. Students in this study are primarily from the 8th grade. The data from only four out of eight different test forms used in the TIMSS-R (1999) are used in the RSM analyses. For these analyses, Booklets 1, 3, 5, and 7 are selected because the other forms (Booklets 2, 4, 6 & 8) show an uneven distribution of the attributes (that is, few or no items measuring certain attributes). The description about student sample size is listed in Table 4.

Based on the literature review, five teacher variables included in the original TIMSS-R data are used as the indictors of teachers' qualification: teachers' educational level, years of experiences, whether or not his/her major of study is mathematics, whether or not his/her major of study is education, and whether or not he/she is qualified for teaching.

In this study, six student background variables are used as control variables: student's gender, age, parents' educational level, index of Home Educational Resources, index of Students' Self-Concept in Mathematics, and index of Positive Attitudes towards Mathematics. Among the six student variables, the last three variables are derived variables combined by TIMSS (Gonzalez & Miles, 2001) from the original items included in the TIMSS 1999 Student Questionnaires. Based on the TIMSS user's guide, all these kinds of derived variables are scaled into three categories: high, medium, and low. The high level of a derived variable corresponds to conditions or activities that generally associated with higher academic achievement (see the Appendix for definitions of derived variables). In this study, we did not control student background variables at school level, because all student variables used in this article were derived from the original items, and there is no information about how such variables are derived in the TIMSS user's guide. So it may cause problems to control student background variables at school level.

RSM is a good approach to transform students' achievement score into a set of attribute mastery probability scores. However, one limitation of the attribute mastery

probability score is that such score does not distribute normally. Based on Tatsuoka's rationale, one way of overcoming this problem is to develop subscales. In the present study, we use three subscales created by Tatsuoka et al. (Tatsuoka, Corter, & Tatsuoka, 2003) as measures of students' cognitive abilities:

Sub-scale 1: p1+p2+p3+p4+p5+p6+p7+p9+p10 (process skills)
Sub-scale 2: s11+p1+p10. (reading skills)
Sub-scale 3: s6+p3+p5+p6 (higher level mathematical thinking skills)

The reason for selecting these three subscales is that they are basic skills for student mathematics learning process. In order to validate these subscales, three experts on mathematical instruction were asked to assess three measures based on the TIMSS-R Mathematical Attributes List. They reached a high agreement to Tatsuoka's compositions. We also use the scaled achievement score that is included in the original TIMSS-R data as the indictor of students' overall achievement.

Results

Table 2: Descriptive Statistics of Teacher Qualification (categorical type)

Variables	USA	Korea	Japan	Netherlands
Educational level				
Under secondary school				
Secondary school				2
BA or equivalent	157	166	104	102
MA/PhD	162	27	4	8
Other post-secondary				
Qualification				
Yes	-	185	118	86
No		7	9	15
Major: Mathematics				
Yes	149	107	115	82
No	283	85	11	26
Major: Education				
Yes	57	8	5	29
No	275	184	122	84
Sample size	392	193	146	126

Table 3: Descriptive Statistics of Teacher Qualification (numeric type)

Countries	Years of experience		Years of pre-training	
	Mean	*SD*	Mean	*SD*
USA	15.88	10.08	1.64	1.56
Korea	12.65	8.36	3.71	1.68
Japan	14.44	8.03	3.27	1.38
Netherlands	16.42	10.27	4.59	1.71

From the two descriptive tables, Tables 2 and 3, several patterns loom obvious. All the teachers in these four countries have a Bachelor's or Master's degree (except for two cases in the Netherlands). These teachers are mostly certified teachers, but the certification types are not identified. The US shows a very different pattern in teacher major composition. The math and science teachers in the US are dominantly non-math majors, while in the other three countries math and science are mostly taught by professional math graduates, especially in Japan. The US also has a larger proportion of math and science teachers with an education major background than the other three countries. All the four countries have very experienced teachers, with the Netherlands teachers receiving the longest pre-training.

