Commentary

Measuring and Reporting School and District Effectiveness[1]

James L. Phelps

Introduction

The federal *No Child Left Behind Act (NCLB)*, the 2001 reauthorization of Title I, requires states to assess students in grades three through eight in reading and mathematics, and students in three grades in science. *NCLB* further requires states to evaluate schools on the basis of their aggregate performance on these examinations. Specifically, schools are required to show "adequate yearly progress" (AYP) for each student subgroup represented in the school in each subject tested and, ultimately, bring *every* student to proficiency by 2013-2014.

Under *NCLB*, schools and districts failing to make AYP for two or more consecutive years are required to undergo a set of reforms and sanctions. These include the offering of transfer within the district to children whose parents desire a school change, the provision of supplementary educational services, the replacement of school staff, and the conversion of the school to charter status. Additional district sanctions include the withholding of Title I funds, replacement of district staff, and district reorganization. In response to these mandates, each of the 50 states has implemented an accountability plan that specifies curriculum content standards by grade level and achievement levels on tests to measure attainment of those standards. According to the U.S. Department of Education (2005, p. 7), no two state accountability plans are identical. As the U.S. Department of Education notes, "...within each state context–considering diversity of student populations, number of schools, size of schools, and other factors–states must strike a fair balance when making school accountability decisions. States must design accountability systems that are both valid (accurately identifying schools not reaching their academic goals for all students) and reliable (with accountability judgments based on sound data)" (U.S. Department of Education, 2005, p. 8)

In response to this federal mandate and the public's call for incentives to improve the quality of teaching and learning in our public schools, states have adopted outcome goals for schools and students, implemented student testing programs, and used the test results to gauge school effectiveness. The stakes are high. Not only do states attach financial rewards and public recognition to superior school performance, but school and district enrollments and corresponding revenue are also contingent on school test scores; school choice programs often allow high performing schools to attract residents of neighboring districts.

The value of these school accountability programs as both indicators of school performance and incentives for school improvement depends crucially on several characteristics. The accountability program must be: (1) understandable by policymakers, practitioners, and the public; (2) statistically valid and reliable; and (3) operational by departments of education. Understanding of and confidence in an accountability system are essential. Policymakers, practitioners, and the public must have a general understanding of the key decision-making factors, how the system works, and where the respective schools stand on these key factors. While there will be large volumes of data available for analysis, these data must be reduced to the core—the key elements—while maintaining accuracy. In essence, the system cannot be so complicated that it cannot be easily implemented and reported.

State efforts have varied considerably in rigor and sophistication, ranging from simple school performance measures such as average student test scores or percentage of students surpassing a specified proficiency level to "change scores" and "adjusted performance measures" (APMs) that explicitly account for the often wide disparities in resources and student characteristics across schools. APMs are derived from school-level regression equations in which school performance measures, generally test scores, are regressed over a set of independent variables representing school and student characteristics beyond the school's control. The APM is the residual obtained from the regression, or the difference between each school's actual and estimated performance level. Clearly, the APM approach is preferred to simple performance measures once agreement is reached on a standard set of adjustment parameters.[2] The calculation of APMs is also quite feasible for states refining their school accountability plans, requiring routinely collected school-level administrative data.[3]

In contrast, scant attention has been given to the task of identifying effective school districts, despite the considerable emphasis placed on district as well as school performance in *NCLB*.[4] This joint focus on schools and districts raises the question: How much do district policies, leadership, and support services influence the quality of teaching and learning in public schools? These district attributes generally go unobserved in empirical studies of school performance and effectiveness, but their influence could be substantial.

The strategy for this project was selected after a review of other more complicated alternatives: Data envelopment analysis; mathematical programming; and hierarchical linear modeling (HLM). The strategy also evolved from an earlier effort. This model is based on what is commonly called "fixed effect estimation" in econometrics for which there are several alternatives (Schwartz & Zabel, 2005). This model was developed as a hybrid to meet the criteria identified above.

The purpose of this article is to illustrate how a valid and reliable state accountability system could be developed that identifies effective schools and school districts in a comprehensive, understandable, and practical way. Section two presents an overview of the strategy used in the analysis. The third discusses the use of education production functions and to assess school effectiveness. Section four presents a model of education production. The data are described in the fifth section while the analysis process is described in section six, and the empirical results are presented in section seven. The

James L. Phelps is former Education Assistant to the Governor of Michigan and Deputy Superintendent in the Michigan Department of Education.

results from the presented method are contrasted with results from a "change score" approach in the eighth section with conclusions and implications for school accountability policy are presented in the final section.

Strategy

The strategy for building the accountability system is largely based on the several definitions of the word "Par." The components of the accountability system are identified and converted to a common and understandable "currency" to form an educational profile. The profile includes the various components of achievement, school resources, and student/community characteristics. A unique target achievement score is then determined for each school based upon the school resources and student/community characteristics contained in the profile. The target achievement score is compared with the actual score over time to determine what schools consistently under- or over-perform their individual "Par." Those schools consistently performing better than expected—better than their unique par—are considered effective. The degree to which schools exceed or fall short of "Par" becomes an index of effectiveness. All the key information regarding the accountability system is contained in an educational profile; it is the centerpiece for reporting to policymakers, practitioners, and the public.

Accountability

Once a potential measure of effectiveness is constructed, it is critical to determine if the measure is valid. In this case, the question is whether the effectiveness measure identifies individual schools randomly or systematically based on their performance. Schools should be held accountable only for those actions under their control and not for random occurrences. Distinguishing between random and systematic occurrences is accomplished by evaluating the performance of individual schools over time; one observation is insufficient. The difference between random and systematic may be best illustrated by a golfing analogy. Because the objective is to putt the ball into the hole, an individual who consistently misses the target to the same side is performing systematically and it is reasonable to expect a corrective action. On the other hand, someone who consistently putts the ball into the cup (hits the target) and only sometimes misses just a little to either side is performing randomly with no specific corrective action indicated (except more practice).

