International and Interstate Gaps in Value-Added Math Achievement: Multilevel Instrumental Variable Analysis of Age Effect and Grade Effect

JAEKYUNG LEE
State University of New York at Buffalo

REVA M. FISH
State University College at Buffalo

Merging National Assessment of Educational Progress and Trends in International Math and Science Study math data, this study examines the extent and sources of variation in value-added academic growth patterns among different nations and U.S. states. Applying hierarchical linear models with an instrumental variable method, the study disentangles age effect and grade effect on growth and accounts for both international and interstate variations. Math achievement gains between grades 4 and 8 are significant with positive age and grade effects and negative age-by-grade interaction. The study finds convergence of the age effect and divergence of the grade effect among the nations and states. While the United States lags further behind East Asian nations in math achievement during middle school years, a few of the U.S. states match East Asian gains. The initial or widening gap between the U.S. states and East Asian countries is more closely related to school factors, whereas family factors play a greater explanatory role for achievement gaps among the U.S. states.

Research suggests that a key element in the human capital of a nation is the quality of its schools as measured by math and science skills (Barro 2001; Hanushek and Kimko 2000; Hanushek and Woessman 2007). Since math and science achievement gaps loomed large between students in the United States and other industrial countries, particularly East Asian countries, school reform movements during the past two decades were driven partly by international math and science test results (Baker 2003; Lee 2001). According to the results of 2003 and 2007 Trends in International Math and Science Study...
(TIMSS) math assessments, U.S. fourth graders and eighth graders scored above the international average, but they were consistently outperformed by participating Asian countries, including Chinese Taipei, Hong Kong, Japan, Korea, and Singapore (Mullis et al. 2004, 2008).

Previous studies mostly focused on snapshot comparisons of national average math achievement at single time points by using cross-sectional international test data. However, a few studies conducted quasi-longitudinal analysis of same cohort groups’ math achievement gains, revealing cross-national variations in academic progress (Gonzales 2001; Mullis et al. 2008). For example, among 17 nations that participated in both 1995 fourth-grade TIMSS and 1999 eighth-grade TIMSS, there were substantial variations in the amount of relative math achievement gain: some countries improved their relative performance significantly (Canada, Hong Kong, Iran) or maintained their relative status (Australia, Korea, Japan, England); whereas others dropped their relative performance (Czech Republic, Italy, Netherlands, United States). The U.S. students’ relative math performance deteriorates from grade 4 to grade 8. These findings raise further questions: Do these growing gaps indicate deterioration of American schools’ value-added contribution to students’ math achievement? Why do American students lag further behind their peers in other nations during middle-level schooling?

American educational policy makers, practitioners, and researchers have raised concerns about middle schools’ academic rigor and the effectiveness of activities designed to help early adolescents develop in nonacademic realms (Alt et al. 2000; Carnegie Council on Adolescent Development 1989; McEwin et al. 1996). The targets of criticisms range from school structures to teacher preparation and classroom practices. Some pointed out that fragmented grade configurations might disrupt the continuity of student learning during the transition from elementary schools to middle schools (Coladarci and Hancock 2002). Some criticized the impersonal and fragmented learning environment in middle schools for failing to meet early adolescents’ developmental needs (Eccles et al. 1991). Comparative studies found differences between American and East Asian school systems in terms of middle-level math curriculum and teaching practices (Schmidt et al. 1997; Stigler and Hiebert 1999).
Despite the fact that the United States lags behind East Asian countries in average student math achievement, there remains a controversy about when the gap starts and how much of the achievement gap is attributable to the effects of family and student characteristics as opposed to school effects (Bracey 1999; Stevenson and Stigler 1992; Uttal 1997; Wang and Lin 2005). Previous comparative studies suggested that the math achievement gap between East Asian (e.g., Chinese, Japanese, and Korean) and U.S. students exists even before their school entry and that the gap tends to widen as students move to the upper level of schooling (Stevenson and Stigler 1992; Uttal et al. 1988; Wang and Lin 2005). If we assume that family and other social conditions are mostly responsible for children’s cognitive development before school entry (whereas schools are primarily responsible for academic progress later on), the growing achievement gap might place greater blame on schools rather than families. However, it remains to be examined how student math achievement would grow with schooling versus without schooling, as policy makers are concerned about value-added contributions of schooling to the math achievement of students above and beyond family and other out-of-school influences on their cognitive development and academic progress.

This study also addresses the concern that aggregate international comparisons obscure interstate variations within the United States. Most American states are not only comparable to many countries in size or population, but also each state, like most nations, is in charge of its own educational system, including compulsory education age rules. Studies linking the International Assessment of Educational Progress and National Assessment of Educational Progress (NAEP) mathematics data showed that differences in achievement among the American states were as great as the differences among the countries examined and that there was greater achievement variation within countries than across countries and states (Johnson and Slegendorf 1998; Phelps and Smith 1996; Salganik et al. 1993). The comparison of individual states with other nations in the mathematics proficiency scores of students revealed that the highest-performing states in the United States perform as well as the highest-performing countries, whereas the lowest-performing states in the United States perform as poorly as the lowest-performing countries in the world (Matheson 1996). While prior research examined variations in math achievement growth patterns between the U.S. states (Coley 1998, 2003), there was no comparison of individual states with other nations in terms of value-added school effects on those gains.

The purposes of this study are to examine the nature and sources of variations in nation- and state-level growth of average math achievement, particularly value-added school effects, in accounting for both (1) international and expectations (LeTendre 2000), out-of-school learning (Lee 2007; Paik 2002), and teacher qualifications (Akiba et al. 2007; Mullis et al. 2000).
International and Interstate Gaps

achievement gaps between the United States and other nations and (2) interstate achievement gaps between American states. The study does not address specific issues of American educational policies, but it has potential implications for international policy benchmarking efforts. Since U.S. school reform policy is often aimed at benchmarking the high-performing East Asian education system—a model of common rigorous education standards and test-driven accountability—it needs to be informed by comparative research on underlying global or idiosyncratic mechanisms of academic growth and educational development for achieving high standards.

Method

Analytical Approaches

There are two major analytical approaches to disentangling different sources of influence on student achievement and sorting out school effects to account for international student achievement gaps. One predominant statistical approach is the use of analysis of covariance or multiple regression methods that involve statistical control for observable covariates (Baker et al. 2002; Hanushek and Luque 2003; Harris 2007; Heyneman and Loxley 1983; Lee and Barro 2001). Those correlational studies are designed to examine “relative” school effects (e.g., how much of the achievement gap is explained by observed differences in school-related vs. family-related factors between countries). Despite an array of variables available in large national data sets, this line of research has limitations due to its reliance on observable measures of family and school factors.

