RESEARCH REPORT

September 2001
RR-01-19

Teacher Classroom Practices and Student Performance: How Schools Can Make a Difference

Harold Wenglinsky

ETS®
Educational
Testing Service

Statistics & Research Division
Princeton, NJ 08541

Teacher Classroom Practices and Student Performance:

How Schools Can Make a Difference

Harold Wenglinsky

Educational Testing Service

September 2001

Research Reports provide preliminary and limited dissemination of ETS research prior to publication. They are available without charge from:

Research Publications Office
Mail Stop 10-R
Educational Testing Service
Princeton, NJ 08541

Abstract

Quantitative studies of school effects have generally supported the notion that the problems of U.S. education lie outside of the school. Yet such studies neglect the primary venue through which students learn, the classroom. The current study explores the link between classroom practices and student academic performance by applying multilevel modeling to the 1996 National Assessment of Educational Progress in mathematics. The study finds that the effects of classroom practices, when added to those of other teacher characteristics, are comparable in size to those of student background, suggesting that teachers can contribute as much to student learning as the students themselves.

Key words: instructional practices, NAEP, mathematics, teacher quality

Acknowledgements

The author thanks the Milken Family Foundation which, through its generous financial support, made the research upon which this article is based possible. The author also thanks Paul Barton, Daniel Eignor, Claudia Gentile, Drew Gitomer, Robin Henke, Nancy Knapp, John Mazzeo, and Lewis Solmon for their helpful input at various stages of the project. An earlier version of this paper was presented at the Annual Meeting of the American Educational Research Association in Seattle, Washington, in April 2001. The views expressed are solely those of the author.

Introduction

Much of the discussion in educational reform hinges on the question of whether schools matter. Over the past two decades, policymakers have called for improvements in the academic performance of U.S. students. Many educational reformers, particularly those associated with the standards movement, hold that the key to improving student performance lies in improving schools. If academic standards are rigorous, curriculum and assessments are aligned to those standards, and teachers possess the skills to teach at the level the standards demand, student performance will improve. However, this perspective is to some extent at odds with another that has emerged from the discussion about school improvement, namely that it is students rather than schools that make the difference. Hence a *New York Times* story on how to improve the academic performance of low-income students can include the headline: "What No School Can Do" (Traub, 2000). Or, as Laurence Steinberg puts it in *Beyond the Classroom: Why School Reform has Failed and What Parents Need to Do*, "neither the source of our achievement problem, nor the mechanism through which we can best address it, is to be found by examining or altering schools" (Steinberg, 1996, p. 60). In this view, it is the social backgrounds of students that play the key role in their ability to learn, and only by moving outside of the educational system and attacking the pervasive economic inequalities that exist in the United States can student performance be improved.

Quantitative research on whether schools matter has generally supported the notion that the problems of U.S. education lie outside of the schools. Some research finds that when the social backgrounds of students are taken into account, school characteristics do not seem to influence student outcomes, suggesting that schools do not serve as avenues for upward mobility, but instead reinforce existing social and economic inequalities (Coleman et al., 1966; Jencks et al., 1972). Other researchers contend that school characteristics can have a greater effect on student outcomes than would be expected based upon student background (Lee, Bryk, & Smith, 1993). But while the research in support of this contention does find significant effects for school characteristics, the magnitudes of these effects tend to be modest, far overshadowed by the effects of student background characteristics.[1]

A possible reason for the lack of large school effects in quantitative research is the failure of such research to capitalize on an insight from qualitative research: the central importance of the classroom practices of teachers. As far back as Willard Waller (1932), qualitative researchers have noted that the interaction that occurs between teachers and students in the classroom is greater than the sum of its parts. Students can leave the classroom with their knowledge and attitudes dramatically altered from what they were before they entered. Quantitative research neglects this dimension of schooling by treating it as a “black box,” not worthy of study (Mehan, 1992). Often teaching is not studied at all, and when it is, only the characteristics of teachers that are easily measured but far removed from the classroom (such as their level of educational attainment) are included.

The current study seeks to fill this gap in the literature by using quantitative methods to study the link between student academic achievement and teacher classroom practices, as well as other aspects of teaching, such as the professional development teachers receive in support of their classroom practices and the more traditional teacher background characteristics, referred to here as “teacher inputs.” Such a study is made possible by the availability of a large-scale nationally representative database, the National Assessment of Educational Progress (NAEP), which includes a comprehensive set of classroom practices along with student test scores and other characteristics of students and teachers. For this study, the 7,146 eighth-graders who took the 1996 assessment in mathematics are studied along with their mathematics teachers. The statistical technique of multilevel structural equation modeling (MSEM) is employed to address the major methodological shortcomings of the quantitative literature, namely the failure to distinguish between school- and student-level effects, to measure relationships among independent variables, and to explicitly model measurement error. The study finds that classroom practices indeed have a marked effect on student achievement and that, in concert with the other aspects of teaching under study, this effect is at least as strong as that of student background. This finding documents the fact that schools indeed matter, due to the overwhelming influence of the classroom practices of their teachers.

Background

Much of the quantitative literature linking school characteristics to student outcomes focuses on the impact of economic characteristics, or school resources. These studies are known as production functions. One of the earliest of these studies was the Equality of Educational Opportunity Study, commonly referred to as the Coleman Report (Coleman et al., 1966). This study applied ordinary least squares (OLS) regression analysis to nationally representative samples of elementary and secondary school students to relate school resources such as per-pupil expenditures to student academic achievement and other outcomes. The study found that, on average, when student background was taken into account, school resources were not significantly associated with student outcomes. Nearly 400 additional production function studies have since been conducted. Meta-analyses tabulating the results of such studies between 1964 and 1994 reached divergent conclusions. Some concluded that these studies showed no consistent relationship between school resources and student achievement (Hanushek, 1989, 1996a, 1996b, 1997), while others concluded that the studies showed a consistent, albeit modest, positive relationship (Greenwald, Hedges, & Laine, 1996; Hedges & Greenwald, 1996; Hedges, Laine, & Greenwald, 1994).[2]

Another line of inquiry into the impact of schooling on students, focusing on the social and organizational characteristics of schools, also emerged from the Coleman Report. This body of research, known as effective schools research, sought to identify common characteristics of schools in which students performed above what would be expected based upon their backgrounds (Austin & Garber, 1985; Brookover, Beady, Flood, Schweitzer, & Wisenbaker, 1979; Edmonds, 1979). While the earliest of these studies tended to be small in scope, later studies using large-scale databases confirmed many of their basic findings (Chubb & Moe, 1990; Lee et al., 1993). These studies found that certain characteristics of schools, such as the leadership qualities of the principal, the disciplinary environment of the school, and the size of the student body, all had an effect on student outcomes. In comparison to student background, however, these effects appeared quite modest.

Much of the quantitative research that focused specifically on teaching conformed to a similar pattern, finding little relationship between teacher inputs and student achievement. The Coleman Report measured seven teacher characteristics: years of experience, educational attainment, scores on a vocabulary test, ethnicity, parents' educational attainment, whether the teacher grew up in the area in which he or she was teaching, and the teacher's attitude toward teaching middle-class students. For most students, this study found these characteristics to explain less than 1% of the variation in student test scores. The findings of the meta-analyses of production function studies were just as mixed for teacher inputs as for other school resources. They found that less than one-third of the studies could document a link between student outcomes and teacher experience, less than one-quarter could do so for teacher salaries, and just 1 in 10 could do so for educational attainment; from such mixed results, the meta-analyses came to divergent conclusions, some suggesting a positive relationship and some suggesting no relationship.

More recent research on teaching has confirmed the lack of a clear relationship between student outcomes and teacher inputs, but with two exceptions: the amount of course work the teacher had pursued in the relevant subject area and the teacher's scores on basic skills tests. Two analyses of large-scale databases revealed that exposure teachers received to college-level courses in the subject they were teaching led to better student performance. Monk (1994) analyzed 2,829 high school students from the Longitudinal Study of American Youth. These students were tested in mathematics and science in 10th, 11th, and 12th grades, and filled out questionnaires on their background characteristics. Their mathematics and science teachers were also surveyed. The study related teacher characteristics to student test scores, taking into account students' earlier test scores, background characteristics, and teacher inputs. The study found that the more college-level mathematics or science courses (or math or science pedagogy courses) teachers had taken, the better their students did on the mathematics and science assessments. The more traditional teacher inputs that had been measured in the earlier production function studies, such as teacher experience or educational attainment, proved unrelated to student achievement. Similar results were obtained in a study by Goldhaber and Brewer (1995). They analyzed data on 5,149 10th-graders, 2,245 mathematics teachers, and 638 schools drawn from the National Educational Longitudinal Study of 1988 (NELS:88). Of the various inputs studied, the only one found to make a difference was the proxy for college-level mathematics course taking, namely whether the teacher had majored in mathematics.