As a consequence, if the effectiveness measure is judged to be a random occurrence, it is an inappropriate accountability measure because it is uncontrollable by school officials. If, however, the effectiveness measure is determined to be systematic, it is a valid and reliable accountability measure because it indicates that "effectiveness" is indeed under the control of the school organization (and corrective action is warranted). In sum, the occurrence is considered random when there is an equal likelihood of performing above or below the expected level. The occurrence is considered systematic when the performance is consistently either above or below the expectation. The systematic/random likelihood is estimated through regression analysis comparing school performance over time.

Conceptual Categories

The data variables for the accountability system are selected purposefully: because they fit into the conceptual categories of student/community characteristics (SES or socioeconomic status), staff quantity, staff qualifications, and instructional materials. States regularly collect data in the categories of staffing roles, staff qualifications, instructional material expenditures, and student characteristics because these categories are commonly acknowledged as being related to student achievement. (The non-instructional and facilities categories are not included because they are thought not to make a substantial contribution to achievement and they would add undue complexity.) In other words, the individual variables for possible use in the analysis were not selected because of their unique conceptual value; they were selected only because of their membership in one of the compelling categories.

The justification for grouping individual variables into conceptual categories, what is hereafter called "factors," is based on factor theory, a fundamental principle of regression. Briefly, the statistical variance of conceptually and statistically related variables is divided into three types: (1) the common variance shared by all variables (sometimes called the g-factor); (2) the unique variance of each individual variable; and (3) the error variance. When measuring and reporting individual variables, it is not clear how much of the variance is "common" and how much is "unique" because some of the variance is shared by other variables. Instead of trying to distinguish among the common and unique variances for each individual variable, a better alternative is to measure and report the total variance—common and unique—for the entire factor. Operationally, the total contribution of the regression equation is reported as being the factor rather than the contribution of the individual variables.

The individual variables within each of the previously identified school factors are substantially correlated because they share common variance (g-factor). This is supported by the general observations: (1) all instructional staff roles combine to produce a comprehensive instructional environment; (2) teacher qualifications are an integrated combination of traits; and (3) instructional materials work as an amalgamation. All these are reasonable illustrations of gestalt, a set of variables working together conceptually, operationally, and statistically to produce a larger product.

SES is commonly reported in research papers as a single factor even though it is most certainly comprised of several variables. Individual variables are combined via regression to represent the concept of SES as a proxy. Similarly, there is no single data variable representing the other factors: Staff quantity; staff qualifications; and instructional materials. Individual variables must be combined to form proxies for the factors. The variables identified for inclusion in each proxy and their weightings are based first on their membership within the conceptual category, and then on their relationship with achievement and their inter-correlations as a part of the regression process.

This strategy evolved based on the shortcomings of a previous analytical effort, which utilized individual variables rather than related variables combined into factors. In the previous analysis, different combinations of variables accounted for the relationship with the several achievement outcomes. This was due to the high correlation among the explanatory variables causing the order of entry into the regression equations to change frequently. The assumption that an ever-changing set of variables with an ever-changing set of weights explains student achievement was difficult to sustain. It is more reasonable to assume that consistent variables with consistent weights are related to achievement. Therefore, it was prudent to use a common variable set with common weightings to form factors across the various achievement equations. By inspecting the regression results

for each factor, those variables consistently making a contribution are easily identified because the weightings were similar. These are the reasons why this analysis is conducted in terms of factors rather than individual variables.

The goal is to develop a single number for each factor that is a "good" predictor of achievement. The first step is to build a series of regression models predicting the various grade/subject achievements for each of the factors across the several years identifying the variables with consistent predicting powers. Using only these variables, the next step is to select the weightings producing the "good" predictor factor formulae. This is accomplished by combining (averaging) the respective variable weightings. The weightings can be combined for only years, resulting in a unique set of factor formulae for each achievement variable, or combined for years and grade/subject achievements for a common factor formula across the multiple achievement measures. The common factor set alternative was selected in order to reduce the number of comparisons required to present the results. In addition, it avoids the question of why individual schools would rank differently on each of the factors for each of the grade/subject achievement tests. The final step is to insert the data for each of the observations into each factor formula to obtain the factor scores. Now, a few key numbers "explain" achievement, rather than too many numbers to contemplate.

Importantly, the actual school values of the factors are different for each year because the data change every year, even though the definitions remain constant. Most importantly, what little explanatory variance is "lost" by combining the variable weightings is later "recouped" when the factors are entered into the equations predicting the achievement levels for each grade and subject. In essence, the explanatory variance is moved from the individual variables to the factors. With this transformation, the results are easily understood as the product of four achievement measures against four common factors (16 comparisons), rather than sixteen factors (different each year) against the four achievement measures (64 comparisons), or a multitude (23) of individual variables and the achievement variables (92 comparisons).

Before being used in the equations, the factors scores are first transformed into standard scores and then into percentiles (area under the normal curve), standard statistical procedures. (Standard regression coefficients are produced when the variables are in standard scores.) In addition to normalizing, the transformation adjusts for the undue influence outliers may have on the results. This process creates a consistent, common, and easily understood measurement scale for every factor—the common "currency" of percentiles. All the elements in the educational profile are then directly comparable.

Testing the Transformation

The amount of explanatory variance was calculated for the transformed (factors) and non-transformed (variable sets) forms of the equations, and the results were virtually identical (.02 or less in the amount of explained variance). Therefore, the transformation process neither materially diminished nor augmented the statistical results. Thus, the factors are available as a comprehensive and comprehensible profile of school performance and resources.

Analysis Strategy

The factors were entered into the regression equations. Using the factors, regressions yielded the predicted achievement levels and residuals. Residuals (the difference between the predicted and actual levels) are normally reported in terms of standard scores so the transformation to percentiles is straightforward. Therefore, all of the factors are in a standard "currency" or index.

The next part of the strategy is to analyze the residual. By definition, the residual is normally distributed around the standard score of zero; the chance of being above or below the mean is virtually equal. However, the residual is actually comprised of random and systematic error. There is a critical difference between random and systematic error: random error is random over time; systematic error is not (Taylor, 1982, p. 81). Analyzing the residual for each observation over time identifies the systematic-error portion of the residual. In this context, the error analysis addresses the question, what schools consistently—or "on average"—perform above or below the expected level? Regressing the time-averaged residual against the dependent variable identifies the systematic portion of the residual. If the amount of variance explained by the averaged residual is zero, then there is no systematic occurrence. If the explained variance is not zero, then there is systematic occurrence. The random portion can be measured because the sum of the two types of error equals the residual.