Another statistical approach is regression-discontinuity design, a quasi-experimental approach, in order to estimate causal effects of schooling. Previous studies attempted to address the limitations of the correlational approach by tapping into “absolute” school effects on student achievement (i.e., effects of schooling vs. no schooling) that would capture the effects of both observable and unobservable school factors (Cahan and Cohen 1989; Luyten 2006). These studies capitalized on mandatory school entry cutoff date rules in selected countries to decompose academic growth between two adjacent grades into grade effects (the effect of having one more year of schooling at the same age) and age effects (the effect of being a year older in the same grade). They found that the grade effects (allegedly an indicator of schooling effects) were bigger than the age effects—presumably an indicator of any nonschooling effects such as the influences of child and family characteristics (Cahan and Cohen 1989; Luyten 2006). Although this line of quasi-experimental research is stronger than correlational research for drawing causal inferences about
school effects, those previous studies had limited generalizability by restricting
the data range to two consecutive grades in countries with school entrance
age cutoff rules. The previous studies also failed to explain the sources of age
versus grade effects, as they did not account for international or interstate
variations in those effects.3

In light of those concerns, this study builds on a combination of the two
approaches above. It attempts to measure absolute effects of schooling unique
to each nation and state, first, by disentangling math achievement gains into
age effect and grade effect and, then, by exploring the sources of variations
in growth patterns among the nations and states.

Operational Definition of Age Effect and Grade Effect

This study decomposes total math achievement gains (T) from grade 4 to
grade 8 into two components: age effect (A) and grade effect (G). Graphically,
these two effects and their relationships are illustrated in figure 1. If we simply
measure math achievement gain from grade 4 to grade 8, total gain score is
T. Assuming a consistently positive relationship between age and math achieve-
ment across grades, extrapolating the age-achievement regression line from
grade 4 to grade 8 would capture the age effect (A), that is, the portion of
academic growth related to age. The difference between total math achieve-
ment gains (T) and age effect (A) captures grade effect (G), that is, the portion
of growth that occurs with progression from one grade to the next grade after
we partial out the age effect.

The operational definitions of age effect, grade effect, and total gain are as
follows:

Age effect (A) = math achievement gains attributable to age differences.
This age effect estimation would address the counterfactual question of
“How much would children’s math achievement grow if they did not go
to school?”

Grade effect (G) = math achievement gains attributable to grade dif-
ferences. This grade effect estimation taps into the question of “How much
would children’s math achievement grow as a result of their exposure to
schooling during those middle school grades above and beyond the age
effect?”

Total gain (T) = total math achievement gains between grade 4 and
grade 8. This total gain is sum of the age effect and the grade effect above
(T = A + G).

Theoretically, the portion of growth attributable to age effect would capture
growth that would be observed without schooling (e.g., cognitive development
FIG. 1.—Analytical framework for the decomposition of total math achievement gain from grade 4 to grade 8 into age effect and grade effect.

by maturation under home and other out-of-school influences on that process), whereas the portion of growth attributable to grade effect would be something observed with schooling. However, the interpretation of estimated age effect in this study needs to consider the fact that the actual estimation is based on age-related cumulative learning differences by the time of testing that involve time both before and after school entry; these two different periods of age effects have been called “entrance age effect” and “age-at-test effect,” respectively (Elder and Lubotsky 2006). Because it is not possible to identify these two effects separately, the estimate of age effect reported in this study refers to the combined effect of entrance age and the age at the time of the test. Since the grade effect is obtained after the age effect is parsed out from total growth, the interpretation of the grade effect counts on the definition of
the age effect. While the grade effect is supposed to capture value-added achievement gains due to schooling, the grade effect may involve multiple factors that influence learning gains beyond the age effect.

Samples

This study used repeated cross-sectional data by merging samples from TIMSS 1995 fourth-grade and 1999 eighth-grade math assessment data with samples from NAEP 1996 fourth-grade and 2000 eighth-grade math assessment data. Specifically, this merged data set is designed to address the question, How do the NAEP state average fourth- to eighth-grade math achievement gain scores (as measured by cohort-based gain scores from the 1996 fourth-grade assessment to the 2000 eighth-grade assessment) compare with the TIMSS country average math achievement fourth- to eighth-grade gain scores? The NAEP and TIMSS testing schedules with a four-year interval allow for tracking the same cohort group of students’ academic growth from grade 4 to grade 8. The TIMSS U.S. sample does not consider individual states’ representation or identification. In contrast, the NAEP state assessment provides representative samples for participating states.

Comparison of academic growth among different countries and states requires linking the TIMSS math scale to a U.S. NAEP data set that uses developmental scales of child achievement. NAEP uses cross-grade (vertical) scaling for its math assessment, which facilitates interpretations of scores in terms of grade equivalents and affords comparisons of growth rates; this was made possible through test equating based on item response theory modeling. Since TIMSS scores were not on a developmental scale like NAEP, they were converted into the NAEP scale through statistical moderation (see “Math Achievement” in app. A, available in the online version of American Journal of Education, for NAEP-TIMSS linking procedures).

There were commonalities between NAEP and TIMSS that enhanced the validity of linking the scores (Beaton and Gonzalez 1993; Johnson and Owen 1998). First, both NAEP and TIMSS were based on similar curricular frameworks; although the two assessments have no common items, content analyses of both assessments suggest that they are sufficiently similar to warrant linkage for global comparisons (Johnson and Owen 1998; McLaughlin et al. 1997; NCES 2006). Further, both NAEP and TIMSS used item response theory to estimate proficiency scores in math for each student. However, these proficiency scores are latent variables conditional on the student’s responses to several cognitive and background items and are not directly observed. Because the proficiency scores are estimated, there is some amount of uncertainty or variance associated with them. Thus, rather than having a single observed
math score, there is a range or distribution of plausible values for each sampled student’s proficiency. Plausible values were developed as a computational approximation to obtain consistent estimates of population characteristics in assessment situations in which individuals are administered too few items to allow precise estimates of their ability (Mislevy 1991; Mislevy et al. 1992). In both NAEP and TIMSS, there are five such plausible values for each sampled student, resulting from five random draws from the conditional distribution of proficiency scores for each student.

For international comparison, an analytical subset of the TIMSS data was drawn from the entire pool of participating countries. There were only six countries, including three East Asian and three non-East Asian countries (Canada, Cyprus, Czech Republic, Japan, Korea, Singapore), that met the following two selection criteria: (1) participated in TIMSS assessments in both 1995 fourth-grade math and 1999 eighth-grade math and (2) had an established cutoff birth date for students to be enrolled in school. Only participants who were in grade 4 in 1995 and participants who were in grade 8 in 1999 were included in the analysis. Students who were in other grades were removed from the data file. In order to check the accuracy of the cutoff age rule, the 12 most frequent ages in months for each grade, for each country, were determined. The percentage of students who were out of those appropriate age ranges varied among countries, ranging from 0.0 to 1.5 percent, with an average of 0.4 percent out of range.