Another series of recent studies suggested that, in addition to the teacher's course work in the relevant subject making a difference, so too did the teacher's proficiency in basic skills as measured by standardized tests. Ferguson (1991) analyzed data on nearly 900 Texas districts, representing 2.4 million students and 150,000 teachers. He related the district average of various teacher inputs to average student scores on a basic skills test, taking into account student background. All of the school variables taken together accounted for from 25% to 33% of the variation in average student test scores, and one input, teachers' scores on the Texas Examination of Current Administrators and Teachers, a basic skills test, accounted for the lion's share of this effect. Similar results were obtained by Ferguson and Ladd (1996) in their study of Alabama school districts. Another district-level analysis, this time of 145 North Carolina school districts (Strauss & Sawyer, 1986), found a relationship between average teacher scores on a licensure test, the National Teacher Examination, and student scores on two different assessments taken by high school juniors, taking into account other school and student characteristics. The Coleman data have even been reanalyzed, finding a link between teacher scores on a vocabulary test and student scores on tests in various subject areas (Ehrenberg & Brewer, 1995). That study aggregated data to the school level, analyzing samples of 969 elementary schools and 256 secondary schools. The study calculated a dependent variable, a "synthetic gain score," as the difference between mean student scores in the sixth and third grades for elementary school students and in the twelfth and ninth grades for high school students. The study related teachers' educational attainment, experience, and scores on a vocabulary test to synthetic gain scores and found only the latter to be consistently related to student performance.

Although large-scale quantitative research studied those aspects of teaching that are easily measurable, such aspects tend to be far removed from what actually occurs in the classroom. To study teacher classroom practices and the kinds of training and support pertinent to these practices that teachers receive, it is necessary to draw primarily on the findings of qualitative research.

The qualitative literature on effective teaching emphasizes the importance of higher-order thinking skills (McLaughlin & Talbert, 1993). Teaching higher-order thinking skills involves not so much conveying information as conveying understanding. Students learn concepts and then attempt to apply them to various problems, or they solve problems and then learn the concepts that underlie the solutions. These skills tend to be conveyed in one of two ways: through applying concepts to problems (applications) or by providing examples or concrete versions of the concept (simulations). In either case, students learn to understand the concept by putting it in another context. In the case of an application, this might mean solving a unique problem with which the student is unfamiliar. In the case of a simulation, this might mean examining a physical representation of a theorem from geometry or engaging in a laboratory exercise that exemplifies a law from chemistry. While both lower-order and higher-order thinking skills undoubtedly have a role to play in any classroom, much of the qualitative research asserts that the students of teachers who can convey higher-order thinking skills as well as lower-order thinking skills outperform students whose teachers are only capable of conveying lower-order thinking skills (see also Langer & Applebee, 1987; Phelan 1989).

The qualitative research also emphasizes three additional classroom practices: individualization, collaboration, and authentic assessment. Individualization means that teachers instruct each student by drawing upon the knowledge and experience that that particular student already possesses. Collaborative learning means that teachers allow students to work together in groups. Finally, authentic assessment means that assessment occurs as an artifact of learning activities. This can be accomplished, for instance, through individual and group projects that occur on an ongoing basis rather than at a single point in time (Golub, 1988; Graves & Sunstein, 1992; McLaughlin & Talbert, 1993).[3]

The qualitative research suggests that this set of classroom practices can produce qualitative improvements in the academic performance of all students, regardless of their backgrounds. The focus on higher-order thinking skills is not only appropriate for advanced students; even those in need of more basic skills can benefit from understanding the conceptual basis of these skills. And individualization of instruction does not simply mean using special techniques for low-performing students; techniques developed to address the problems of low-performing students can often help high-performing students as well. Regardless of the level of preparation students bring into the classroom, the qualitative research asserts, decisions that teachers make about classroom practices can either greatly facilitate student learning or serve as an obstacle to it.

Qualitative studies are, by their nature, in-depth portraits of the experiences of specific students and teachers. As such, they provide valuable insight into the interrelationships between various aspects of teacher practice and student learning. However, because they focus on one specific setting, it is difficult to generalize the results of these studies to broader groups of students and teachers. This suggests the need for large-scale quantitative studies that can test the generalizability of the insights from qualitative research.

Yet there has been little quantitative research into whether classroom practices, in concert with other teacher characteristics, have an impact on student learning that is comparable in size to that from background characteristics. Two notable exceptions are a study of the classroom experiences of the nation’s students using NELS:88 (National Center for Education Statistics, 1996) and a study of the professional development experiences and classroom practices of

California's teachers (Cohen & Hill, 2000). The NELS:88 study related a few classroom practices to student achievement in mathematics and science and found that a focus on higher-order thinking skills had a positive effect in math but not in science. The California study related a few professional development experiences of teachers to their classroom practices, and related both of these to student scores on the state assessment. The study found positive relationships between reform-oriented classroom practices and student achievement, as well as between reform-oriented professional development and reform-oriented classroom practices, although these relationships were marginal (mostly significant at the .15 level). While these two studies represent an important departure from production function studies, in their inclusion of measures of classroom practice and professional development, the usefulness of their findings is limited by their data and method. The measures of classroom practice in the NELS:88 and California databases are hardly comprehensive. Neither database has, among other things, a measure of hands-on learning activities. And the California study combines its few classroom practices into two variables, reform-minded and traditional practice, making it difficult to gauge the effectiveness of particular practices. The NELS:88 data also lack measures of most aspects of professional development, and hence professional development was not included in the NELS:88 study. The California data lack measures of social background for individual students, and hence the California study relied upon the percentage of students in the school who received a free or reduced-price lunch, a weak measure. The two studies also relied upon regression analysis, which, as shall be seen, is problematic in the study of school effects.

These two exceptions notwithstanding, quantitative research has tended to find that the effects of student background on student achievement and other outcomes far overshadows school effects. Some of the research has found no school effects at all, while other research has found effects that are, at best, modest. Specifically in terms of teaching, such research has found that most characteristics of teachers do not matter, and the few that do are not as important as student background. Yet such studies ignore qualitative work that suggests that certain classroom practices are highly conducive to student achievement. If this is the case, then classroom practices may indeed explain a substantial portion of the variance in student achievement. The current study seeks to explore this possibility through the analysis of a national database that includes an unprecedentedly comprehensive set of classroom practices.

Hypotheses, Data, and Method

The study tests two hypotheses concerning teacher quality. Teacher quality has three aspects: the teacher's classroom practices, the professional development the teacher receives in support of these practices, and characteristics of the teacher external to the classroom, such as educational attainment. The first hypothesis is that, of these aspects of teacher quality, classroom practices will have the greatest impact on student academic performance, professional development the next greatest, and teacher inputs the least. The rationale for this expectation is that the classroom is the primary venue in which students and teachers interact; hence, decisions by teachers as to what to do in this venue will most strongly affect student outcomes. Teacher inputs will be least likely to influence student academic performance because they do so less directly, through encouraging classroom practices conducive to high student performance. Professional development falls somewhere between classroom practices and teacher inputs. It does occur outside the classroom, but is more closely tied to specific classroom practices than are teacher inputs. Second, it is hypothesized that teacher quality is as strongly related to student academic performance as student background characteristics. When the effects for all three aspects of teacher quality are added together, the result will be comparable in size to that of student background. The rationale behind this expectation is that, as the qualitative literature suggests, student learning is a product of the interaction between students and teachers, and both parties contribute to this interaction.

To test these hypotheses, this study makes use of NAEP, which can measure all three aspects of teacher quality as well as student performance and other potential influences on student performance. NAEP is administered every year or two in various subjects to nationally representative samples of fourth-, eighth-, and twelfth-graders. The subjects vary, but have included at one time or another mathematics, science, reading, writing, geography, and history. In addition to the test itself, NAEP includes background questionnaires completed by the student, the principal, and the teacher in the relevant subject area. The results from NAEP are used to measure trends in student performance over time and to compare performance among various subgroups of students, such as males and females (for an overview of NAEP, see Johnson, 1994).