In essence, this method is based on the identical algebra commonly known in econometrics as "fixed effect estimation," with the systematic portion of the residual being the "fixed" or "school effect." This portion of the residual is called "fixed" because of the assumption that it changes little, if at all, over a reasonably short period of time and can be best estimated by the average.

Econometric Models

There are specialized computer programs for conducting "fixed effect" analysis that are effective under certain conditions: (1) There are a small number of variables under consideration; and (2) the primary interest is in the statistical inference of the variables; and (3) the audience has a sophisticated understanding of econometrics. These conditions do not appear to be present in the situation at hand. So rather than using a "black box" computer model, the product from each step of the analytical process is presented in order to provide understanding and confidence to those who are in judgment of the final product—an index of school effectiveness. In other words, this method combines a myriad of variables into a comprehensible profile of school performance and calculates the components of "fixed effects estimation" for those individuals who are not knowledgeable in the field of econometrics. It culminates with an index of school effectiveness within the educational profile.

Assessing School Efficiency

One approach to developing school effectiveness measures relies upon the concept of production efficiency and techniques for measuring such efficiency. This approach utilizes the economist's notion of a production function.[5] Production models have three parts: The outcomes sought; the necessary ingredients or inputs; and the process transforming the inputs into outcomes. These three parts are linked together by a mathematical function. This production function reveals the maximum amount of outcome possible for various combinations of inputs. If the levels of the inputs and the function are known, the maximum level of outcome (i.e., production) can be determined. Anything short of maximum attainable output indicates technical inefficiency.

A second dimension to production efficiency involves input costs. Assuming an organization makes the best possible use of a set of inputs—that is, it is technically efficient—the least-costly input combination is required to achieve allocation efficiency. Put another way, production efficiency requires both technical and allocation efficiency.

A third dimension of production efficiency involves the process portion of the production function. Assuming that technical and allocation efficiency have been achieved, the process must also be efficient before the maximum attainable outcome is achieved. This aspect is discussed in more detail later. Together these three dimensions combine to yield production efficiency. For a more detailed discussion of the educational production function, see Monk (1990). Notwithstanding some difficulties, various notions of the production function receive political support across the states and serves as the basis of many school accountability systems.

An accurate estimate of the effectiveness or "quality" of a school (the school's contribution to student learning) must account for the relative contributions of SES and school resources to student learning. Put another way, accountability systems should not confound school quality with other fundamental determinants of student performance, particularly when assessments of school quality trigger school rewards and sanctions.

The production function approach estimates the marginal educational contributions of identified educational inputs, both "controllable" and "uncontrollable," and identifies those controllable inputs with positive marginal weightings. These estimated weightings can then be compared with corresponding input costs to improve allocation efficiency. The production function approach can also be used to identify school districts and schools that consistently produce levels of student achievement that exceed (or fall short of) levels predicted by the identified inputs. These consistently higher or lower than predicted performance levels can be attributed to the process component of the production function for which data are usually unavailable.

The process component is difficult to measure and thus is generally excluded in educational production function studies. Staff and organizational behavior are frequent process topics. Murnane and Phillips (1981), in a study of elementary schools, included a set of teacher behavior variables in a model of vocabulary test performance. The variables included the percentage of time the teacher used subgroups, demonstrations, and individualized work, and whether the teacher felt responsible for explaining the subject matter. These authors found that the process behaviors explained a larger proportion of test score variance than teacher qualification characteristics. School climate, another process variable, may also enhance the quality of teaching and learning (Mortimore, et. al., 1988).

Leibenstein's (1966) seminal article on X-efficiency in businesses contends that incentives and other generally unmeasured organizational attributes of the firm make a greater contribution to process efficiency than the marginal reallocation of inputs. Building on the same idea, Levin (1997) suggested that unmeasured and often unobserved school practices and organizational characteristics—the process component of the production function—can be very important to school performance. Levin did not provide estimates of the magnitude of X-efficiency. Actually, there are few empirical studies regarding X-efficiency in schools. While there are some general ideas as to why some schools consistently produce higher or lower than predicted performance, the specific behaviors and organizational characteristics are largely unknown.

A Model of Education Production

In this section, a production function model is used as an approach to estimate the magnitude of the unobserved school characteristics influencing student performance—the X-efficiency factor. The basic notion of the model is:

$$\text{Output} = \text{Input} + \text{Process}$$

Hanushek proposed a framework for an educational production function that distinguishes among family background, peer, and school inputs (Hanushek, 1979). A simplified version of this production function is:

$$A = f\,(B,P,S)$$

where A represents outcomes; B represents family background inputs; P represents peer inputs; S represents school resources; and $f\,(\ )$ is the function, or production process transforming the inputs into outcomes. This framework is modified slightly, combining the family and peer inputs into a single SES element and includes the process X-efficiency factor. The theoretical school-level model of education production becomes:

$$A = f\,(SES, S, X)$$

When the different aspect of school resources are identified and the process portion or X-efficiency is included, the expanded production function becomes:

$$A = f\,(SES, SQN, SQL, IM, X, E)$$

where A is the school achievement level; SES represents the student/community characteristics; SQN represents the staff quantity; SQL represents the staff qualifications; IM represents the funding for instructional materials; X represents the unobserved X-efficiency behavior and policy attributes; and E represents the random error. The SES and school resource factors are the inputs, for which there are data, and the unobserved X-efficiency factors along with the error are in the residual (i.e., the difference between actual and predicted performance levels for each school). Additionally, prior school resources and SES could have an influence on later achievement levels and could be considered a part of the production function. These prior values were incorporated into the production function analysis, discussed later.

What is not measured directly is concealed in the residual term along with the measurement error. Of particular interest is the portion of the residual term attributable to a missing variable; that is, X-efficiency. Accordingly, the model is estimated and the residuals divided into random and X-efficiency components. In essence, this analysis measures indirectly the "process" portion of the production function from estimates of the outcomes and inputs based on the following logic:

If Outcome = Input + Process, then Process = Outcome – Input.