For interstate comparison, an analytical subset of the NAEP data was drawn from all participating states. There were only 28 U.S. states that met two selection criteria: (1) participation in both the grade 4 math assessment in 1996 and the grade 8 math assessment in 2000 and (2) an established statewide cutoff date for students to reach age 5 in order to be enrolled in kindergarten. The states included in analyses were Alabama, Arizona, Arkansas, California, Connecticut, Georgia, Hawaii, Kentucky, Maine, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Mexico, New York, North Carolina, North Dakota, Oregon, Rhode Island, South Carolina, Tennessee, Texas, Virginia, West Virginia, and Wyoming. As with the TIMSS data, the 12 most frequent ages in months were determined for each grade in each state in the NAEP data. The empirical distribution of age was cross-checked against the theoretical distribution of age based on cutoff age rule. Students who were not one of the 12 most frequent ages in months were likely to have been retained. The percentage of cases that deviated from the 12-month range was 0.3 on average, but it varied among states, ranging from 0.0 to 0.5 percent.

Table B1, available in the online version of American Journal of Education, shows sample sizes for each country and state, with information on the number of missing values for assigned age and assigned grade variables. In general,
the size of missing data was quite small, falling within a 0–2 percent range among all the nations and states. The NAEP data from 28 U.S. states and the TIMSS data from six countries were merged. The total number of cases in the analytical sample was 193,086, including 101,835 fourth graders and 91,251 eighth graders.

Instrumental Variable Method

Conventional ordinary least squares (OLS) regression analysis of the relationship between child age or grade and math achievement could be misleading due to the possibility of selection bias. Student age may influence parent and teacher educational decisions, which in turn can affect the status and progress of grade levels in school and consequently academic achievement outcomes (Bedard and Dhuey 2006). Even in nations and states with school entry cutoff dates, the rules may not be strictly followed so that one cannot assume random assignment of students to grades on the basis of birth dates. This fuzzy cutoff problem poses a threat to estimating the causal impact of age on achievement. When a significant fraction of children defer school entry by a year, these “redshirted” children are not randomly selected, and their academic readiness at the time of school entry may be different from those who enter school at a typical age (Datar 2006). Children who are young at school entry are more likely to repeat a grade, whereas relatively older students may be more cognitively mature and score higher on achievement tests.

Therefore, both age and grade variables cannot be treated as exogenous, and the OLS regression would give biased estimates of the age effect and the grade effect. In other words, bias can occur due to the omission of important covariates such as child characteristics (e.g., cognitive ability) in the model that are associated with both age and child achievement. This study uses an instrumental variable (IV) estimation method in order to address such endogeneity and omitted-variable problems.

A good IV should be associated with the endogenous variable (i.e., age and grade) but be uncorrelated with the outcome variable (i.e., math achievement), except through its indirect relationship with those endogenous variables (Schneider et al. 2007). Assigned age—the relative age at which children should be observed on the basis of their birth date relative to the school cutoff date—meets these conditions (Bedard and Dhuey 2006). Likewise, assigned grade—the grade in which the students would be expected to be enrolled, based on their birth date relative to the school cutoff date—can be used. This study uses both assigned age and assigned grade as the instruments and employs two-stage regression analysis in order to compute IV estimates.
Hierarchical linear models (HLMs) were used to examine within-nation/state and between-nation/state variations in the TIMSS and NAEP cohorts’ math achievement gain from grade 4 to grade 8 (Raudenbush and Bryk 2002). Because students in the grade 4 sample and the grade 8 sample do not come from the same schools and their schools cannot be matched between the two grades, a school-level model was not possible for this multilevel analysis. Adapting the procedures of two-stage least squares regression for conventional unilevel IV analysis, two successive stages of HLM regression analysis were conducted for this multilevel IV analysis. The first-stage HLM regression analysis was conducted to obtain predicted values of age and grade for students in each country and state on the basis of their assigned age and assigned grade. Specifically, the first-stage HLM analysis involved (1) the regression of age on assigned age and assigned grade and (2) the regression of grade on assigned age and assigned grade. Then, the second-stage HLM regression analysis was conducted to account for variations in math achievement with the five plausible values of student math achievement as dependent variables (see equations below for final HLM level 1 and level 2 models). NAEP and TIMSS have five plausible values for each sampled student, resulting from five random draws from the conditional distribution of proficiency scores for each student. The parameter estimates from the HLM analyses are based on the average parameter estimates from analyses of the five plausible values (Raudenbush et al. 2005).

The level 1 (student-level) predictors included instrumented versions of age and grade variables, that is, the predicted values of age and grade variables as obtained from the first-stage HLM analysis level 1 residual file, and their interaction term. Although the age effect is expected to reflect the influences of student, family, and preschool characteristics on growth during the year before school entry, it may also include school effects if schools function in a way that strengthens or weakens initial achievement gaps between older and younger students. In contrast, the grade effect is expected to capture value-added school effects, above and beyond growth that would be expected even before school entry. The level 1 coefficients of age, grade, and the age-by-grade interaction become outcome variables at level 2. The level 2 (nation/state-level) predictors included a series of dummy code indicators of whether the source of the data was a U.S. state, an East Asian country, or a non-East Asian country.

The level 1 model (student level) is

\[ Y_{ij} = \beta_{0j} + \beta_{1j} age_{ij} + \beta_{2j} grade_{ij} + \beta_{3j} (age \times grade)_{ij} + e_{ij}, \]

where \( age_{ij} \) is the instrumented age of student \( i \) in nation/state \( j \), \( grade_{ij} \) is
the instrumented grade of student $i$ in nation/state $j$, $(\text{age}' \times \text{grade}')_j$ is the product of instrumented age and grade variables for student $i$ in nation/state $j$, and $\epsilon_{ij}$ is a level 1 random effect that represents the deviation of student $y$'s score from the predicted score based on the student-level model. The level 2 model (nation/state level) is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{East Asia}_j + \gamma_{02}\text{non-East Asia}_j + r_{0j};$$

where East Asia is a dummy variable for Japan, Korea, and Singapore; non-East Asia is a dummy variable for Canada, Cyprus, and Czech Republic; and $r_{0j}$ is a level 2 random effect for the level 1 intercept and slopes ($p = 0, 1, 2, 3$).

A sequence of HLM models was applied to the analytical sample of combined NAEP and TIMSS data. First, a base model without any predictors showed an intraclass correlation of .08; 8 percent of the variance in math achievement occurred between nations/states, and 92 percent of the variance occurred within nations/states. Second, a random coefficient model with instrumented age and grade as the level 1 predictors provided estimates of the grade effect, age effect, and age-by-grade interaction effect for individual nations and states. These predictors were grand-mean centered for adjustment of between-nation/state differences in age and grade distributions. Finally, a fully conditional model included level 2 predictors for comparing American states with East Asian and non–East Asian countries.