For this study, data on the 7,146 eighth-graders who took the 1996 mathematics assessment are analyzed. Eighth-graders are used for this analysis because they are exposed to a wider range of subject matter than fourth-graders, and teacher questionnaires are not available for twelfth-graders. Student performance is measured from test scores on the assessment. Student background is measured utilizing six questions from the student background questionnaire: the father's level of education, the mother's level of education, whether there are 25 or more books in the home, whether there is an encyclopedia in the home, whether the family subscribes to a newspaper, and whether the family subscribes to a magazine. The three aspects of teacher quality are measured from a background questionnaire, completed by the mathematics teacher. Three teacher inputs are measured: the teacher's education level, whether the teacher majored or minored in the relevant subject area (mathematics or math education), and the teacher's years of experience. Ten measures of professional development are used: the amount of professional development teachers received last year and whether teachers received any professional development in the past five years in the topics of cooperative learning, interdisciplinary instruction, higher-order thinking skills, classroom management, portfolio assessment, performance-based assessment, cultural diversity, teaching special-needs students, and teaching limited-English-proficient (LEP) students. Finally, 21 classroom practices are utilized: addressing algebra, addressing geometry, addressing unique problems, addressing routine problems, using textbooks, using worksheets, having students talk about mathematics, having students write reports, having students solve problems that involve writing about math, having students work with objects, having students work with blocks, having students solve real-world problems, having students hold discussions in small groups, having students write a group paper, having students work with partners, assessing student progress from tests, assessing student progress from multiple-choice tests, assessing student progress from tests involving constructed responses, assessing student progress from portfolios, assessing student progress from individual projects, and the amount of homework assigned. One school characteristic not pertaining to teacher quality is also drawn from the teacher questionnaire, the number of students in the class (see Tables 1-3 for a complete list of variables).

The method employed in this study is intended to address key methodological problems in the prior literature. Much of school effects research (including most production function studies as well as the NELS:88 and Cohen & Hill studies of classroom practice) relies upon OLS regression techniques. One problem with such techniques is that they are not sensitive to the multilevel nature of the data. School effects involve relating variables at one level of analysis, the school, to another level of analysis, the student. Studies using OLS tend either to aggregate student data to the school level or to disaggregate school data to the student level. The first approach can introduce aggregation biases into the models, the second approach can seriously underestimate standard errors, and both approaches can miss important information about the nature of the school effects (Bryk & Raudenbush, 1992; Goldstein, 1995). A second problem with regression techniques is their failure to take measurement error into account. These techniques assume that the variables in the models are perfectly measured by the observed data. Yet the operationalizations of most variables are subject to substantial error, both because the operationalization does not correspond perfectly to the model (e.g., parents' income as a proxy for socioeconomic status) and because data-collection procedures are error-prone. Failing to take measurement error into account can lead to biased estimates of model coefficients. A third problem is that regression techniques are not adept at measuring interrelationships among independent variables. School effects often involve a multistep process in which one school characteristic influences another that may, in turn, influence the outcome of interest. While it is possible to run a series of models that regress each independent variable on the others, such models tend to be cumbersome and lack statistics measuring the overall fit of the series of models. Because of these difficulties, school effects research often neglects the indirect effects of various school characteristics.

One way to address these problems is through the technique of multilevel structural equation modeling (MSEM). Structural equation modeling (SEM) involves two components: factor models and path models (Hayduk, 1987; Jöreskog & Sörbom, 1993). The factor models relate a series of indicators, known as manifest variables, to a construct of those indicators, known as a latent variable. The path models then relate the latent variables to one another. The estimation procedure for both the factor and path components involves three steps. A set of hypothesized relationships is specified by the researcher. Then, through an iterative process, differences in the covariance matrix those relationships imply ($\mathbf{\Sigma}$) and the covariance matrix of observed data ($\mathbf{S}$) are minimized. The resulting estimates include coefficients for the hypothesized relationships, *t* tests for their statistical significance, and statistics for the goodness of fit between $\mathbf{\Sigma}$ and $\mathbf{S}$. SEM can be adapted to handle multilevel data by employing the estimation procedure separately for the two levels of analysis (Muthén, 1991, 1994). The researcher hypothesizes a student-level factor model, a student-level path model, a school-level factor model, and a school-level path model. These models can be used to generate two implied covariance matrices, $\mathbf{\Sigma_{B}}$, a between-school matrix computed as the school deviations around the grand mean, and $\mathbf{\Sigma_{W}}$, a within-school matrix computed as student deviations around group means. The observed data can be similarly partitioned into between-school and within-school covariance matrices ($\mathbf{S_{B}}$ and $\mathbf{S_{W}}$).

MSEMs can address the three problems in the prior literature. First, they do distinguish between schools and students; separate models are specified for each level of analysis and related to one another through a constant. Second, these models take measurement error into account in two ways. For one, the factor models explicitly measure the amount of variance in the latent variables unexplained by the manifest variables. In addition, factor models can actually reduce measurement error by generating latent variables from multiple manifest variables. Third, the path models estimate interrelationships among independent variables, allowing for the estimation of indirect effects. The effect sizes and *t* scores of the indirect effects are produced, as well as statistics that measure the overall goodness of fit of models that simultaneously specify these interrelationships.[4]

The current study produces three MSEMs. Analyses are conducted using AMOS 3.6 (Arbuckle, 1997), an SEM software package, along with STREAMS 1.8, a pre- and post-processor that simplifies the syntax and output for multilevel models (Gustafsson & Stahl, 1997). In preparation for the preprocessor, the preexisting student-level data variable labels are reduced to six characters, and missing values are replaced with means for the pertinent variable. The software then aggregates the student-level data to the school level, and creates both a school-level covariance matrix and a pooled matrix of residual student-level covariances.[5] The first MSEM relates teacher inputs to student academic performance, taking into account student socioeconomic status (SES) and class size (Figure 1). The student-level factor model generates an SES construct from the six measures of student background, and an academic performance construct from a single test score. The student-level path model simply measures the covariance between SES and student academic performance. The school-level factor model generates an SES construct from school means of the six measures of student background and an academic construct from the school mean of the single test score. In addition, class size, teachers' years of experience, educational attainment, and major are constructed from individual measures that correspond to these constructs. The school-level path model treats student academic performance as a function of the other constructs.

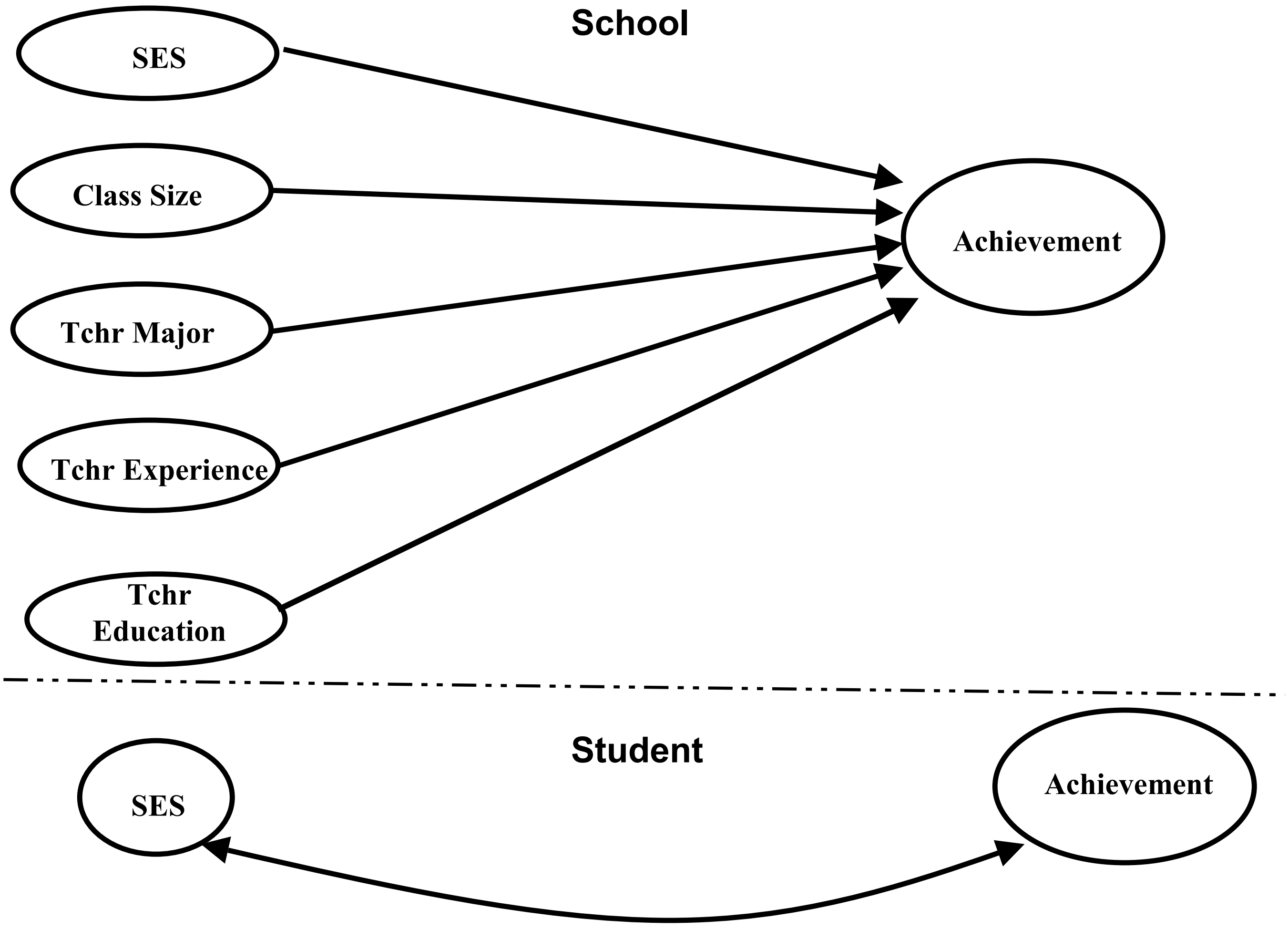

Figure 1.
Teacher inputs path model

The second MSEM relates professional development and teacher inputs to student academic performance and one another, taking into account student SES and class size (Figure 2). The student-level factor and path models are the same as in the teacher inputs model. Early versions of the school-level factor and path models include SES and student academic performance, constructed as before; teacher inputs, which prove significantly related to student academic performance, constructed from a single corresponding measure; the amount of time in professional development, constructed from a single corresponding measure; and all nine professional development topics. For the sake of parsimony, the final school-level factor and path models include only those professional development topics significantly related to student academic performance. These are professional development in higher-order thinking skills, constructed from a single corresponding measure, and professional development in teaching different populations of students, constructed from professional development in cultural diversity, professional development in teaching LEP students, and professional development in teaching students with special needs. The parsimonious school-level path model relates each professional development construct to student achievement, and the teacher input, class size, and SES both to student achievement and to each professional development construct.