Table 4: Sample Sizes for Booklets in Selected Countries

Country	Booklet 1	Booklet 3	Booklet 5	Booklet 7	Total
Japan	584 (585)	589 (590)	597 (599)	601 (601)	2371
Korea	758 (766)	763 (765)	761 (762)	763 (765)	3045
Netherlands	377 (379)	369 (369)	365 (367)	369 (370)	1480
United States	1104 (1132)	1110 (1144)	1117 (1147)	1080 (1100)	4411

In Table 5 we present the estimated between- and within-teacher variances of student achievement and abilities. The variance decomposition is achieved by using an "intercept-only" hierarchical linear model (HLM). A very clear pattern emerges. In the US and the Netherlands, between-teacher student achievement variance is greater than within-teacher variance. More than 50% and 66% of the variances are accounted for by between teacher variances for the US and the Netherlands respectively. In other words, even though student performance significantly varies within the same class taught by the same teacher, teacher differences added more performance gaps between students from different classes. The within-teacher variance is due to factors other than teacher quality.

Table 5: Estimated Between- and Within-Teacher Variances of Student Achievement and Abilities Among Four Countries

Dependent variables	USA		Japan		Netherlands		Korea	
	Between	Within	Between	Within	Between	Within	Between	Within
Achievement	4255.11	3688.68	827.61	5524.73	3333.35	1457.24	780.66	5497.11
	(347.81)	(80.84)	(108.50)	(161.31)	(440.38)	(55.93)	(129.74)	(144.39)
Process skill	.59(.05)	.88(.02)	.04(.01)	.72(.02)	.53(.07)	.39(.02)	.09(.01)	.66(.02)
Reading skill	.05(.00)	.15(.00)	.01(.00)	.15(.00)	.02(.00)	.07(.00)	.00(.00)	.08(.00)
Math thinking	.22(.02)	.32(.01)	.01(.00)	.19(.01)	.20(.03)	.20(.01)	.03(.01)	.35(.01)

Note: Standard errors in the brackets.

Table 6: Comparison of Multilevel Coefficients of Teacher Qualifications on Students' Achievement Among Four Countries

Type of effect	USA		Japan		Netherlands		Korea	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed effect								
Intercept	452.89	28.44	204.98	78.79	553.04	42.16	383.33	54.72
Student background								
Gender	6.73***	1.99	3.71	3.16	-2.67	2.32	-2.45	3.02
Age	-6.90***	1.81	19.98***	5.32	-4.74*	2.31	-.42	3.61
Parents edu. level	-2.41**	.92	-	-	-.53	.85	-8.37***	1.29
Home Resources	14.61***	2.60	-	-	1.94	4.25	37.46***	3.39
Math self-concept	35.22***	1.79	36.21***	3.85	24.59***	2.24	43.59***	3.44
Math attitude	4.30**	1.62	30.59***	2.82	3.51	2.03	25.29***	2.25
Teacher qualifications								
Years of teaching	.77*	.35	.51	.30	.09	.54	-.24	.22
BA vs. Master	-.94	7.01	-51.81***	14.25	-8.23	20.89	9.05	4.95
Major: Math	12.39	6.60	-22.99	17.96	3.94	13.21	-3.13	3.66
Major: Education	5.84	8.87	15.48	11.60	-15.63	12.47	12.51	9.02
Certificate	-	-	22.49	20.14	-	-	17.19	9.37
Random effect	V.C.	SE	V.C.	SE	V.C.	SE	V.C.	SE
Between teacher	2974.69	283.06	328.18	79.97	3085.82	454.57	280.68	67.31
Within teacher	3186.31	79.05	4874.36	155.89	1331.48	57.94	4480.51	118.60

Notes: V.C. stands for Variance Component; * $p < .05$; ** $p < .01$; *** $p < .001$. – denotes that this variable is not included in the particular country.