Predictors of International and Interstate Variations

In order to differentiate American states by their pace of growth, we classified American states into subgroups according to their math achievement gain scores relative to two references. We employed both criterion-referenced and norm-referenced classifications to account for interstate variations. Because the United States does not have any formal national standards of academic growth in math, we drew inferences about this velocity standard by interpolating student performance standards between grade 4 and grade 8 on the basis of NAEP math “proficient” achievement level cutoff scores (249 for grade 4 and 299 for grade 8); this derived standard of growth demands a 50-point gain on the NAEP scale during the four years of middle-level education. Second, we also grouped states on the basis of their average gain scores relative to the national norm, classifying states in the top quartile as fast track and the rest in the bottom three quartiles as regular track. We classified states into two groups, for the sake of contrast, using 56 points, the minimum gain score of the top-quartile states, as a threshold for the fast-track group.

Further, we selected commonly available aggregate family and school variables on the basis of prior research in order to account for the sources of
both international and interstate variations (see app. A for a full description of the level 2 variables). For family and socioeconomic background variables, we used parents’ education and national income as indicators. We chose in-field teaching and school year length as common indicators of school quality. Prior research demonstrated the importance of school time (AERA 2008; Berliner 1990; Fisher et al. 1980; National Education Commission on Time and Learning 1994; Smith et al. 2005) and of in-field teaching for student learning (Darling-Hammond 2000; Goldhaber and Brewer 1997). In comparison with the East Asian countries in the sample, the U.S. states have mixed conditions with a relatively higher level of parental education but a shorter school year and fewer qualified math teachers (see table B2, available in the online version of *American Journal of Education*). On average, the U.S. states report parents with a high school education or above at about 20 percent higher than do the East Asian countries. The U.S. states have an average of 179 days per school year, and 49 percent of math teachers have a math major, whereas the East Asian countries report 209 days in the school year (more than one month longer) and 76 percent of in-field math teachers (27 percentage points more).

**Limitations**

There are several caveats in interpreting the results of this study. First, the study examined academic growth in a single subject area, mathematics, during middle-level school years only. It is still possible that relatively stronger or weaker school effects may be observed in primary school years or in high school years. Second, the study restricted its analytical sample to nations and states with birthday cutoff dates for school entrance, so that it produces estimation of local average treatment effects. Although there is no systematic pattern of any known differences between those with and those without school entrance age cutoff rules, applying the study findings to excluded nations and states is not warranted. Third, while the study focuses on the national/state average math achievement growth trajectory for excellence, we acknowledge that it is only one indicator of educational system performance. Particularly, it is desirable to examine how achievement gaps within each nation/state change over the course of schooling, for equity.

Further, we acknowledge a possible threat to interstate comparison of achievement gains due to confounding effects of interstate mobility between the U.S. fourth- and eighth-grade cross-sectional student samples. If there had been significant moves in and out of each state between 1996 and 2000, those assessed in a state for the first assessment (fourth-grade NAEP) would not have been the same population at the second assessment (eighth-grade NAEP).
Further, those in a state at a particular assessment would not have been subject to that specific current state’s kindergarten age cutoff policy, if they had been in a different state at the time of school entry. To address this, we checked how the fourth-grade population matches the eighth-grade counterpart in each state on the basis of available key sociodemographic markers (race/ethnicity and poverty status) and correlated those state-specific measures of any sociodemographic changes with their fourth- to eighth-grade math achievement gains in the same period.

While there were overall changes in the U.S. fourth- and eighth-grade student populations, such as decreases in the percentage of poor students and increases in the percentage of Hispanic and Asian students between 1996 and 2000 across the states, the results of correlation analyses show that variations among states in the degree of sociodemographic changes were not associated with the size of their math achievement gain scores ($r = -0.13$ for change in percent poor; $r = 0.07$ for change in percent white; $r = 0.19$ for change in percent black; $r = -0.14$ for change in percent Hispanic; $r = -0.17$ for change in percent Asian/Pacific Islanders). Further, independent samples $t$-tests for differences between fast-track and regular-track states show that the two groups did not differ in terms of the measures of sociodemographic changes during 1996–2000, implying that their observed differences in math achievement gains were not attributable to changes in student family background characteristics ($t = -0.69$ for change in percent poor; $t = -1.23$ for change in percent white; $t = -0.02$ for change in percent black; $t = 0.64$ for change in percent Hispanic; $t = 0.75$ for change in percent Asian/Pacific Islanders). Although the same issue may arise with any international mobility between TIMSS fourth- and eighth-grade samples during 1995–94, the same kind of analysis was not possible with TIMSS, which lacked commonly reliable sociodemographic markers between its fourth- and eighth-grade student samples; we assumed that the proportion of such international movers is very small, and its effect would be negligible.

Finally, we acknowledge that the decomposition of achievement gains into age effect and grade effect does not pinpoint specific causes of those gains. Age effect reflects more than maturation and cognitive development, as it is derived from the comparison of younger and older children in the same grade, and this is likely to originate from any preschool learning differences before school entry. In other words, it tells how much students would learn math even without formal schooling, as they would have done in preschool years. Depending on different combinations of preschool learning settings and experiences among children in different countries and states, it may reflect not only home-schooling effects (for kids who stay home in their prekindergarten year) but also preschool effects (e.g., private child care center or public school prekindergarten classes). This gives a baseline gain that one may expect of
**International and Interstate Gaps**

students if no schooling were offered. By the same token, the grade effect is supposed to capture any value-added contribution of schooling above and beyond what children might learn without going to schools. However, the grade effect is unlikely to capture school effects only. Particularly in East Asian countries, private tutoring supplements public schooling for a majority of students, so that part of the grade effect is likely to reflect out-of-school learning benefits (Baker et al. 2001; Lee 2007). Therefore, this study is highly exploratory in identifying the sources of international and interstate variations in both age effects and grade effects on the basis of a limited set of variables commonly available for the nations and the U.S. states.

**Results**

**Descriptive Analysis**

Descriptive analysis of the merged NAEP and TIMSS data set shows that, on average, students make substantial math achievement gains from grade 4 to grade 8 across all six countries and 28 American states (57-point gain from the grade 4 average score of 225 to the grade 8 average score of 282). The average annual math gain of 14 points is equivalent to approximately .5 SD on the U.S. NAEP scale. However, there are substantial international and interstate variations in the size of those math achievement gains (see fig. 2). In the United States, the average fourth- to eighth-grade math gain score ranges from 45 points in Mississippi to 60 points in Montana. All of the 28 American states lag behind most of the TIMSS nations in our sample, including the three East Asian countries, in terms of their average math achievement gains. The average four-year math gain score among our subsample of 28 American states was 50 points, whereas the average gain score among the three East Asian countries was 65 points. The average difference between U.S. states and East Asian nations in their math gain scores was about 15 points on a 0–500 scale, which translates into an effect size of .5 SD. In other words, typical students in the American states grow slower than their counterparts in East Asian nations during the middle school years, so that their cumulative learning gap is worth one year of schooling.