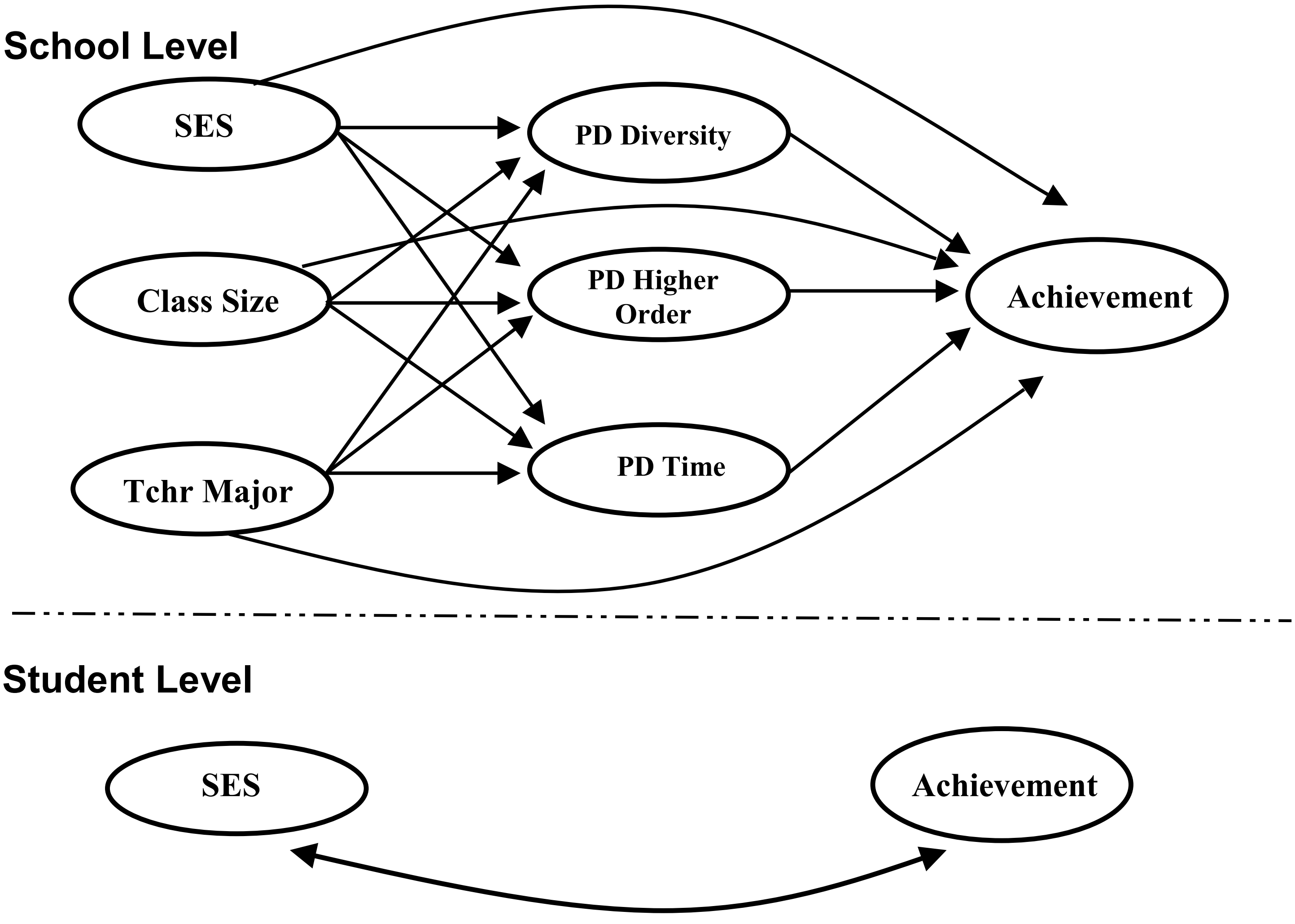

Figure 2.
Professional development path model

The third MSEM relates classroom practices, professional development, and teacher inputs to student academic performance and to one another, taking into account SES and class size (Figure 3). Student-level factor and path models remain the same as in prior models. Early versions of the school-level factor and path models include SES, class size, teacher inputs that prove significant in the teacher input model, the amount of time in professional development, the topics of professional development that prove significant in the professional development model, and all 21 classroom practices. For the sake of parsimony, the final school-level factor and path models include only those classroom practices that prove significantly related to student achievement. The final school-level factor model constructs teaching of higher-order thinking skills from a single measure (solving unique problems), teaching of lower-order thinking skills