Table 7: Comparison of Multilevel Coefficients of Teacher Qualifications on Students' Process Skill Among Four Countries

Type of effect	USA		Japan		Netherlands		Korea	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed effect								
Intercept	7.425	.446	5.105	.911	9.390	.638	6.847	.623
Student background								
Gender	.082*	.032	.074*	.037	-.068	.038	-.024	.034
Age	-.137***	.029	.150*	.062	-.167***	.038	-..024	.041
Parents edu. level	-.032*	.015	-	-	-.012	.014	-.071***	.015
Home resources	.151***	.042	-	-	.059	.070	.320***	.039
Math self-concept	.435***	.029	.302***	.045	.282***	.037	.346***	.039
Math attitude	.053*	.026	.291***	.033	.041	.033	.229***	.026
Teacher qualifications								
Years of teaching	.007	.004	.003	.003	.002	.009	-.003	.002
BA vs. Master	-.010	.085	-.446**	.141	-.119	.258	.087	.055
Major: Math	.168*	.802	-.252	.179	.045	.164	-.031	.041
Major: Education	.021	.108	.161	.116	-.172	.154	.134	.100
Certificate	-	-	.289	.201	-	-	.078	.104
Random effect	V.C.	SE	V.C.	SE	V.C.	SE	V.C.	SE
Between teacher	.392	.041	.022	.008	.457	.071	.033	.008
Within teacher	.812	.020	.666	.021	.361	.016	.581	.015

Notes: V.C. stands for Variance Component; * $p < .05$; ** $p < .01$; *** $p < .001$. – denotes that this variable is not included in the particular country.

Table 8: Comparison of Multilevel Coefficients of Teacher Qualifications on Students' Reading Skill Among Four Countries

Type of effect	USA		Japan		Netherlands		Korea	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed effect								
Intercept	3.158	.181	1.109	.424	3.381	.248	2.616	.227
Student background								
Gender	-.003	.013	.040*	.017	.001	.016	-.014	.012
Age	-.058***	.012	.091**	.029	-.050**	.016	-..017	.015
Parents edu. level	-.019**	.006	-	-	-.005	.006	-.017**	.005
Home Resources	.037*	.017	-	-	.002	.029	.078***	.014
Math self-concept	.117***	.012	.114***	.021	.066***	.016	.094***	.014
Math attitude	.007	.011	.114***	.015	-.000	.014	.050***	.009
Teacher qualifications								
Years of teaching	.002	.001	.001	.001	-.000	.001	-.001	.001
BA vs. Master	.037	.025	-.178**	.062	-.015	.063	.046**	.017
Major: Math	.034	.024	-.119	.079	.035	.040	.000	.013
Major: Education	.009	.033	.034	.051	-.058	.038	.019	.032
Certificate	-	-	.096	.089	-	-	.024	.033
Random effect	V.C.	SE	V.C.	SE	V.C.	SE	V.C.	SE
Between teacher	.030	.004	.004	.002	.032	.005	.002	.001
Within teacher	.141	.004	.145	.005	.066	.003	.079	.002

Notes: V.C. stands for Variance Component; * p< .05; ** $p < .01$; *** $p < .001$. – denotes that this variable is not included in the particular country.

Table 9: Comparison of Multilevel Coefficients of Teacher Qualifications on Students' Mathematical Thinking Skill Among Four Countries

Type of effect	USA		Japan		Netherlands		Korea	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed effect								
Intercept	2.459	.304	2.277	.545	3.823	.513	2.000	.505
Student background								
Gender	.038	.022	.012	.022	-.070*	.033	.022	.026
Age	-.059**	.020	.061	.037	-.075*	.032	-..005	.034
Parents edu. level	-.024*	.011	-	-	-.033**	.012	-.050***	.012
Home resources	.156***	.029	-	-	.033	.059	.248***	.032
Math self-concept	.273***	.020	.143***	.027	.181***	.031	.243***	.032
Math attitude	.011	.018	.136***	.020	.013	.028	.186***	.021
Teacher qualifications								
Years of teaching	.005	.003	.001	.002	.001	.004	-.001	.002
BA vs. Master	-.033	.052	-.243**	.081	.021	.016	.059	.040
Major: Math	.088	.049	-.235*	.104	.039	.101	.001	.029
Major: Education	.030	.066	.100	.066	-.091	.095	.136	.073
Certificate	-	-	.265*	.117	-	-	.111	.075
Random effect	V.C.	SE	V.C.	SE	V.C.	SE	V.C.	SE
Between teacher	.137	.016	.006	.003	.161	.027	.011	.004
Within teacher	.295	.009	.186	.007	.194	.010	.302	.009