Despite these international gaps in math achievement gains, it should be noted that there were variations among the U.S. states. The top three states in the fast track—Minnesota, Montana, and Oregon—gained 57 points on average. The remaining 25 states in the regular track gained 51 points on average. As shown in figure 2, the U.S. fourth-grade NAEP math proficiency standard is as high as the average math scores of the highest-performing East Asian countries, but the U.S. eighth-grade NAEP math proficiency standard
becomes lower than the average of the highest-performing East Asian countries. This suggests that not only American students’ actual math achievement but also national expectations as measured by the common NAEP math proficiency standards fall behind those of East Asian counterparts at the middle school grade level.

**HLM Instrumental Variable Analysis**

The first-stage HLM analysis shows that both age and grade (as endogenous variables) are well explained by the two instruments (assigned age and assigned grade). For age, the level 1 analysis shows strong positive effects of assigned age ($\gamma = .93, p < .001$) and assigned grade ($\gamma = 12.00, p < .001$); the level 1 model explains 99 percent of the within-nation/state variance. Second, the level 1 analysis of grade also showed similarly strong positive relationships for the effect of assigned age ($\gamma = .03, p < .001$) and for the effect of assigned grade ($\gamma = .95, p < .001$); the level 1 model explains 96 percent of the within-
International and Interstate Gaps

TABLE 1
Hierarchical Linear Model Random Coefficient Model: Age and Grade Effects on Math Achievement

<table>
<thead>
<tr>
<th></th>
<th>Fixed-Effects Coefficient</th>
<th>Random-Effects Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (average math score)</td>
<td>249.86***</td>
<td>116.05***</td>
</tr>
<tr>
<td></td>
<td>(1.85)</td>
<td></td>
</tr>
<tr>
<td>Age effect</td>
<td>.60***</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>(.04)</td>
<td></td>
</tr>
<tr>
<td>Grade effect</td>
<td>5.73***</td>
<td>5.07***</td>
</tr>
<tr>
<td></td>
<td>(.63)</td>
<td></td>
</tr>
<tr>
<td>Age × grade effect</td>
<td>−.14***</td>
<td>.004***</td>
</tr>
<tr>
<td></td>
<td>(.01)</td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—Standard errors of the coefficients are in parentheses. Age effect = average math gain per month based on the difference in math achievement between youngest and oldest in the same grades. Grade effect = average math gain per grade based on the difference in math achievement between grade 4 and grade 8 students of the same ages. Age × grade effect = interaction between age effect and grade effect.

*** p < .001.

nation/state variance. A multiparameter omnibus test for the effects of assigned age and grade together also reveals that they are strong instruments for the model of age ($χ^2(2) = 267,714, p < .001$) and the model of grade ($χ^2(2) = 124,652, p < .001$). Further, the Hausman test of endogeneity rejects the null hypothesis that age and grade variables are exogenous; a multiparameter omnibus test for the effects of residuals shows significance of possible omitted variables ($χ^2(2) = 8,272, p < .001$). These results together support the use of this IV estimation method.

The second-stage HLM analyses used predicted age and grade values obtained from the above first-stage analyses in order to account for math achievement as an outcome variable. Table 1 summarizes the results of the HLM random coefficient model regarding age effect, grade effect, and age-by-grade interaction effect. The fixed effects part of table 1 shows statistical significance of all three effects ($γ = .60, p < .001$, for age effect; $γ = 5.73, p < .001$, for grade effect; $γ = −.14, p < .001$, for age-by-grade interaction effect). The level 1 model explains about 40 percent of the within-nation/state variance in math achievement.8

The estimate of age effect suggests that students gain .60 points per month, on average, across all nations and states in the sample. While the average age effect was estimated across grade 4 and grade 8, the negative value of the age-by-grade interaction effect suggests that this age effect would be greater at the lower grade level but smaller at the upper grade level. For example, at the time of kindergarten entry (grade = −1), the estimated average age effect would be about 1.58 (age effect + [grade − centering value] × grade effect).
FIG. 3.—Scatter plot of grade effect versus age effect (10-month grade equivalent) estimates among six countries and 28 U.S. states (equivalence line drawn for the same values of age effect and grade effect).

Multiplying age effect (monthly gain) by 10 (number of months for one school year) would give an equivalent of grade effect of 6 points ($0.60 \times 10 = 6$). The size of this average age effect across all countries and states in

The random effects in table 1 show significant international and interstate variations in all of the level 1 parameters except for age effect. Variances for the intercept (average math achievement), grade effect, and age-by-grade interaction effect were all statistically significant at the .001 level. Correlations among these random effects suggest that the high-performing nations and states have relatively stronger grade effects ($r = 0.56$) than their low-performing counterparts, but there are no systematic differences in age affects ($r = 0.04$). At the same time, the nations and states with stronger grade effects have relatively weaker age effects ($r = -0.60$). Figure 3 shows the distributions of grade effects and age effects (10-month grade effect equivalent).

Effect, or $0.60 + [1 - 6] \times -0.14 = 1.58$), but the age effect would diminish by 0.14 for each additional year of schooling. Likewise, the average grade effect of 5.73 would be smaller for older children but larger for younger children.
### TABLE 2

Hierarchical Linear Model Intercepts and Slopes as Outcomes Model 1: Comparison of U.S. States versus East Asian and Other (Non–East Asian) Nations

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Age Effect</th>
<th>Grade Effect</th>
<th>Age × Grade Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States (all states combined)</td>
<td>246.06***</td>
<td>.59***</td>
<td>5.17***</td>
<td>−.14***</td>
</tr>
<tr>
<td></td>
<td>(1.53)</td>
<td>(.06)</td>
<td>(.76)</td>
<td>(.01)</td>
</tr>
<tr>
<td>U.S. gap against East Asian nations</td>
<td>34.40***</td>
<td>.02</td>
<td>4.00*</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td>(4.93)</td>
<td>(.12)</td>
<td>(1.63)</td>
<td>(.04)</td>
</tr>
<tr>
<td>U.S. gap against non–East Asian nations</td>
<td>3.25</td>
<td>.10</td>
<td>2.14</td>
<td>−.04</td>
</tr>
<tr>
<td></td>
<td>(4.93)</td>
<td>(.14)</td>
<td>(1.78)</td>
<td>(.04)</td>
</tr>
<tr>
<td>U.S. fast-track states’ gap against regular track</td>
<td>−9.99***</td>
<td>.17</td>
<td>−4.03*</td>
<td>−.03</td>
</tr>
<tr>
<td></td>
<td>(4.47)</td>
<td>(.24)</td>
<td>(2.99)</td>
<td>(.04)</td>
</tr>
<tr>
<td>U.S. fast-track states’ gap against East Asian nations</td>
<td>17.80***</td>
<td>.18</td>
<td>.37</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>(5.98)</td>
<td>(.26)</td>
<td>(3.24)</td>
<td>(.06)</td>
</tr>
<tr>
<td>U.S. fast-track states’ gap against non–East Asian nations</td>
<td>−8.89</td>
<td>.28</td>
<td>−1.71</td>
<td>−.07</td>
</tr>
<tr>
<td></td>
<td>(5.97)</td>
<td>(.26)</td>
<td>(3.32)</td>
<td>(.06)</td>
</tr>
</tbody>
</table>