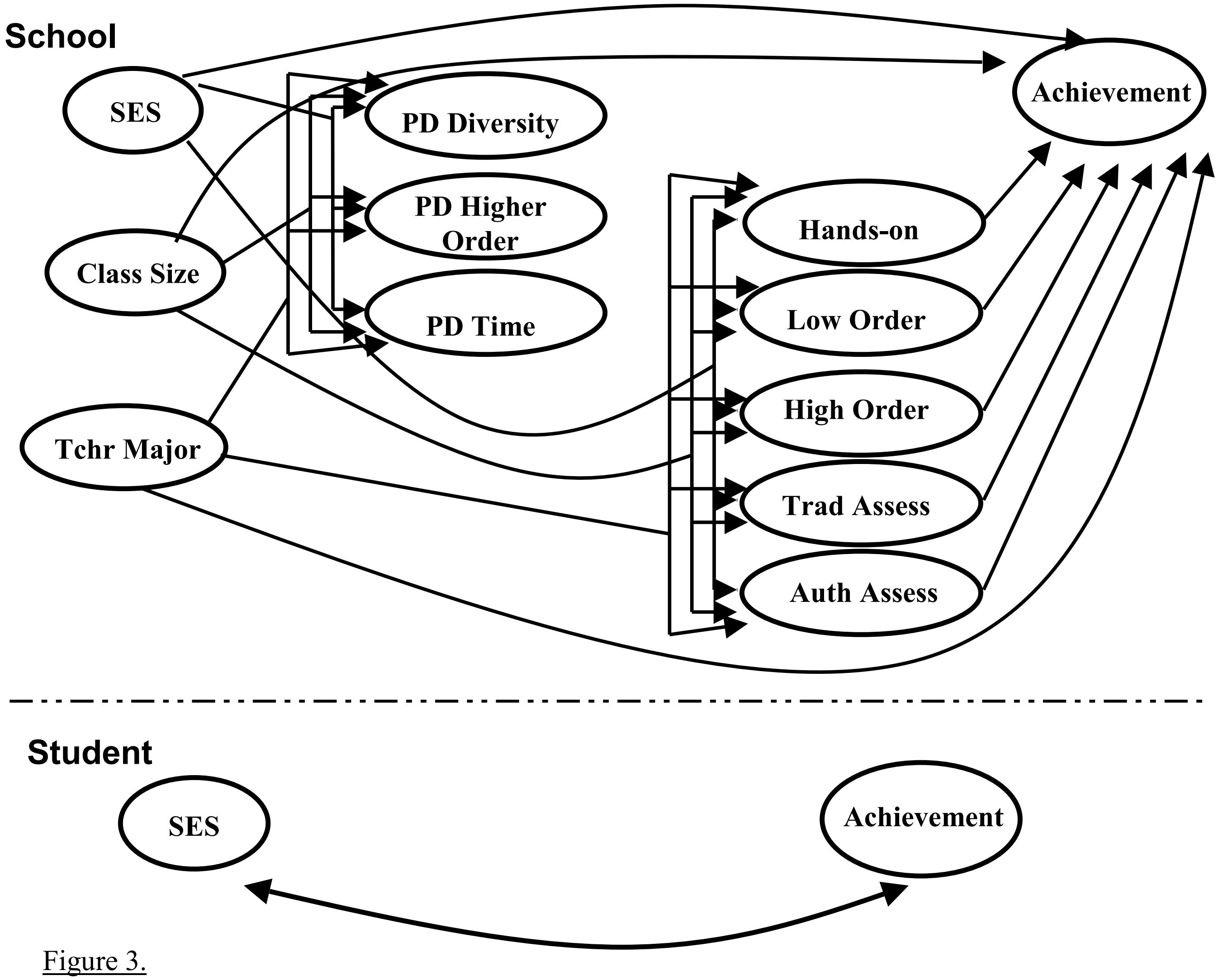

Figure 3.
Classroom practice model

from a single measure (solving routine problems), engaging in hands-on learning from three measures (working with blocks, working with objects, and solving real-world problems), assessing student progress through traditional testing from two measures (multiple-choice testing and the overall frequency of testing), and assessing student progress through more authentic assessments from three measures (portfolio assessments, individual projects, and constructed-response tests). The SES, class size, teacher input, and professional development constructs are handled as in the professional development model. The school-level path model relates these classroom practice constructs to the student achievement constructs, relates the professional development constructs to the classroom practice constructs, and relates teacher inputs, SES, and class size to the professional development classroom practice and student achievement constructs.

These procedures are modified in two ways to take the design of NAEP into account. First, design effects are employed. NAEP is a stratified, clustered sample. Secondary analyses of NAEP that treat it as a simple random sample will underestimate standard errors, making significance tests overly liberal. One procedure recommended to address this problem is to inflate standard errors estimated assuming a simple random sample by a certain factor, known as a design effect (O'Reilly, Zelenak, Rogers, & Kline, 1996). This study uses a design effect of 2, calculated by estimating the proper standard error for select values in the first MSEM and choosing the most conservative one. Cutoff points for all significance tests are increased by 41% (the increase in standard errors attributable to the square root of the design effect).[6] Second, each MSEM is estimated multiple times, once for each "plausible value" of the student test score, and the resulting parameters and standard errors are pooled. Because each student answers only a small subset of the assessment items, it is not possible to estimate a single student score. Instead, five estimates are provided based upon the items the student did not answer and background information about the student and the school. The appropriate procedure for secondary analyses using these five estimates, which are known as plausible values, is to estimate five separate models for each of the plausible values, pool their point estimates by taking their means, and pool their standard errors as the sum of the mean standard error and the variance among the five plausible values, weighted by a factor of 1.2 (Johnson, Mislevy, & Thomas, 1994).[7] The current study employs this technique, producing a total of 15 sets of estimates, five for each of the MSEMs.[8]

Results

Before discussing the results from the three MSEMs, it is worthwhile to summarize what the NAEP data reveal about the prevalence of classroom practices, professional development, and teacher inputs.[9] The data on teacher inputs indicates that eighth-grade math teachers are most likely to possess less than a master's degree, have majored or minored in mathematics or math education, and have 10 or more years of experience teaching (Table 1). Approximately 40% of eighth-graders have teachers who possess a master's degree or more, with the remainder possessing a bachelor's degree or less. Approximately 70% of eighth-graders have teachers who majored or minored in mathematics or math education; the rest have teachers who are teaching off-topic. And approximately 60% of eighth-graders have teachers with more than 10 years of experience.

The data on professional development indicate that while most teachers receive some professional development in some topics, that professional development tends not to be of long duration, and certain topics tend to be neglected (Table 1). Most eighth-graders have teachers who have received some professional development within the past five years in the most common topics, such as cooperative learning or interdisciplinary instruction. But only one-third of eighth-graders have teachers who received professional development in cultural diversity, one-quarter have teachers who received professional development in teaching students with special needs, and one-tenth have teachers who received professional development in teaching LEP students. And regardless of the topic of professional development, only a minority of students have teachers who received more than 15 hours of professional development last year.

The prevalence of classroom practices varies greatly (Table 2). While much of the material covered in eighth grade involves issues of operations and measurement, teachers do cover more advanced topics. More than half of all students are exposed to algebra, and one-quarter to geometry. The kinds of problems students are taught to solve tend to involve a routine set of algorithms; four out of five students commonly work with such problems, as opposed to about half of students working with problems that involve unique situations. All students report taking a math test at least once a month. The nature of the test varies, however. Typically, students take tests that involve extended written responses (more than half do so at least once a month). About one-third of students take multiple-choice tests. Students are also assessed through individual projects and portfolios (also about one-third of students at least once a month). Hands-on learning activities appear quite infrequent. Just one-quarter of students work with objects and just one-tenth work with blocks. Problems with a concrete or practical bent, that address real-world situations, are fairly usual, however, with three-quarters of students encountering such problems as least once a week. Writing about mathematics is fairly uncommon, with just one-third of students doing so at least once a week. Group activities vary in their frequency; most students discuss math in small groups, but only a minority of students solve problems in groups or work on a problem with a partner. Finally, textbooks and homework are ubiquitous in eighth-grade classrooms; nearly all students use a textbook at least once a week, and most do some homework every day.

This description of teacher inputs, professional development, and classroom practices says little about their effectiveness. The fact that certain practices are uncommon may be bad or good, depending upon their impact on student outcomes. It is the role of the series of MSEMs to gauge the effectiveness of these three aspects of teacher quality.

For all three MSEMs, the student-level factor models are similar (Table 4). The factor models show the two student-level characteristics—SES and achievement—to be well measured. All of the indicators of SES have standardized factor loadings ranging from .24 to .33, suggesting that each plays a role in constructing the variable. The construct for achievement consists of a single indicator, and hence has a loading fixed at 1 and an error fixed at 0. The path model consists simply of the covariance between student SES and student achievement, and this covariance proves significant, with a correlation coefficient of .35 for all models. It should be remembered that this covariance pertains only to the student-level component of the models, meaning that variations in SES among students in the same school are associated with variations in their mathematics scores within that same school. Variations in average SES and achievement between schools is the purview of the school-level models.

The school-level factor models also have indicators that contribute substantially to their constructs (Table 5). The loadings for SES range between .17 and .25.[10] Hands-on learning has loadings ranging from .46 to .79. Traditional assessment has loadings ranging from .37 to .57. And authentic assessment has loadings ranging from .41 to .73. All of the constructs generated from a single indicator have loadings fixed at 1 and errors fixed at 0 and so, by definition, their indicators contribute substantially. The one construct for which the indicators do not all contribute substantially is professional development in teaching special populations. Here, two of the indicators (cultural diversity and teaching LEP students) load strongly on the construct, but the third (teaching special-needs students) does not. (A sensitivity analysis was conducted in which this indicator was excluded, without significant impact on the model.)

Table 1

Descriptive Statistics for Teacher Inputs and Professional Development

Teacher Inputs	**M**	**SD**
Teacher's Education Level (From 1=<B.A. to 4=>M.A.)	2.38	.46
Teacher Majors in Mathematics (1=yes, 0=no)	.69	.43
Teacher's Years of Experience (From 1= 2 or less to 5=25 or more)	3.53	1.17
Professional Development		
Classroom Management (1=yes, 0=no)	.44	.46
Cooperative Learning (1=yes, 0=no)	.68	.44
Cultural Diversity (1=yes, 0=no)	.32	.43
Higher-Order Thinking Skills (1=yes, 0=no)	.45	.46
Interdisciplinary Instruction (1=yes, 0=no)	.50	.47
Limited English Proficiency (1=yes, 0=no)	.12	.47
Performance-Based Assessment (1=yes, 0=no)	.12	.35
Portfolio Assessment (1=yes, 0=no)	.36	.45
Special-Needs Students (1=yes, 0=no)	.26	.41
Time in Professional Development Last Year (From 1=none to 5=35+ hours)	3.30	1.14

Table 2

Descriptive Statistics for Classroom Practices

Classroom Practices	**M**	**SD**
Address Algebra (From 1=none to 4=a lot)	2.51	.59
Address Geometry (From 1=none to 4=a lot)	2.00	.61
Address Solving Routine Problems (From 1=none to 4=a lot)	2.78	.43
Address Solving Unique Problems (From 1=none to 4=a lot)	2.44	.56
Assessment Using Multiple-Choice Questions (From 1=never to 4= a lot/ twice a week)	1.99	.83
Assessment Using Short/Long Answers (From 1=never to 4= a lot/twice a week)	2.49	.92
Assessment Using Portfolios (From 1=never to 4= a lot/twice a week)	1.87	.79
Assessment Using Individual Projects (From 1=never to 4= a lot/twice a week)	2.19	.81
Work with Blocks (From 1=never to 4=almost everyday)	1.52	.58
Work with Objects (From 1=never to 4=almost everyday)	2.09	.77
Solve Real-Life Problems (From 1=never to 4=almost everyday)	2.93	.74
Write Reports (From 1=never to 4=almost everyday)	1.39	.49
Write about Math (From 1=never to 4=almost everyday)	1.97	.79
Take Math Tests (From 1=never to 4=almost everyday)	2.49	.47
Do Worksheets (From 1=never to 4=almost everyday)	2.65	.82
Talk about Math (From 1=never to 4=almost everyday)	2.70	1.02
Solve Problems with Other Students (From 1=never to 4=almost everyday)	2.84	.83
Discuss Math with Other Students (From 1=never to 4=almost everyday)	3.31	.70
Work with Partner (From 1=never to 4=almost everyday)	2.98	.82
Do Homework (From 1=never to 4=almost everyday)	2.93	.75
Use Textbooks (From 1=never to 4=almost everyday)	3.63	.65