Notes: V.C. stands for Variance Component; * $p< .05$; ** $p< .01$; *** $p<.001$. – denotes that this variable is not included in the particular country.

Those factors often include family incomes, parent education levels and other individual characteristics. The between-teacher variance, however, cannot be cleanly attributed to teacher quality differences alone. The average academic readiness when students enrolled into classes and their SES might be related to teacher selections. Therefore, large between-teacher variance might indicate that teacher quality matters, or it might imply that student attributes are more heterogeneous between classrooms than within classrooms.

Very interestingly, the opposite pattern is found with the two East Asian countries. There, between-teacher variance is much smaller than within-teacher variance. The between variance only accounts for about 1/8 of the total variance while 7/8 of the variance happens among students taught by the same teacher. Assuming that teachers have similar student bodies that are the same in both individual and family attributes, such a between-within variance ration indicates teacher plays an ignorable role in the school. One theory that can potentially explain such a patter in Japan and Korea can be found within the private tutoring literature (Bray, 1999; Bray & Kwok, 2003; Shafiq, 2002). Studies have found that in many East Asian schools, teachers are responsible for teaching the very basic knowledge and skills. What decides student performance is after-school activities. A very high percentage of family education expenditure is spent on private tutoring. Russell (1997) found that nearly 70% of all students had received tutoring by the time they had completed middle school. In Seoul, 82% of elementary, 66% of middle and 59% of academic high school students received tutoring (Paik, 1998). Private tutoring is most intensive in Asia, Africa, Eastern Europe and Latin America, regardless of the per capita GDP of those countries, and the least intensive in Western Europe, North America and Australia (Bray, 1999). By comparison, for example, only 25% of the US students have attended private tutoring outside of school (OECD: 2000). In terms of private tutoring expenditures, Japan spent an equivalence of $14,000 million in the mid-1990s annually (Russell, 1997, p. 153), and the Republic of Korea is reported to have spent $25,000 million on private tutoring in 1996, an equivalence of 150% of the government's budget (Bray, 1999).

Table 6 shows the estimates of a multilevel model examining the achievement scores. Tables 7-9 show the results using decomposed ability categories as the dependent variable. We added teacher attributes to explain between-teacher variance. In general, between-teacher variance reduced minimally after teacher quality is controlled. Teaching experience, teacher's level of education, certification status and major do not have effect on student math and science achievement scores. They also have no relation with process skills, reading skills, or mathematical thinking skills development. Years of teaching has positive significant impact on achievement scores for the US, but not for the three decomposed cognitive development indices separately. Teacher's highest level of education has strong positive impact in the Japanese case for the achievement scores as well as all the cognitive achievements. As pointed out in Wayne and Young (2003), teacher experience is a variable that is very hard to generalize. It is not only affected by the teacher labor market condition, but also by teacher motivation and personal time constraint. Our insignificant result on experience is consistent with those authors' concerns. Different from Goldhaber and Brewer's (2000) findings, teachers majoring in math do not have a significant edge in helping their students to obtain higher mathematical skills as compared with teachers with other majors. However, our results are not necessarily contradictory with earlier findings.

After all, the Goldhaber-Brewer study focuses on high school students. While focusing on elementary students, Eberts and Stone (1984) find no association between teacher's major field and student scores in that field. Finally, the insignificant relation between teacher certification and student skills is at least partially attributable to the lack of information on which subject the teacher is certified to teach.