**NOTE.**—U.S. fast-track states include three states that belong to the top quartile of total math achievement gain scores, whereas regular-track states are all remaining states. Standard errors of the coefficients are in parentheses.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

The sample appears to be slightly larger than the size of its corresponding grade effect of 5.73, but the difference is not statistically significant. Further, a closer look at individual nations and states’ profiles based on HLM empirical Bayes estimates reveals variations; the grade effect exceeds the age effect in some countries (Canada, Singapore, Korea), whereas the two effects are almost equal in most American states. The ratio of grade effect to total gain is about 47 percent in American states and 54 percent in other nations, on average. However, direct comparison of age and grade effects is complex, as the sizes of those effects are likely to change over the course of child development and schooling.

The results of HLM analysis for comparison of the U.S. states with East Asian and non–East Asian nations are summarized in table 2. The U.S. states’ average age effect and grade effect were .59 ($p < .001$) and 5.17 ($p < .001$), respectively.
respectively. At the same time, the U.S. average age-by-grade interaction effect was \(-.14\) \((p < .001)\). This means that the average age effect (monthly math gain) for fourth-grade American students can be as high as \(.87\) (age effect + [grade – centering value] × grade effect, or \(.59 + [4 - 6] \times -14 = .87\)), whereas the average age effect for eighth-grade American students can be as low as \(.31\) (age effect + [grade – centering value] × grade effect, or \(.59 + [8 - 6] \times -14 = .31\)). These age effects for a school year (10 months) are about .3 and .1 SD (U.S. SD = 30 for grade 4 and 35 for grade 8).

Differences between the U.S. and East Asian nations vary by the type of effect (see table 2). The grade effect is significantly smaller in the United States than in East Asian countries on average \((\gamma = 4.00, p < .05)\), whereas there are no significant differences in age effect \((\gamma = .02, p = .02)\) and age-by-grade interaction effect \((\gamma = .04, p = .38)\). It appears that the gap for American states arises not because their age effects are weaker than other countries but because their grade effects are weaker. However, there are no significant differences between non–East Asian countries and the United States in all of those effects.

Differentiation of U.S. states into two groups (fast track and regular track) shows that the differences relative to East Asian nations in grade and age effects depend on how states are grouped and compared. When we grouped states by using the 50-point gain threshold (derived from the NAEP proficiency standard cutoff scores), the results remained largely unchanged from those of the aggregate U.S. comparison with other nations. However, when we identified fast-track states by using the 56-point gain threshold (the minimum gain scores of the top-quartile group), no significant differences were detected between those states and the East Asian counterparts for both age and grade effects (table 2). While the top-quartile American states did not fall behind East Asian countries in the grade effect \((\gamma = .37, p > .05)\), they exceeded other states in the grade effect \((\gamma = -4.03, p < .05)\).

Finally, we ran a full model with aggregated family and school variables to account for both international and interstate variations at level 2 (see full sample in table 3). For the intercept (average math achievement across grades 4 and 8), both school year length and teacher qualification variables were positively associated with the higher achievement \((\gamma = 6.87, p < .01,\) for school year length; \(\gamma = 4.32, p < .05,\) for teacher qualification), whereas both parent education and per capita income were not significant predictors. For the age effect, none of the predictors were significant. For the grade effect, school year length was the only significant predictor of math achievement gains \((\gamma = 1.19, p < .01)\). For the age-by-grade interaction effect, both parent education and school year length were significant predictors \((\gamma = .10, p < .01,\) for parent education; \(\gamma = .05, p < .001,\) for school year length). Supplementary analysis of the same model with the U.S. sample only showed
International and Interstate Gaps

**Table 3**

*Hierarchical Linear Model Intercepts and Slopes as Outcomes Model 2: Effects of Aggregate Family and School Variables in Full Sample versus U.S. Sample Only*

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Age Effect</th>
<th>Grade Effect</th>
<th>Age × Grade Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full sample:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per capita income</td>
<td>−.19</td>
<td>−.05</td>
<td>.34</td>
<td>−.01</td>
</tr>
<tr>
<td>(2.27)</td>
<td>(.07)</td>
<td>(.93)</td>
<td>(.01)</td>
<td></td>
</tr>
<tr>
<td>Parental education</td>
<td>.77</td>
<td>.13</td>
<td>−2.61</td>
<td>.10**</td>
</tr>
<tr>
<td>(5.72)</td>
<td>(.18)</td>
<td>(.29)</td>
<td>(.04)</td>
<td></td>
</tr>
<tr>
<td>School year length</td>
<td>6.87**</td>
<td>−.02</td>
<td>1.19**</td>
<td>.05***</td>
</tr>
<tr>
<td>(2.00)</td>
<td>(.05)</td>
<td>(.66)</td>
<td>(.01)</td>
<td></td>
</tr>
<tr>
<td>Math teacher qualification</td>
<td>4.32*</td>
<td>.09</td>
<td>−.91</td>
<td>.01</td>
</tr>
<tr>
<td>(2.12)</td>
<td>(.06)</td>
<td>(.86)</td>
<td>(.01)</td>
<td></td>
</tr>
<tr>
<td><strong>U.S. sample:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per capita income</td>
<td>1.09</td>
<td>−.06</td>
<td>.90</td>
<td>−.01</td>
</tr>
<tr>
<td>(1.42)</td>
<td>(.10)</td>
<td>(1.11)</td>
<td>(0.0)</td>
<td></td>
</tr>
<tr>
<td>Parental education</td>
<td>25.79**</td>
<td>.24</td>
<td>.18</td>
<td>.12*</td>
</tr>
<tr>
<td>(6.35)</td>
<td>(.25)</td>
<td>(2.92)</td>
<td>(0.05)</td>
<td></td>
</tr>
<tr>
<td>School year length</td>
<td>−9.72</td>
<td>−.24</td>
<td>1.83</td>
<td>−.09</td>
</tr>
<tr>
<td>(5.33)</td>
<td>(.30)</td>
<td>(3.73)</td>
<td>(0.07)</td>
<td></td>
</tr>
<tr>
<td>Math teacher qualification</td>
<td>1.50</td>
<td>−.08</td>
<td>.52</td>
<td>.00</td>
</tr>
<tr>
<td>(1.95)</td>
<td>(.08)</td>
<td>(1.07)</td>
<td>(0.02)</td>
<td></td>
</tr>
</tbody>
</table>

**Note.** — Standard errors of the coefficients are in parentheses.

*  \( p < .05 \)
**  \( p < .01 \)
***  \( p < .001 \)

Different patterns: parental education was significant for the intercept and age-by-grade interaction effect, whereas none of the school factors were significant (see U.S. sample only in table 3).