Data

A panel of school-level data was obtained from the Minnesota Department of Children, Families and Learning for elementary schools for the years 1998 through 2001. All schools reporting to the state

were included in the study. Reporting of school-level data was optional in 1998, and 506 schools participated that year. Participation rose to 671 schools in 1999, 690 in 2000, and 694 in 2001, including all elementary schools in the state. Data for all variables were reported by participating schools, except for "teachers' average years of teaching experience" for 1998. For that year, each school's 1999 figure was used. A complete panel of data was available for 476 schools. Achievement data consisted of building-averaged scores on statewide assessments for reading and mathematics in grades three and five for each of the four years.[6] Definitions for the set of school-level variables are given in Appendix A.

Analysis Process

The analysis process began with the construction of a set of indices based on the factors in the production function. Indices for staff qualifications, staff quantity, and instructional supplies and materials (non-personnel instructional expenditures) were constructed from sets of component variables.The purpose of the regression-based method was to maximize the proportion of variance in student achievement explained by the variance in the respective indices. Importantly, by maximizing the explanatory variance in the factors, the residual, and therefore the school effect is minimized to avoid any over estimation. These school resource indices and their component variables are summarized in Table 1.

Specifically, the achievement measures were regressed against the component variables of each index. The estimated coefficients for each variable were then averaged over the four years, and this average was used as the weighting for the variable in the construction of the index. For the same reason the "fixed" or school effect is assumed to be constant, the weightings are assumes constant and their best estimate of the "true" value is the mean over the time periods (Wooldridge, 2000, p. 441-2). (Analytical research cannot be conducted without the assumption that the same laws exist in space/time, also a basic principle in physics. The relationships between inputs and outputs were assumed to be the same wherever the school is located, the space component. Likewise, the relationships were assumed to be the same regardless of when the measurements are made, the time component.)

The amount of explanatory variance from each index was calculated and compared with the variance using the component variable sets in order to verify that the indexing process did not substantially change the results. The comparisons were made for each index, for each achievement measure, and for each year, for a total of 16 comparisons per index. The actual variance values for the respective indices were similar and the average differences between the two methods (indices and component variable sets) were small: .024 for staff quantity; .018 for staff qualifications; and -.013 for instructional materials. The average weighting method for SES, however, produced a larger difference, .062. Because this level was considered too high, an alternative method was tested; instead of averaging the regression coefficients, the individual variables were weighted based on the inverse of the standard deviations.[7] This method produced a result more similar to the average variable set method, the difference being .014.

Finally, each school's index and achievement levels were converted into a percentile ranking. This scaling did not change the statistical character, but did reduce the undue influence of outliers (Wooldridge, 2000). At this point, there was a profile of four school achievement measures and four resources measures in a common scale or "currency," meeting the two previously identified criteria of an accountability model: The components of the accountability system are understandable by policymakers, practitioners, and the public; and they are statistically valid and reliable. Without the factors and indexing, there would still be four achievement measures, but twenty-three explanatory variables all in different metrics are hardly "user friendly."

Using the achievement and school resource indices of the profile, the model was estimated by ordinary least squares (OLS).[8] Separate

Table 1
Indices of School Resources

Index	Component Variables
Staff Qualifications (Teachers)	(1) Average length of teaching experience; (2) average salary; (3) average age; (4) percentage with a Master's degree; (5) percentage of new teachers
Staff Quantity (Instructional Staff Only)	(1) Administrative staff (licensed and unlicensed); (2) licensed staff (teachers); (3) licensed support staff; (4) non-licensed instructional staff (teacher aides), all per 1,000 students
Non-Personnel Instructional Expenditures (Instructional Materials)	(1) Supplies and materials; (2) capital outlay and debt; (3) other instructional non-personnel expenditures
Student/Community Characteristics (SES)	(1) Percentage of children in the school who are eligible for free or reduced-price lunch; (2) percentage of children who are minority; (3) percentage of children who are in special education; (4) reported disciplinary incidents as a percent of building enrollment; and (5) intra-district mobility rate. Four other variables were excluded because they did not add to the explanatory power.

regressions were run for each of the outcome measures (READ3, READ5, MATH3, and MATH5) for each of the four years. Because the focus was on the residuals and not the estimated coefficients of the indices, the complete regression results are not reported. Moreover, there is no attempt to make statistical inferences regarding the indices. At this point, the school profile is further developed; a predicted achievement level, in percentiles, is added in order for it to be compared with the actual achievement level. The other byproduct of the regression is the residual, the dwelling of the school effect—the final piece of the profile puzzle.

Analysis of Residuals

The object of the residual analysis is to partition the explanation of the achievement levels across the factors of the production function. This is accomplished by first partitioning the amount of variance (the R^2 or coefficient of determination) explained by the SES and school resource factors from the residual, and then separating the random error from the systematic error within the residual. The systematic error portion of the residual is considered to be the school effect. An upper bound for the magnitude of the residuals is: 1 minus the coefficient of determination ($1-R^2$). The R^2 for each outcome measure, averaged over the four years, was: MATH3 = .532; MATH5 = .635; READ3 = .712; and READ5 = .706 with an average of .646. Therefore, the random and systematic error must share the difference between this number and 1, or .354.

To obtain an estimate of the magnitude of the systematic error, the residuals were examined to identify schools and districts that consistently over- or under-performed compared with predicted outcome levels. A school that consistently exceeded its target performance, as predicted by its students' characteristics (SES) and resource levels, was presumed to benefit from unobserved school attributes, or X-efficiency. Specifically, the averaged residual represents the systematic error and is the estimate of school X-efficiency. School residuals were averaged for each outcome (i.e., grade level and subject) over the four years. In essence, the averaged residual became a new variable representing the effectiveness of each school. The effectiveness variable was entered into the regression equations to determine if it was associated with the achievement variable, controlled for the other factors. The magnitude of the association was measured.

If the effectiveness variable (the four-year averaged residual for each school) represented only random error, the regression coefficient would be zero, and it would account for no additional variance (R^2). In other words, schools had the same chance of being above the target level as below. If this were the case, the conclusion must be that the effectiveness variable has no statistical validity. If, on the other hand, the coefficient was greater than zero, the magnitude of statistical validity of the effectiveness variable is measured by the percent of variance explained (R^2). In this case, the conclusion must be that there is some underlying reason why schools consistently either under-achieve or over-achieve their predicted targets. The statistical results are substantial. By including the effectiveness variable in the equations, the percent of variance explained (R^2) increased for all subject/grade combinations, with an average increase from .646 to .928 and a change of .282 out of a maximum possible .354 (see Table 2).