Table 3

Descriptive statistics for other characteristics of schools and students

Other Characteristics of Schools and Students	**M**	**SD**
Class Size (From 0=more than 36 student to 4=1 to 20 students)	2.54	.88
Student’s Family Gets Newspaper (1=yes, 0=no)	.74	.43
Student’s Family Has Encyclopedia (1=yes, 0=no)	.82	.38
Student’s Family Gets Magazine (1=yes, 0=no)	.83	.37
Student’s Family Has More than 25 Books (1=yes, 0=no)	.95	.21
Father’s Education Level	2.91	.94
Mother’s Education Level	2.85	.96
Math Score: Plausible Value #1	272.45	35.85
Math Score: Plausible Value #2	272.64	35.89
Math Score: Plausible Value #3	272.36	36.35
Math Score: Plausible Value #4	272.45	35.85
Math Score: Plausible Value #5	272.54	35.56

Table 4

Student-Level Factor and Path Models

Factor Model	**Input Models**			**P.D. Model**			**Practices Model**		
	SES	**Ach**	**Err**	**SES**	**Ach**	**Err**	**SES**	**Ach**	**Err**
Mother's Education Level	2.91* .31		1.00 .86	2.92* .31		1.00 .86	2.91* .31		1.00 .86
Father's Education Level	2.78* .31		1.00 .85	2.80* .31		1.00 .84	2.79* .31		1.00 .85
Family Gets Newspaper	1.00* .24		1.00 .92	1.00* .24		1.00 .92	1.00* .24		1.00 .92
Family Has Encyclopedia	.92* .26		1.00 .94	.92* .26		1.00 .94	.92* .26		1.00 .94
Family Gets Magazine	1.16* .33		1.00 .90	1.16* .33		1.00 .90	1.16* .33		1.00 .92
Family Has More than 25 Books	.60* .30		1.00 .92	.60* .30		1.00 .92	.60* .30		1.00 .85
Plausible Value #1		1.00* .77	1.00 .28		1.00* .77	1.00 .28		1.00* .77	1.00 .28
Plausible Value #2		.99* .77	1.00 .28		.99* .77	1.00 .28		.99* .77	1.00 .28
Plausible Value #3		1.00* .77	1.00 .28		1.00* .77	1.00 .28		1.00* .77	1.00 .28
Plausible Value #4		.99* .77	1.00 .28		.99* .77	1.00 .28		.99* .77	1.00 .28
Plausible Value #5		.98* .77	1.00 .28		.98* .77	1.00 .28		.98* .77	1.00 .28
Path Model									
Covariance between SES and Achievement	1.15* .35			1.15* .35			1.15* .35		

Note. Cells contain unstandardized and standardized coefficients, in that order.

*p<.05.

Table 5

School-level Factor Model: Classroom Practices

	SES	Ach	Class Size	PD Time	PD Hi Order	PD Diversity	Tchr Major	Error
Mother's Education	2.66* .23							1.00 .23
Father's Education	2.89* .25							1.00 .25
Newspaper	1.00* .19							1.00 .16
Encyclopedia	.78* .17							1.00 .07
Books	.50* .20							1.00 .10
Magazine	1.05* .23							1.00 .10
Plausible Value #1		1.00* .57						1.00 .03
Plausible Value #2		1.00* .58						1.00 .02
Plausible Value #3		1.00* .58						1.00 .03
Plausible Value #4		1.01* .58						1.00 .04
Plausible Value #5		1.00* .58						1.00 .02
Class Size			1.00* 1.00					1.00 .00
PD Time				1.00* 1.00				1.00 .00
PD Hi order					1.00* 1.00			1.00 .00
PD Cultural						1.00* .76		1.00 .46
PD LED						.65* .55		1.00 .59
PD Special						.26* .20		1.00 .69
Tchr Major							1.00* 1.00	1.00 .00

Note. Cells contain unstandardized and standardized coefficients, in that order.

*p<.05.

(table continues)

Table 5

School-level Factor Model: Classroom Practices (continued)

	Hands-On Learning	**Trad Assess**	**Auth Assess**	**Lower Order**	**Higher Order**	**Error**
Real-world Problems	.64* .46					1.00 .63
Work with Objects	1.00* .66					1.00 .53
Work with Blocks	.83* .79					1.00 .43
Take Tests		.35* .37				1.00 .66
Assess through Multiple-Choice Tests		1.00* .57				1.00 .58
Assess through Extended Response Tests			1.01* .65			1.00 .54
Assess through Projects			1.00* .73			1.00 .48
Assess through Portfolios			.63* .41			1.00 .65
Address Routine Problems				1.00* 1.00		1.00 .00
Address Unique Problems					1.00* 1.00	1.00 .00

Note. Cells contain unstandardized and standardized coefficients, in that order.
*p<.05.

The school-level path model for teacher inputs shows that one of the three inputs, the teacher's major, is modestly associated with academic achievement. The model consists of a single dependent variable, achievement, related to five independent variables, SES, class size, and the three teacher inputs (Table 6). SES has an effect size of .76, which far overshadows those of class size and teacher's major (.10 and .09, respectively). The teacher's level of education and years of experience prove unrelated to student achievement.

Table 6

School-level Path Model: Teacher Inputs

	Ach
SES	198.41** .76
Class Size	3.04* .10
Tchr Major	4.82** .09
Tchr Ed	1.20 .02
Tchr Exp	1.03 .05
Error	1.00 .44

Note. Cells contain unstandardized and standardized coefficients, in that order.
*p<.10. **p<.05.

The school-level path model for professional development finds that two topics, addressing special populations of students and higher-order thinking skills, are substantially related to student achievement. The model indicates that schools with high percentages of affluent students tend to have less time spent on professional development generally, and are less likely to expose their teachers to professional development on working with different student populations (Table 7). Schools with smaller average class sizes are also less likely to do these things. But, schools with more teachers teaching on topic also devote more time to professional development. Of the three aspects of professional development, the amount of time is not significantly related to achievement. Professional development in higher-order thinking skills and dealing with special populations, however, do have significant effects, with standardized coefficients of .12 and .21, respectively.

Table 7

School-level Path Model: Professional Development

	PD Diversity	PD Hi Order	PD Time	Ach
SES	-1.29** -.32	-.58 -.09	-1.87* -.12	213.18** .83
Class Size	-.08** -.17	-.04 -.06	-.20* -.11	4.23** .14
Tchr Major	-.01 -.01	.11 .08	.94** .27	5.05** .09
PD Diversity				13.24** .21
PD Hi order				4.88** .12
PD Time				-.23 -.01
Error	1.00 .66	1.00 .70	1.00 .67	1.00 .41

Note. Cells in the tables above contain unstandardized and standardized coefficients, in that order.

*p<.10. **p<.05.

The school-level path model for classroom practices finds three constructs—hands-on learning, solving unique problems, and avoiding reliance on authentic assessments—to be positively related to student achievement (Table 8). All five of the classroom practice constructs are related to some of the earlier variables (SES, class size, teacher major) or to the three aspects of professional development. Schools with more affluent students are more likely to solve unique problems and less likely to engage in inauthentic forms of assessment. Schools where teachers received professional development in dealing with different student populations are less likely to have students engage in routine problem solving. And schools where teachers received professional development in higher-order thinking skills are more likely to have students engage in hands-on learning. Also, the more time teachers engage in professional development, the more their students engage in hands-on learning and authentic assessment. These practices are associated with student achievement. Schools where students engage in hands-on learning score higher on the mathematics assessment. Schools where students solve unique problems also score higher, as do those schools that do not rely primarily on authentic forms of assessment.