These findings suggest that parent education, as a proxy indicator of family resources and engagement for child education, is a strong predictor of achievement gaps among the U.S. states, but it does not explain gaps between the United States and other nations. In contrast, a longer school year, as a proxy indicator of the intensity of schooling, explains international gaps but not interstate gaps. The restricted range of interstate variations in terms of schooling year length may have contributed to the paucity of observed school effects in the U.S. sample. The only common finding between the full sample and the U.S. sample results is that nations or states with relatively higher levels of parent education have stronger age-by-grade interaction effects (i.e., slower decay of the age effect over the course of schooling). These results are tentative, implying that school-related variables account for more of the international achievement gaps, whereas family-related variables account for more of the interstate achievement gaps.
Discussion

Educational policy makers have raised concerns about the productivity of American middle schools for building human capital as measured by standardized tests of core math knowledge and skills. A recent study of long-term NAEP trends in academic growth imply a tripartite pattern in which American students are gaining ground at the pre-/early primary school level, holding ground at the middle school level, and losing ground at the high school level (Lee 2010). Although middle-level education is not necessarily the weakest link in the American school system, it appears to be a crucial turning point of educational transition in which students’ academic progress in math begins to slow down relative to other nations, particularly those in East Asia.

While there is no shortage of comparative research that has studied this topic, the previous aggregate national comparisons obscured interstate variations within the United States. Further, they examined test results at different grades separately and were not able to measure value-added academic progress due to the cross-sectional nature of TIMSS design. Merging the U.S. NAEP state data and TIMSS international data on a common scale of math achievement across fourth and eighth grades, this study examines the same data in creative ways. It attempts to give new insights into these old questions: How much do American students improve their math achievement during middle school years, how well do different states in the United States fare against other nations in terms of those math achievement gains, and how much of the U.S. math achievement gap relative to high-performing nations such as those in East Asia is attributable to value-added school effects?

This study reveals significant student progress in math during middle school years across all TIMSS countries and U.S. NAEP states. On average, the U.S. students make substantial gains from grade 4 to grade 8 (48-point gain on a 0–500 scale, or 1.6 SD). There exist substantial variations among American states and other countries in the amount of yearly math achievement gains between grades 4 and 8; it ranges from an 11.6-point gain per year (.39 SD) in Mississippi to a 16.9-point gain per year (.56 SD) in Singapore. Clearly, most American states have relatively smaller academic progress than East Asian countries in math between fourth and eighth grade. The U.S. states whose average math achievement is at the NAEP proficiency standard at grade 4 are expected to gain 50 points in order to be on track to proficiency by grade 8. However, even this amount of math gain, as derived from the NAEP proficiency standards, would put American states behind high-performing East Asian countries by grade 8.

This study identifies key growth factors by decomposing total math achievement gains into age effect, grade effect, and age-by-grade interaction effect. The HLM IV analyses showed that across the nations and the states, math
achievement gains between grade 4 and grade 8 are attributable to both grade effects (95 percent confidence interval of the grade effect = 4.5 ∼ 7.0-point gain per grade) and age effects (95 percent confidence interval of the age effect = 5.2 ∼ 6.8-point gain per grade equivalent). This shows strong value-added school effects but, at the same time, the persistence of comparable age effects. There was a significantly negative association between age and grade effects across the nations and the states, which implies a potential trade-off between school effects and other sources of influence on student achievement, such as student and family effects. The countries or states where children’s ages have greater influence on the development of math achievement may have relatively weaker school effects in math education. The negative interaction effects between age and grade may also imply that the growing influences of schools occur with the diminishing influences of families over time.

This study’s estimates of age effects for the U.S. NAEP sample were .3 SD (fourth grade) and .1 SD (eighth grade), and they were highly comparable to previous studies’ estimates. For example, Bedard and Dhuey’s (2006) estimate of age effect per month on the basis of their IV analysis of the Early Childhood Longitudinal Study (ECLS) third-grade data and the National Education Longitudinal Study (NELS) eighth-grade data was .298 for grade 3 and .235 for grade 8; since both ECLS and NELS test scores were standardized in the study (mean = 50 and SD = 10), the 10-month grade equivalents of age effects were .3 and .2 SD. Applying the same IV method to the same sets of data with different covariates, Elder and Labotsky (2006) reported similar age effects: .4 for grade 3 and .1 for grade 8. Based on Reardon’s (2003) estimate of age effect with ECLS data, standardized 10-month gains are .5 for the prekindergarten year. In this study, there is no direct estimate of the age effect at the prekindergarten year, but the estimated average age effect for this age level by extrapolation would be .5 also. The common patterns of age effects in previous studies as well as this study call for attention to the academic advantages of being older at school entry (and possibly the benefits of longer preschool education) and decay of the age effect or preschool learning effect over the course of formal schooling.

This study also demonstrates both strikingly converging age effects across the nations and the states and, at the same time, substantially diverging grade effects. In fact, the size of the aforementioned age effects as observed in the U.S. states was highly comparable to corresponding age effects in high-performing East Asian countries. In contrast, the size of grade effects observed in the United States was relatively smaller. Specifically, the HLM analysis showed a significant U.S.–East Asia gap in their grade effect: on the average, the U.S. average standardized grade effect was .15, whereas the East Asian average standardized grade effect was .27. If this yearly difference (.27 − .15 = .12) accumulates over the course of schooling, the U.S.–East Asian gap
owing to the grade effect would exceed 1 SD by the time students graduate from high school. Only a few American states were able to match East Asian countries in terms of the grade effect as a proxy indicator of value-added school contribution.

While the math achievement gap between the United States and East Asian countries tends to widen during middle school years, it appears that a significant part of observed change in the gap is attributable to grade effects as opposed to age effects. This finding raises an outstanding question as to why American students experience slower academic growth than their East Asian counterparts during their middle school years. The findings of relatively smaller math achievement gains with weaker grade effects among the U.S. states may be related to their disadvantages in terms of key schooling conditions such as school year length and teacher qualifications. The evidence does not support the claim that the U.S. students’ math learning gap occurs before school entry, largely owing to their relatively weaker family effects, but rather suggests that the U.S. achievement gap occurs mostly after school entry and widens due to school effects as well as family-school interaction effects.