The effectiveness variable has the same distinctive properties as the residual. It has no correlation with the other variables in the equation; i.e., it is not associated with SES or any of the school resources variables. If, for example, a variable representing staff qualifications is incorporated into the regression equation, it must be substantially independent of the other qualifications variables (experience and training) included in the staff qualifications index in order to have an impact on the results. No candidates for variables associated with the factors with statistical independence come immediately to mind. Therefore, additional variables and better data will improve the predictions, but it is highly unlikely that they would account for a major portion of the amount of variance that can be explained by the effectiveness variable. Put simply, a better specification of the model may reduce the influence of the effectiveness variable but would not eliminate it.

Of equal interest is the relationship between the effectiveness variable and achievement. For any single time period, there is no correlation between the residual and achievement. Only when the residual

Table 2
Decomposition of Residuals into School and District Fixed Effects

Outcome	Coefficient of Determination (R^2)				Error ($1-R^2$)
	Without Residual	With Residual	Difference		
			District	School	
MATH3	0.532	0.913	0.212	0.168	0.087
MATH5	0.635	0.932	0.155	0.142	0.068
READ3	0.712	0.935	0.128	0.095	0.065
READ5	0.706	0.932	0.107	0.119	0.068
Mean	0.646	0.928	0.151	0.131	0.072

is averaged over time does the relationship emerge. It is the averaging process that separates the random and systematic error and provides for the estimate of the effectiveness variable. A longer time period yields a more accurate measure.

The next step was to divide the total effectiveness variable into school and district components to obtain estimates of a school effect and a district effect. To do this, the school effectiveness measures were averaged within each district. The district mean was interpreted as the upper bound for effectiveness attributable to the district—the district effect. The differences between the district average and each school effectiveness measure were considered the school effects. As a result, there were two effectiveness variables, one for the school and one for the district.

To estimate the magnitude of these school and district effects on student achievement, the regressions were re-run for each achievement measure with the school and district effectiveness variables, the SES factor, and the school resource factors. The contributions these effectiveness variables made to the coefficient of determination (R^2) are presented in Table 2.

At this point, another consideration was also addressed. Prior school resources may have an impact on the results because they could have a longer-term influence. This was tested. Regressions were run inserting prior SES, staff qualifications, staff quantity, and instructional materials factors into the equations as lag variables. There was a slight increase in the total R^2 and a small decrease of the R^2 for the X-efficiency effect. The increase in the total R^2 and the discounting for the X-efficiency amounted to about .010 for Math 3, .013 for Math 5, .008 for Read 3, and .005 for Read 5, for an average of about .009. While this is important to note, it increases the precision of the X-efficiency effect only slightly but has little effect on the substantial magnitude.

Discussion

As the results in Table 2 indicate, the district effect accounted for between 10% of the variance in measured achievement for fifth grade reading and 21% for third grade mathematics, averaging about 18% for mathematics and 12% for reading. The estimated school effect ranged from 10 percent for third grade reading to 17% for third grade mathematics, averaging about 16% for mathematics and 11% for reading.

The finding of greater school and district effects on math achievement than on reading achievement is intuitive. Parents may spend considerable time reading with their young children, while mathematics instruction is left largely to the school system.

These school and district effects are substantial. They reflect unobserved qualities of school administrators, faculty, support staff, and the educational climate they create, along with other unobserved variables. More importantly, the personal and professional qualities of these educators interact in ways that produce effective curricula, pedagogy, and instructional programs. The translation of these qualities into effective educational practice is important, but not illuminated by this quantitative analysis. The only way to identify these school effectiveness characteristics is to conduct case studies based on this type of analysis.

On the other hand, this analysis identifies the sources of much of the variation in elementary school student achievement. The R^2 changes associated with school and district effects can be added to the R^2 changes associated with SES and school resources to obtain an estimate of the total explained variance in student achievement (R^2_{total}). The unexplained variance is estimated as ($1-R^2_{total}$) and is attributable to noise in the data and random error. On average, the proportion of the variance in student achievement that remains unexplained is a mere 7%, a remarkably low figure when compared to other education production function studies.

One may expect that these unobserved school and district effects would be roughly consistent across grades and subjects; that is, a good elementary school is good in all grades and subjects. To further examine the consistency of these effects across subjects and grades, correlation coefficients were calculated across subjects and grades. These correlation coefficients are presented in Table 3.

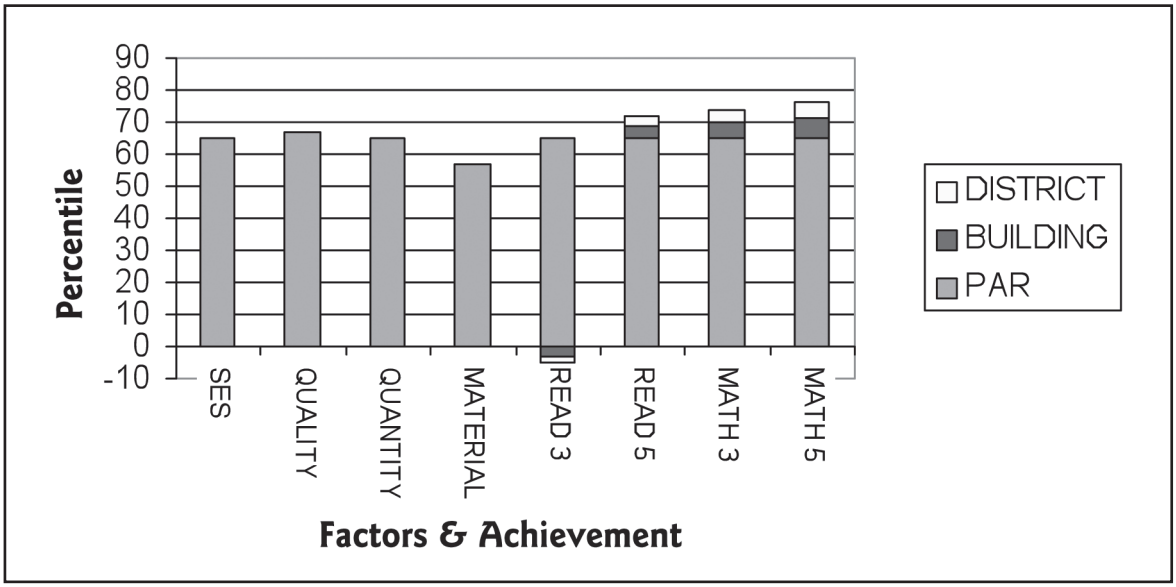
The correlations are relatively high, confirming that the fixed effects, or levels of X-efficiency, within a school tend to be consistent across subjects and grades over the four-year period examined. The effects of such unobserved variables as climate, communications, leadership, and performance incentives appear to be reflected throughout the school and not restricted to particular grades and subjects.[9]