Comparisons among the three school-level path models help to gauge the impact of teaching on student achievement. First, all of the models explain a similar amount of variance. While the residual variance goes from .44 in the teaching model to .41 in the professional development model and .40 in the classroom practices model, these differences are slight. Thus, rather than explain more variance, the more complex models simply reallocate variance among explanatory variables. Second, the three models show the total effect of each teacher quality variable. The total effect is the sum of all direct and indirect effects, and is measured for each aspect from the sum of the effect sizes of the variables in directing that aspect in the model in which that aspect is related to achievement without mediating variables.[11] Thus, the effect size of the one significant teacher input is .09, taken from the teacher inputs model; the effect sizes for the statistically significant aspects of professional development total is .33, taken from the professional development model; and the effect sizes for the classroom practices total is .56, taken from the classroom practices model. Third, all of the models fit the data well, with goodness of fit indices at the .99 or 1.00 level and root mean squared errors of approximation at the .014 level or better.

In sum, it appears that the various aspects of teacher quality are related to student achievement when class size and SES are taken into account. In particular, the following five variables are positively associated with achievement:

- Teacher major
- Professional development in higher-order thinking skills
- Professional development in diversity
- Hand-on learning
- Higher-order thinking skills

Before discussing further the implications of these results, however, it is necessary to note some shortcomings of the study.

Table 8

School-level Path Model: Classroom Practices

	PD Diversity	PD Hi Order	PD Time	Hands-On Learning	Lower Order	Higher Order	Trad Assess	Auth Assess	Ach
SES	-1.11** -.05	-.43 -.07	-1.83* -.12	-1.02 .14	-.32 -.06	1.15** .17	-2.35** -.34	-.79 -.09	192.26** .74
Class Size	-.09** -.17	-.04 -.06	-.20* -.11	.03 .03	.03 .05	.03 .03	.01 .01	-.09 -.10	2.33 .08
Tchr Major	-.01 -.01	.11 .08	.94** .27	-.04 -.02	-.06 -.05	.02 .01	.19 .13	-.05 -.03	4.19* .07
PD Diversity				-.23 -.14	-.30* -.24	-.24 -.16	-.13 -.09	-.14 -.07	
PD Hi order				.34** .30	.01 .01	.12 .11	.21 .19	.23* .18	
PD Time				.13** .27	.03 .08	.05 .12	-.14** -.32	.14** .26	
Hands-On Learning									8.88** .25
Lower Order									-3.85 -.08
Higher Order									4.82** .13
Trad Assess									1.23 .03
Auth Assess									-5.73** -.18
Error	1.00 .67	1.00 .70	1.00 .67	1.00 .63	1.00 .68	1.00 .68	1.00 .62	1.00 .66	1.00 .40

Note. Cells contain unstandardized and standardized coefficients, in that order.
*p<.10. **p<.05.

Methodological Caveats

The study suffers from four basic shortcomings. First, the data are cross-sectional. The information about aspects of teacher quality is collected at the same time as student test scores. Consequently, it is not possible to draw inferences about the direction of causation for the relationships that were discovered. It may be that a focus on higher-order thinking skills causes increased student performance, or it may be that having high-performing students drives teachers to focus on higher-order thinking skills. The likelihood of the latter scenario is somewhat reduced in that the models take SES and class size, both proxies of prior academic performance of the student and school, into account. Nonetheless, to confirm the causal direction hypothesized in this study, subsequent research should replicate the results using longitudinal data.

Second, the study covers only one grade level in one subject. It is possible that different sets of classroom practices will prove effective for other subjects and at other grade levels. Third, this study does not measure the link between aspects of teacher quality and the relationship between student test scores and student SES. MSEM measures student-level covariances by pooling each school's within-school covariance matrix. Consequently, while it is possible to measure the relationship between a school variable and a student outcome, it is not possible to measure the relationship between a school variable and the relationship between two student characteristics. Other multilevel techniques, such as Hierarchical Linear Modeling, while unable to perform certain analyses that MSEM can perform (e.g., confirmatory factor analysis), are able to accomplish this. Subsequent research should supplement the findings of this study by measuring the impact of classroom practices and other aspects of teacher quality on the relationship between student test scores and student background characteristics. Such analyses will make it possible to know not only how teachers can affect the average performance of their class, but how they can affect the distribution of performance within the class.

Finally, better indicators of the constructs used in this study are needed. The SES construct lacks indicators of parents' income or occupation, as well as noneducational materials in the home, such as a microwave or washer and dryer, indicators that prior research has found to be an important component of SES. Exposure to each topic of professional development is measured as whether the teacher had received any exposure within the past five years, making it impossible to distinguish between professional development that is rich and sustained and a lone weekend seminar. Given that professional development in working with different student populations is so important, it would be useful to include a measure of classroom practices that involves this activity. And while many of the classroom practices are measured through multiple indicators, some, such as higher-order thinking skills, are not. Additional indicators for single-indicator constructs should be introduced to increase the reliability of the constructs.[12]

Conclusion

Despite these methodological shortcomings, the current study represents an advance over previous work. The first model to some extent exemplifies the traditional approach to gauging the impact of teaching and other school characteristics on student achievement. Although the model differs from most production function studies in including a measurement component and being multilevel, it is otherwise similar. Like OLS, the model relates a single dependent variable to a series of independent variables. The independent variables consist of teacher inputs and a class-size measure, controlling for student background. Like most of the prior research, this model finds no significant relationship to test scores for most of the characteristics, with the exception of the teacher's college-level course work as measured by major or minor in the relevant field. And like all of the prior research, all school effects are overshadowed by the effect of student SES.

The subsequent models move beyond the first by introducing measures of what teachers actually do in the classroom, the training they receive to support these practices directly, and by modeling interrelationships among the independent variables. They are able to do so because the NAEP database includes a comprehensive set of classroom practices, and because MSEM can model all of the relevant interrelationships. And all of the models, including the teacher inputs model, move beyond most prior research in their ability to take into account measurement error and the multilevel nature of the data. Through these innovations it was possible to confirm the two hypotheses regarding the role that teaching plays in student learning.

The first hypothesis—that, of the aspects of teacher quality, classroom practices will have the greatest effect—is confirmed by the models. The effect sizes for the various classroom practices total .56, those for the professional development topics total .33, and the effect size for the one teacher input found to have a statistically significant impact is .09. As the qualitative literature leads one to expect, a focus on higher-order thinking skills is associated with improved student performance. Applying problem-solving techniques to unique problems is a key component of such skills. Hands-on learning can be understood in this way as well, in that it involves the simulation of concepts, moving the student from the abstract to the concrete. Also suggested by the qualitative literature, individualizing instruction seems to be effective. Students whose teachers received professional development in learning how to teach different groups of students substantially outperformed other students. One apparent inconsistency between the findings of this study and the qualitative literature is in the area of authentic assessment in that the study documents the importance of using some form of traditional testing in assessing student progress. This finding, however, merely suggests that ongoing assessments such as portfolios and projects are not sufficient; they need to be supplemented with tests that occur at a distinct point in time.

The second hypothesis—that the total impact of the teaching variables will be comparable to that of student SES—is also confirmed. The sum of the effects from the three aspects of teacher quality is .98. The effect sizes for SES range from .74 to .83, with a value of .76 in the model where all three aspects of teacher quality are included (the classroom practices model). Thus, the impact of teaching can be said not only to be comparable to that of SES, but even to be somewhat greater.

In addition to confirming the hypotheses regarding the impact of teaching on student learning, the study uncovers important interrelationships among the aspects of teaching. For one, professional development seems to influence teachers' classroom practices strongly. The more professional development teachers received in hands-on learning, and indeed the more professional development they received regardless of topic, the more likely they are to engage in hands-on learning activities. And the more professional development teachers received in working with special student populations, the less likely they are to engage in lower-order activities. Another important interrelationship involves the trade-off between teacher quality and teacher quantity. Smaller class sizes are negatively associated with teachers receiving substantial amounts of professional development, whereas teacher major and time in professional development are positively associated with one another. These relationships suggest that schools tend to choose between hiring more teachers or investing in improved teacher quality through recruiting teachers with better preservice training and providing teachers with more and better in-service training.

In sum, this study finds that schools matter because they provide a platform for active, as opposed to passive, teachers. Passive teachers are those who leave students to perform as well as their own resources will allow; active teachers press all students to grow regardless of their backgrounds. Passive teaching involves reducing eighth-grade mathematics to its simplest components. All lessons are at a similar level of abstraction; problems are solved in a single step and admit of a single solution; and all students are treated as if they had entered the class with the same level of preparation and the same learning styles. In contrast, active teaching does justice to the complexities of eighth-grade mathematics. Lessons work at multiple levels of abstraction, from the most mundane problem to the most general theorem; problems involve multiple steps and allow multiple paths to their solution; and teachers tailor their methods to the knowledge and experience of each individual student. Schools that lack a critical mass of active teachers may indeed not matter much; their students will be no less or more able to meet high academic standards than their talents and home resources will allow. But schools that do have a critical mass of active teachers can actually provide a value-added; they can help their students reach higher levels of academic performance than those students otherwise would reach. Through their teachers, then, schools can be the key mechanism for helping students meet high standards.