When family and student background variables are added into the first level model (to explain the within variance), however, the between-teacher variance drops dramatically. This phenomenon implies the possibility of teacher selection: higher SES families choose or are matched with better teachers.

Conclusions

The current study is highly exploratory. Three problems should be kept in mind while interpreting the results. First, due to the limitations of the data, we cannot use a "value added" model. A value added model takes the previous period student performance into consideration, generating student performance improvement measures that correspond to the resource investment of that period. Within the TIMSS-99 data, we cannot control for student performance at a starting time point. Indeed, among the enormous amount of empirical studies, Wayne and Young (2003) were only capable of identifying 21 US studies on student achievement that adequately controlled for prior achievement scores. The TIMSS scores and their decomposed ability indices hence cannot be cleanly linked to the characteristics of current schools and current teachers. In addition, if better performing students in the previous time period are assigned to better teachers in the current period, then the error term, with prior performance included in it, would be correlated with all the teacher quality variables and thus bias the estimates. This is a typical simultaneity problem. Without imposing additional exclusion restrictions, the model is not identified. If good students are assigned to good teachers, the bias is upward. But even with such an upward bias we could not find positive correlations between teacher quality and achievement, which should furnish evidence against the existence of any teacher quality effect as measured by teacher's degree level, major field, certification status and experience.

A second caution is that uncontrolled family ground variables might bias the estimates. We only examine the within-country comparison and not the cross-country variance. So country specific unobserved variables like policy environment pose no problems. But ignorance of family background usually leads to biased and inconsistent OLS estimates because high SES families are more likely to choose better teachers based on previous empirical results. Therefore the error and the regressors are correlated. As TIMSS-99 does not provide adequate family background information, the SES controls in our model may not be enough. However, as reasoned above, the bias is upward. So this problem actually strengthened the conclusion that teacher credentials have no impact on achievement.

Finally, different countries have different teacher qualification requirements. Same level of education means very wide knowledge gap among countries. In addition many questions and choices provided for answering those questions in the TIMSS questionnaire are not well defined. For example, the teacher major question does not specify whether the

major should be at the B.A. or at the Master's level. Therefore any direct comparison between countries should be very careful.

Despite the problem of not being able to adequately control for individual and family background variances, the estimates of the multi-level model strongly suggest that teacher attributes that are usually used in employment and pay decisions have no impact on science and math achievements in the four countries. After decomposing the test score into three categories of cognitive skills using the RSM, we found teacher attributes also have no consistent positive impact on any type of cognitive skills. This result, building on previous evidence using test scores, further argues against the wisdom of using teacher credentials, such as degrees or certificates, as the selection standards in the teacher market.

Notes

1. This study has been supported by the National Science Foundation (Rec. No. 0126064).

2. There are attempts to include new teacher quality measures besides credentials. For example, based on Epstein's "separate spheres" theory, teacher effort in establishing a good school family relationship is included in the analysis. Such effort is found to have significant positive impact on student performance (Xu & Gulosino, 2004).

References

Ballou, D., & Podgursky, S. (2000). Reforming teacher preparation and licensing: hat is the evidence? *Teachers College Record, 102*, 5-27.

Belfield, C., & Levin, H.M. (2002). The economics of education on judgment day. *Journal of Educational Finance, 28* (2), 183-206.

Bray, M. (1999). *The shadow education system: Private tutoring and its implications for planners.* Paris: UNESCO International Institute for Educational Planning.

Bray, M., & Kwok, P. (2003). Demand for private supplementary tutoring: Conceptual considerations, and socio-economic patters in Hong Kong. *Economics of Education Review. 22* (6). 611-620.

Burtless, G. (1996). *Does money matter? The effect of school resources on student achievement and adult success.* Washington, DC: Brookings.

Card, D., & Krueger, A.B. (1992). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy, 100* (1), 1-40.