More generally, this consistent pattern of effectiveness across district and school grades and subjects reveals a degree of stability in school and district influences on teaching and learning in the classroom. Not surprisingly, effective schools are found in effective

Table 3
Consistency of School Fixed Effects:
Pearson Correlations Between Estimates Across Grades and Subjects

	MATH3	MATH5	READ3	READ5
MATH3	1.000000			
MATH5	0.725443	1.000000		
READ3	0.656566	0.564673	1.000000	
READ5	0.677272	0.902083	0.614691	1.000000

Figure 1
Example of a School Effectiveness Profile



districts. The pattern reflects the effects of activities, policies, incentives, instructional practices, climate, and other inputs that are consistently present in the schools and districts but are not captured by the SES or school resource measures.

These results support the previously stated criterion of an accountability system: The method is valid (accurately identifying effective schools and school districts) and reliable (based on sound data and analysis). With the inclusion of the school and district effectiveness measures, the school profile is complete. In one easily understood profile, there is the necessary overview information of the accountability system, including school and student/community resources; the predicted achievement levels; individualized par; and the school and district measures of effectiveness. Schools exceeding par are effective and positive, while schools below par are negative. (See Figure 1 for a simplified illustration of a school effectiveness profile.) The yearly production of this profile would provide policymakers, practitioners, and the public with an understandable report of school status and progress in a statistically valid and reliable form. The production of the profile, as outlined, would seem to be within the grasp of state departments of education, the final criterion of an accountability system.

Comparison with a Difference Model of Effectiveness

Many state accountability systems measure school performance by changes in achievement from one year to the next (Figlio, 2005.) Despite some demonstrated shortcomings, this method, sometimes referred to as "difference scores," is attractive to states because it is relatively easy to administer and explain to the public.[10] The "difference scores" methodology can be interpreted as measuring the production function during two time periods. Assuming the previously presented production function, a straightforward algebraic analysis demonstrates that difference scores is actually attributable to the changes in SES, staff qualifications and quantity, instructional materials, and X-efficiency, only a small part of which is under direct school control (Wooldridge, p. 422). This interpretation makes the justification of difference scores difficult to sustain.

Difference scores for Minnesota's elementary schools were calculated and compared with the X-efficiency findings. In seeking to identify the preferred measure of school efficiency, the following criteria were applied: First, the efficiency measure should be neutral with respect to factors over which schools have little or no control (e.g., student SES and school resource levels); second, each school should have the same chance of improving (e.g., a school's likelihood of improving in any given year should not be conditioned upon prior year performance).

Neutrality can be measured by the simple correlations between the efficiency measure and the uncontrollable SES and resource indices. These correlations are virtually zero for the X-efficiency measures by construction. The correlations of difference scores with the SES and resource indices are also near zero, indicating that difference scores satisfy the neutrality criterion.[11]

To assess the independence of difference scores from prior year scores, the Minnesota elementary schools were divided into deciles (ranked by prior year achievement level) and their difference scores were calculated. The findings are presented in Figure 2. There was an inverse relationship between school difference scores and prior year performance level. This result is intuitive, reflecting both an increasing marginal cost of improvement and a regression to the mean for these schools' academic performance. To complete the analysis, correlations between the schools' difference scores and the schools' X-efficiency scores were calculated (both averaged over the four-year period). These correlations were: READ3 = .45; READ5 = .56; MATH3 = .46; and MATH5 = .52. As indicated, these alternative measures of school effectiveness were not closely comparable. The X-efficiency measure was clearly superior according to the criteria discussed above.

Figure 2
Average Value Added by Decile Group

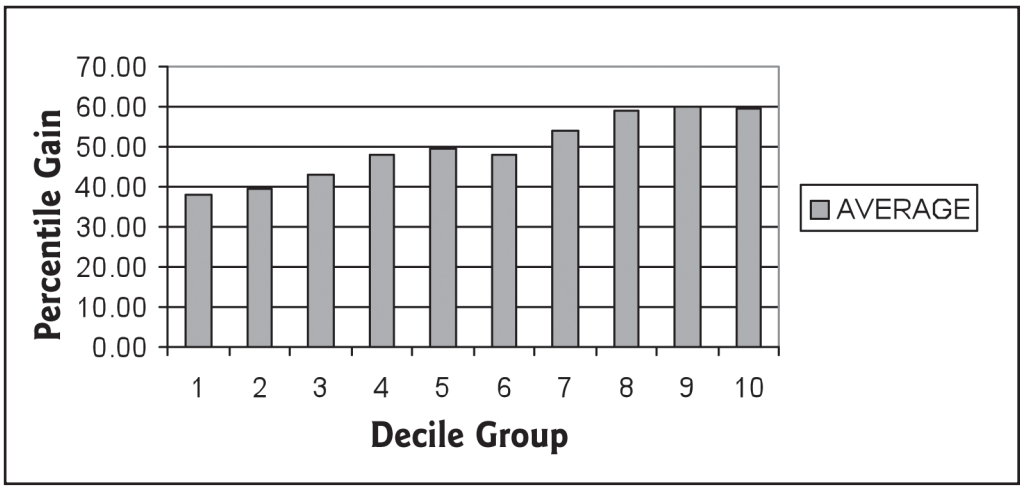
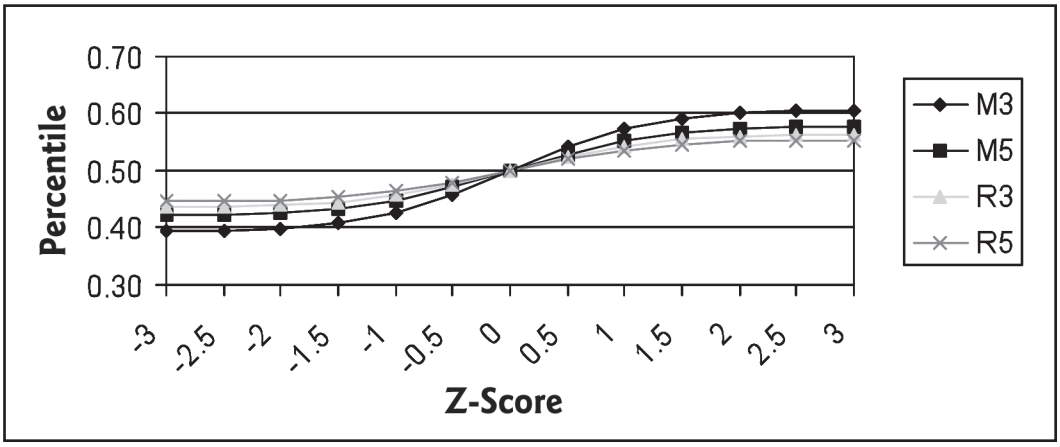