There are some remaining issues due to the limitations of this study. While the NAEP provides a common national benchmark of student performance in math across all U.S. states, it was not designed for linking to TIMSS, which does not necessarily measure the same math content as NAEP. Further, our search for national- or state-level predictors of age and grade effects was restricted by what is commonly available from student and teacher survey questionnaires in both NAEP and TIMSS. With those caveats in mind, our study results call for stronger alignment of NAEP and TIMSS across grades that allows subsequent studies to better track achievement gains and compare educational transition. It is necessary to further examine the sources of international and interstate variations in math learning gains at different levels of ages or grades (pre-/early primary, middle, high school) in the full course of child development. This requires redesigning international and U.S. state data collections for pre-/early primary school grades (below grade 4) and high school grades (above grade 8).

For the relevance of this study using data from the 1990s to today’s policy context, we need to consider changes in the U.S. education policy and practice during the last decade, such as a national movement toward high standards and test-driven school accountability as well as increased school reform directed at improving math achievement across all grades. Recent TIMSS reports (Mullis et al. 2004, 2008) show that while the U.S. average math achievement for both fourth and eighth graders has been on the rise since 1995, the relative performance of the United States remains lower for eighth graders in 2007 than for fourth graders in 2003. Therefore, it seems that the United States still has to cope with the problem of lagging behind other nations during
International and Interstate Gaps

middle school years. Further, the past test-driven school accountability policy did not bring about significant improvement of key schooling conditions such as in-field teaching (Lee and Wong 2004). As more of the U.S. states have moved to adopt a value-added growth model for school accountability, it is critical to address not only variations among the states in the rigor of student performance standards at the same grade levels but also inconsistencies within the states in the rigor of standards between different grade levels. This requires setting increasingly higher standards of student achievement from the elementary school level to the middle school level, accompanied by increased family and school support for this critical stage of academic transition.

Notes

1. A TIMSS test score is not on a developmental scale, so it is not possible to directly evaluate the absolute amount of achievement gain. However, it is possible to assess achievement gain from fourth grade in 1995 to eighth grade in 1999, relative to the international average. The difference from the average across 17 nations (common to both 1995 fourth-grade and 1999 eighth-grade TIMSS) in each year was calculated and then compared. The change in this difference score for each country measures the 1995 fourth-grader cohort’s gain or loss in relative math performance between the fourth and the eighth grades.

2. There are also long-standing debates about the relative importance of genetic versus environmental factors (Flynn 1991; Lynn 1987; Stevenson et al. 1986).

3. Even though the labels of those two effects should evoke the image of maturation effect versus schooling effects, respectively, it is not really clear where these effects actually come from. Prior researchers were vague about the meaning of age and grade effects, as they did not examine the sources of those effects in relation to specific family or school factors.

4. Entrance age effect considers the fact that older children have had more time to acquire human capital (e.g., math knowledge and skills) “before” school entry than did younger children through parent teaching or exposure to preschool learning experiences. It needs to be noted that this portion of the age effect captures the difference in learning between older and younger children in the same grade that would occur during the year before their school entry. Evidence on the effect of school entrance age on academic achievement has been well established (Datar 2006; McDonald 2001). In contrast, age-at-test effect considers the possibility that older (and thus more physically and cognitively mature) children may acquire human capital at a faster rate than do their younger peers “after” school entry until the time of testing (e.g., grade 4), given the assumption that learning begets learning. This age-at-test effect also may occur through schooling experiences that could widen an initial age-related achievement gap (e.g., through ability grouping with differentiated instruction) or reduce the gap (e.g., through active remediation for low-achieving students). Prior research suggests that the phenomenon of an age-related learning gap reflects entrance age effect more than the age-at-test effect since the gap is evident from the very beginning of kindergarten (Elder and Lubotsky 2006).

5. McLaughlin et al. (1997, app. A) point out that the NAEP and TIMSS assessments
Lee and Fish

“are sufficiently similar to warrant linkage for global comparisons but not necessarily for detailed comparisons of areas of student achievement or processes in classrooms.”

Specifically, they noted a few important differences on the basis of content analysis of the two assessments: (1) the TIMSS mathematics assessment was embedded in a combined math and science assessment; (2) the NAEP mathematics assessment included blocks of items on which calculators were available and others on which rulers and cardboard shapes were to be used; (3) there were somewhat more items on geometry in NAEP (19 vs. 13 percent); (4) more TIMSS items involved computation (59 vs. 40 percent), and more involved decimals or fractions (34 vs. 13 percent); (5) more of the TIMSS items were multiple choice (79 vs. 57 percent); and (6) more NAEP items than TIMSS items were difficult, on the basis of percentages of correct responses given by U.S. students. A more recent comparison study (NCES 2006) also noted differences: when NAEP and TIMSS mathematics items were classified to each other’s assessment frameworks in terms of content topics and subtopics, about 20 percent of fourth-grade items and 15 percent of eighth-grade items from each assessment could not be classified to a subtopic in the other’s framework. In view of the differences, therefore, our test-linking results should be interpreted with caution.

6. In order to check for potential bias of international and interstate comparisons due to a different size or pattern of missing data, we conducted t-tests for the significance of attrition rate differences for both assigned age and assigned grade between paired sets of countries and states (e.g., between U.S. states and Asian countries, between fast-track and regular-track U.S. states) that are of interest for comparison in this study, and none of them turned out to be significant at the \( p < .05 \) level.

7. IV estimation requires the exclusion restriction that IVs are not direct determinants of outcome variables and thus are excluded from the outcome variable equation. Assumptions that exclusion restrictions are valid are generally untestable and can be challenged. In this study, however, the assumption that determination of assigned age and assigned grade on the basis of child birth dates is random and unrelated to child ability or other causal predictors of math achievement may not be unreasonable. For a general introduction to the principles, assumptions, utilities, and limitations of IV estimation method in social research, see Foster and McLanahan (1996) and Winship and Morgan (1999).

8. For the sake of comparison, we find that these IV analysis results are quite different from those of ordinary regression analysis without instrumentation, which underestimates the age effect \( (b = .39) \) and age × grade interaction effect \( (b = -.05) \) but overestimates the grade effect \( (b = 20.09) \).

9. Alternatively, it could be 12 months rather than 10 months to calculate the grade-equivalent age effect, but 10 months would be more accurate for comparison because students usually do not go to school during the two-month summer break, and prior research shows little or no learning gains during the summer time.

10. Reardon (2003) gives proper cautions about interpretation. Because the sample includes students who range from roughly 60 to 72 months old at kindergarten entry, the prekindergarten growth rate only applies to children between age 5 and 6. Further, because the prekindergarten gain estimate was based on age-related differences in achievement at the time of school entry, it was likely to be a serious underestimate of actual prekindergarten learning gain to the extent that age is correlated with ability factors due to American parents’ redshirting practice, which delays their children’s school entry.
International and Interstate Gaps

References