Case, A., & Deaton, A. (1999). School inputs and educational outcomes in South Africa. *Quarterly Journal of Economics, 114* (3), 1047-1894.

Cawley, J., Heckman, J.J., Lochner, L., & Vytlacil, E. (2000). Understanding the role of cognitive ability in accounting for the recent rise in the economic return to education. In K. Arrow, S. Bowles & S. Durlauf (Eds.), *Meritocracy and economic inequality* (pp. 230-265). Princeton, NJ: Princeton University Press.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., et al. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Corter, J.E., & Tatsuoka, K.K. (2002). *Diagnostic assessments for mathematics tests grades 6-12.* Unpublished manuscript, Teachers College, Columbia University, New York, NY.

Currie, J., & Thomas, D. (2001). Early test scores, school quality and SES: Longrun effects on wage and employment outcomes. *Worker wellbeing in a changing labor market, 20*, 103-132.

Eberts, R.W., & Stone, J.A. (1984). *Unions and public schools: The effect of collective bargaining on American education.* Lexington, MA: D.C. Health.

Ehrenberg, R.G., & Brewer, D.J. (1994). Do school and teacher characteristics matter? Evidence from high school and beyond. *Economics of Education Review, 13* (1), 1-17.

Ferguson, R.F., & Ladd, H.F. (1996). How and why money matters: An analysis of Alabama schools. In H.F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 265-298). Washington, DC: Brookings Institution.

Figlio, D.N. (1999). Functional form and the estimated effects of school resources. *Economics of Education Review, 18*, 241-252.

Goldhaber, D.D., & Brewer, D.J. (2000). Does teacher certification matter? High school certification status and student achievement. *Education Evaluation and Policy Analysis, 22* (2), 129-146.

Gonzalez, E.J., & Miles, J.A. (2001). *TIMSS 1999 user guide for the international database.* International Study Center, Lynch School of Education, Boston College.

Hanushek, E.A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*, 1141-1177.

Hanushek, E.A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19* (2), 141-164.

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

Hanushek, E.A., & Luque, J.A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review, 22*, 481-502.

Hedges, L.V., Laine, R.D., & Greenwald, R. (1994a). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher, 23* (3), 5-14.

Hedges, L.V., Laine, R.D., & Greenwald, R. (1994b). Money does matter somewhere: A reply to Hanushek. *Educational Researcher, 23* (4), 9-10.

Levin, H.M. (2001). High-stakes testing and economic productivity. In G. Orfield & M.L. Kornhaber (Eds.), *Raising standards or raising barriers?* New York: The Century Foundation Press.

Mullis, I.V.S., Martin, M.O., Gonzales, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 International Mathematics Report.* Chestnut Hill, MA: International Study Center, Boston College.

Murnane. R. (1975). *The impact of school resources on the learning of inner-city children.* Cambridge, MA: Ballinger.

Murnane, R.J., & Cohen, D.K. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review, 56* (1), 1-17.

OECD. (2000). *Additional instruction time and learning time of 15-year-olds*. Paris: Organization for Economic Cooperation and Development.

Paik, S.J. (1998). Personal communication, citing Yim, Y.G. 1997. *1997 survey on current educational issues*. Seoul: Korean Educational Development Institute.

Russell, N.U. (1997). Lessons from Japanese cram schools. In W.K. Cummings & P. Altbach (Eds.). *The challenge of Eastern-Asian education: Lessons for America* (pp. 153-170). Albany: State University of New York Press.

Rice, J.K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes*. Washington, DC: Economic Policy Institute.

Shafiq, M.N. (2002). *The economic burden of private tutoring expenses on households in developing countries: The case of Bangladesh*. Paper presented at the 46th Comparative and International Education Society's Annual Meeting. Orlando, FL.

Summers, A., & Wolfe, B. (1977). Do schools make a difference? *American Economic Review, 67,* 639-652.

Tatsuoka, K.K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 34-38.