References

Arbuckle, J.L. (1997). *Amos users' guide* (Version 3.6). Chicago: Small Waters Corporation.

Austin, G.R., & Garber, H. (Eds.). 1985. *Research on exemplary schools.* New York: Academic Press.

Brookover, W., Beady, C., Flood, P., Schweitzer, J., & Wisenbaker, J. (1979). *School social systems and student achievement: Schools can make a difference.* Brooklyn, NY: J.F. Bergin.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Chubb, J., & Moe, T. (1990). *Politics, markets and America's schools.* Washington, DC: The Brookings Institution.

Cohen, D.K., & Hill, H.C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*(2), 294–343.

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

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership, 37*(1), 15-24.

Ehrenberg, R.G., & Brewer, D.J. (1995). Did teachers' verbal ability and race matter in the 1960s? Coleman revisited. *Economics of Education Review, 14*(1), 1–21.

Ferguson, R.F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal of Legislation, 28*(2), 465–498.

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

Forture, J.C., & O'Neill, J.S. (1994, Summer). Production function analyses and the study of educational funding equity: A methodological critique. *Journal of Education Finance*, *20*, 21-46.

Goldhaber, D.D., & Brewer, D.J. (1995) Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources, 32*(3), 505–520.

Goldstein, H., (1995). *Multilevel statistical models* (2nd ed.). New York: Halsted Press.

Golub, J. (Ed.). (1988). *Focus on collaborative learning.* Urbana, IL: National Council of Teachers of English.

Graves, D.H., & Sunstein, B.S. (Eds.). (1992). *Portfolio portraits.* Portsmouth, NH: Heinemann.

Greenwald, R., Hedges, L.V., & Laine, R.D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361–396.

Gustafsson, J.E., & Stahl, P.A. (1997). *STREAMS user's guide* (*Version 1.7*). Mölndal, Sweden: Multivariate Ware.

Hanushek, E.A. (1989). The impact of differential expenditures on school performance. *Educational Research, 18*(4), 45–51.

Hanushek, E.A. (1996a). A more complete picture of school resource policies. *Review of Educational Research, 66*(3), 397–409.

Hanushek, E.A. (1996b). School resources and student performance. In G.T. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 43–73). Washington, DC: The Brookings Institution.

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

Hayduk, L.A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore: Johns Hopkins University Press.

Hedges, L.V., & Greenwald, R. (1996). Have times changed? The relation between school resources and student performance. In G.T. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 74–92). Washington, DC: The Brookings Institution.

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

Jencks, C., Smith, M., Ackland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1972). *Inequality: A reassessment of the effect of family and schooling in America.* New York: Basic Books.

Johnson, E. (1994). Overview of part I: The design and implementation of the 1992 NAEP. In E. Johnson & J. Carlson (Eds.), *The NAEP 1992 Technical Report* (pp. 9–32). Princeton, NJ: Educational Testing Service.

Johnson, E., Mislevy, R.J., & Thomas, N. (1994). Scaling procedures. In E. Johnson & J. Carlson (Eds.), *The NAEP 1992 Technical Report* (pp. 241–256). Princeton, NJ: Educational Testing Service.

Jöreskog, K.G., & Sörbom, D. (1993). *Structural equation modeling and the SIMPLIS command language.* Chicago: Scientific Software International.

Langer, J.A., & Applebee, A.N. (1987). *How writing shapes thinking.* Urbana, IL: National Council of Teachers of English.

Lee, V.E., Bryk, A.S., & Smith, J.B. (1993). The organization of effective secondary schools. *Review of Research in Education, 19*, 171–267.

Mayer, D.P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29–45.

McLaughlin, M.E,. & Talbert, J.E.(1993). Introduction: New visions of teaching. In M.W. McLaughlin & J.E. Talbert (Eds.), *Teaching for understanding* (pp. 1–10). San Francisco, CA: Jossey-Bass.

Mehan, H. (1992). Understanding inequality in schools: The contribution of interpretive studies. *Sociology of Education, 65*(1), 1–20.

Mislevy, R.J. (1993). Should “multiple imputations” be treated as “multiple indicators”? *Psychometrika*, *58*(1), 79-85.

Monk, D.H. (1992). Educational productivity research: An update and assessment of its role in education finance reform. *Educational Evaluation and Policy Analysis, 14*(4), 307–332.

Monk, D.H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*(2), 125–145.

Muthén, B.O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338–354.

Muthén, B.O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*(3), 399–420.

National Center for Education Statistics. (1996). *High school seniors’ instructional experiences in science and mathematics.* Washington, DC: U.S Government Printing Office.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

O'Reilly, P.E., Zelenak, C.A., Rogers, A.M., & Kline, D.L. (1996). *1994 trial state assessment program in reading secondary-use data files user guide.* Washington, DC: U.S. Department of Education.

Phelan, P. (Ed.). (1989). *Talking to learn.* Urban, IL: National Council of Teachers of English.

Steinberg, L.D. (1996). *Beyond the classroom: Why school reform has failed and what parents need to do.* New York: Simon and Schuster.

Strauss, R.P., & Sawyer, E.A. (1986). Some new evidence on student and teacher competencies. *Economics of Education Review, 5*(1), 41–48.

Traub, J. (2000, January 16). What no school can do. *New York Times Magazine,* p. 52.

Waller, W. (1932). *The sociology of teaching*. New York: Russell & Russell.

Wenglinsky, H.H. (1996). *Modeling the relationship between school district spending and academic achievement: A multivariate analysis of the 1992 National Assessment of Educational Progress and the Common Core of Data*. Princeton, NJ: Educational Testing Service.

Wenglinsky, H. (1997). How money matters: The effect of school district spending on academic achievement. *Sociology of Education, 70*(3), 221–237.

Notes

[1] As is common in the literature, this paper uses the terms “effect” and “school effect” to connote statistically significant associations between variables. These associations need not be causal in nature.

[2] For a discussion of the methodological issues associated with production function research, see Forture and O’Neil (1994), Monk (1992), and Wenglinsky (1997).

[3] For mathematics, the classroom practices are similar to those endorsed by the National Council of Teachers of Mathematics (1989).

[4] It should be noted that some school effects research addresses the problem of the insensitivity of regression analysis to multilevel data through the use of Hierarchical Linear Modeling (HLM). There are trade-offs to using HLM as opposed to MSEM. HLM has the advantage of being able to treat as a dependent variable not only a student outcome, but also the relationship between that outcome and student background characteristics; for its part, MSEM makes it possible to explicitly model measurement error and more fully test relationships among independent variables. While this study uses MSEM, it should be supplemented with an HLM.

[5] In aggregating teacher characteristics to the school level, the values of all teachers in that school for whom there were data were averaged. It was not possible to create a separate teacher level of analysis because there were generally only one or two teachers surveyed from each school, and thus not a sufficient number of degrees of freedom for a third level.

[6] For a fuller discussion of this approach as applied to the 1992 mathematics assessment for eighth-graders, see Wenglinsky (1996).

[7] More generally, the pooled variance can be expressed as:

$$V = U^* + (1 + M^{-1})B$$

Where V is the pooled variance,

U^* is the average sampling variance,

M is the number of plausible values, and

B is the variance among the M plausible values.

[8] One misleadingly compelling alternative to this approach is to treat the five plausible values as multiple indicators of a test-score construct. However, this approach violates the assumption in

structural equation models of independence of errors, and has been shown to distort estimates of residual variances and certain statistics, such as the *R*-squared (Mislevy 1993).

[9] Because NAEP is a sample of students and schools, but not of teachers, descriptive statistics apply to the students rather than the teachers (e.g., 45% of students have teachers who received professional development in higher-order thinking skills, not 45% of teachers received professional development in higher-order thinking skills).

[10] Loadings used here are taken from the classroom practices model (Table 5). For constructs that were also included in other models, the loadings proved nearly identical across models. The output for the two other school-level factor models is not presented here but is available upon request.

[11] Total effects can be calculated in one of two ways. The first is to estimate a single model that includes all relevant variables, both exogenous and endogenous, and to sum each of the direct and indirect effects for each variable. This option can be problematic, however, in that the size of the total effect may be an artifact of the number of paths the model permits. The more paths that are fixed at zero, for a given variable, the lower the total effect. The second option is to estimate successive models, in which only the direct effects of the variables are used. Thus, in the current case, the first model is made entirely of exogenous variables. Their direct effects on achievement are equal to their total effects. The second model adds a set of endogenous variables. They are related to achievement only in a direct manner, however, and hence can be treated as total effects. A final set of endogenous variables is added in the third model. These, too, are only directly related to achievement and hence can be treated as estimates of total effects. The presentation of total effects in this study is thus based upon the direct effects of teacher inputs in the first model, of professional development in the second model, and of classroom practices in the third model.

[12] Mayer (1999) finds that while composite measures of classroom practices drawn from teacher questionnaires are highly reliable and valid, individual measures are problematic.