Figure 3
District X-Efficiency



Note: M3 = MATH3; M5 = MATH5; R3 = READ3; and R5 = READ5.

Discussion and Conclusions

The key to an accountability system is to separate those elements beyond the control of schools (SES and resources) and focus on the elements under their control.

In keeping with a vast research literature on educational productivity, this analysis confirmed that the socioeconomic characteristics of students remain the most influential factor in predicting achievement outcomes. SES exerted a large influence on academic achievement, about 55% of the variance.

Estimating the impact of school resources on student achievement is problematic. First is the simultaneity problem; low-performing schools are given additional, compensatory resources. Second, school resources are correlated with school SES in a U-shaped relationship, where resources are highest in extremely low and high SES schools. Correlations between school SES and school resource measures are: staff quantity, .393; staff qualifications, .320; and non-staff instructional financial resources, .427. Nevertheless, an estimate of the school resources is about 11% of the explanatory variance. This amount includes about 1% due to adding prior school resources as a lag variable into the analysis (the 1% is discounted from the school effect). No attempt was made to distinguish among the relative contributions of the three school resource factors.

The estimates of school district and building effects were substantial, 27% of the variance. This finding was consistent with Leibenstein (1966), who observed in his article on X-efficiency in organizations, that organizational characteristics have far greater implications for efficiency than the allocation of inputs at the margins. The finding was also consistent with Levin's (1997) statement, "...the potential gains from improved allocative efficiency in education are unlikely to be as large as those from creating schools with greater X-efficiency..." p 308.

By these estimates, unobserved district characteristics (district X-efficiency) exerted a substantial influence on achievement outcomes. High X-efficiency districts (i.e., three standard deviations above the mean) were about five to ten percentile points above the mean in achievement, while low X-efficiency districts (i.e., three standard

Figure 4
Building X-Efficiency

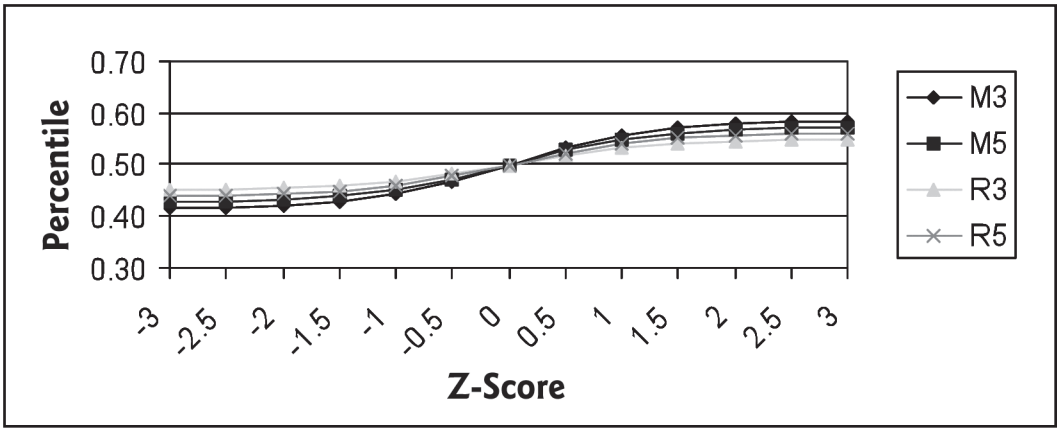
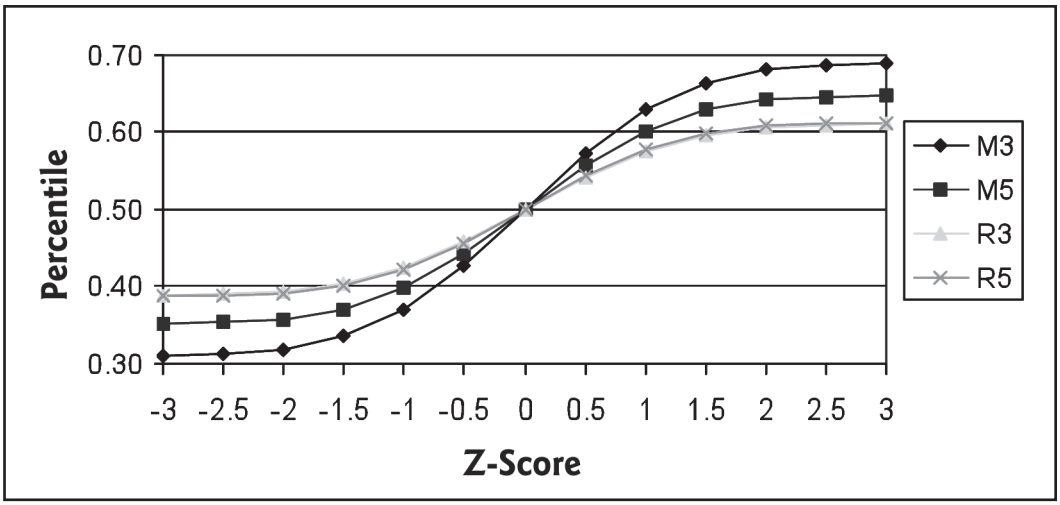


Figure 5
Building and District X-Efficiency



deviations below the mean) were about five to ten points below the mean in achievement. These effects are depicted in standardized units in Figure 3.