Tatsuoka, K.K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12* (1), 55-73.

Tatsuoka, K.K. (1990). Toward an integration of item response theory and cognitive analysis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.C. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 543-588). Hillsdale, NJ: Erlbaum.

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern classification approach. In P. Nichol, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-360). Hillsdale, NJ: Erlbaum.

Tatsuoka, K.K. (1997). Use of generalized person-fit indices for statistical pattern classification. An invited paper for a special issue of person-fit statistics. *Journal of Applied Educational Measurement, 9* (1), 65-75.

Tatsuoka, K.K. (in press). *Statistical pattern recognition and classification of latent knowledge states: Cognitively diagnostic assessment*. Mahwah, NJ: Erlbaum.

Tatsuoka, K.K., Corter, J., & Tatsuoka, C. (2003). *Exploring mathematical thinking skills in TIMSS for twenty countries: Application of the rule space method*. Unpublished manuscript, Teachers College, Columbia University, New York, NY.

Tatsuoka, K.K., & Tatsuoka, M.M. (1987). Bug distribution and pattern classification. *Psychometrika, 52* (2), 193-206.

Tatsuoka, M.M., & Tatsuoka, K.K. (1989). Rule space. In Kotz and Johnson (Eds.), *Encyclopedia of statistical sciences*. New York: Wiley.

Wayne, A.J., & Young, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73* (1), 89-122.

Xu, Z., & Gulosino, C. (2004). *How does teacher quality matter? Effect of teacher-parent partnership on early childhood performance in public and private schools.* Paper presented at the 2004 AERA conference, San Diego, CA.

The Authors

TAO XIN is a professor in the Institute of Developmental Psychology, Beijing Normal University, China. His research interests focus on educational testing, and how schooling effects student's development.

ZEYU XU is a research analyst in the American Institutes for Research, Washington, D.C. He is an education economist, with a focus on household economics and applied econometrics.

KIKUMI TATSUOKA is a research professor in Measurement, Evaluation and Statistics at Teachers College, Columbia University. Her research focuses on Educational Testing theory and Cognitive diagnosis. She is the founder of the Rule-Space Model.

Correspondence: <zx20@columbia.edu>

Appendix: Definition of Composite Variables in TIMSS-99

Index of Home Educational Resources. Index based on students' responses to three questions about home educational resources: number of books in the home; educational aids in the home (computer, study desk/table for own use, dictionary); parents' education. High level indicates more than 100 books in the home; all three educational aids; and either parent's highest level of education is finished university. Low level indicates 25 or fewer books in the home; not all three educational aids; and both parents' highest level of education is some secondary or less or is not known. Medium level includes all other possible combinations of responses. Response categories were defined by each country to conform to their own educational system and may not be strictly comparable across countries.

Index of Students' Self-Concept in Mathematics. Index based on students' responses to five statements about their mathematics ability: 1) I would like mathematics much more if it were not so difficult; 2) Although I do my best, mathematics is more difficult for me than for many of my classmates; 3) Nobody can be good in every subject, and I am just not talented in mathematics; 4) Sometimes, when I do not understand a new topic in mathematics initially, I know that I will never really understand it; 5) Mathematics is not one of my strengths. High level indicates student disagrees or strongly disagrees with all five statements. Low level indicates student agrees or strongly agrees with all five statements. Medium level includes all other possible combinations of responses.

Index of Positive Attitudes towards Mathematics. Index based on students' responses to five statements about mathematics: 1) I like mathematics; 2) I enjoy learning mathematics; 3) Mathematics is boring (reversed scale); 4) Mathematics is important to everyone's life; 5) I would like a job that involved using mathematics. Average is computed across the five items based on a 4-point scale: 1 = *strongly* negative; 2 = *negative*; 3 = *positive*; 4 = *strongly positive*. High level indicates average is greater than 3. Medium level indicates average is greater than 2 and less than or equal to 3. Low level indicates average is less than or equal to 2.