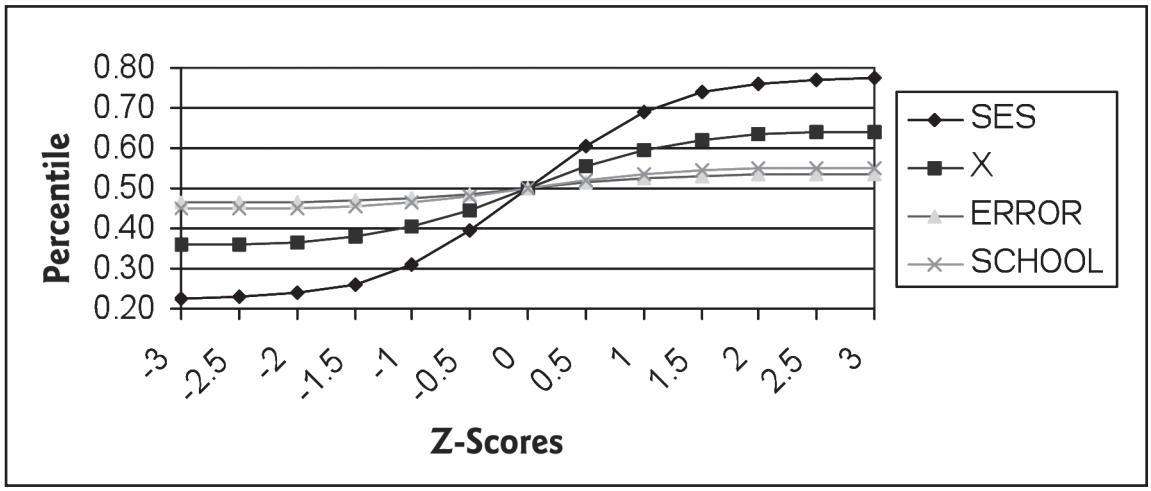
Unobserved building characteristics (building X-efficiency) also exerted an influence on achievement outcomes, with about five to eight percentile points above the mean for buildings at the high end of X-efficiency and about five to eight points below the mean for buildings at the low end. These estimated effects are depicted in standardized units in Figure 4.

Most importantly, the combined X-efficiencies of the building and district were important determinants of student achievement, far exceeding the marginal impacts of observed school resources. (See Figure 5.) Further, the correlation between building and district X-efficiency was .733, strongly suggesting a synergistic relationship between school and district. Their joint influence on achievement ranges from 10 to nearly 20 percentile points at the high end of X-efficiencies and the same at the low end. Effective buildings in effective districts apparently improve student achievement with any given level of resources.

These findings hold several important implications for school accountability policies. First, holding schools accountable for levels of achievement is tantamount to holding them accountable for the SES of the community; unadjusted scores of student achievement say little about school quality. To ascribe high quality to schools in which children attain high scores on achievement tests is to confuse school quality with student attributes. Second, when SES and school resources are taken into consideration, high and low performing schools are found in all SES strata. Holding schools accountable for achievement outcomes after SES and school resources are considered is more logical and appropriate.

While it was not the purpose of this study to examine education costs, the analysis does suggest the availability of substantial efficiencies in education production through the exploitation of school and district X-efficiencies. On average, about 55% of achievement variance is attributable to the SES factor, 27% to school and district X-efficiency, 11% to observed school resources, and 7% to random error. Of course, these estimates are confounded by multicollinearity among the factors, particularly between SES and

Figure 6
Estimate of All Factors



observed school resources. These relative effects are depicted in Figure 6. Nevertheless, the magnitude of X-efficiency substantially exceeded those of school resources. Further, the achievement gains stemming from improved X-efficiency are likely low cost. Logic suggests that time and effort devoted to the identification and dissemination of these X-efficient policies and practices are far more promising for school improvement than increases in, or marginal reallocations of, school resources.

Socioeconomic status, clearly a key determinant of academic performance, is generally thought to be beyond the control of schools. However, not all of the variables commonly used as proxies for SES (e.g., family income, parents' educational levels, etc.) are directly responsible for student achievement. Rather, the observed relationship between SES and student achievement is attributable to "achievement-friendly" behaviors (e.g., parents/guardians reading to their children and showing interest in their schoolwork, limiting television, etc.). Viewed this way, it appears possible for schools, in concert with their communities, to encourage these behaviors. Put another way, schools may have substantially more opportunity to improve student achievement than commonly assumed if families and communities are a fundamental part of any X-efficiency strategy.

A production function model of student achievement identifies school districts and buildings consistently exceeding the performance levels predicted by student and school characteristics. These schools and districts should be the subject of case studies to identify the sources of their X-efficiency. The school profiles, as suggested by this analysis, would be helpful in identifying potential schools for such case studies. Insights gained into school and district climate, policies, operations, and incentives would be invaluable, as states look for ways to improve teaching and learning in their schools in an economic environment that promises little in the way of increased resources in the near future. While leadership and teaching talent cannot always be replicated across schools and districts, effective practices and other elements of X-efficiency probably can. Case studies of this sort are not unusual in education research but are generally not conducted as part of an ongoing and systematic state-level effort to improve teaching and learning in schools. With a concerted effort between departments of education and universities, surely greater knowledge and school effectiveness is possible.

Currently, state departments of education generally do not gather information regarding the behavior, activities, policies, leadership, or instruction at the school district or building levels explaining the sources of the X-efficiency. The educational profile and school effectiveness index could serve as a template for identifying X-efficiency variables influencing student achievement. As these variables and relationships are identified, the accountability model will be enhanced.

The data historically collected by departments of education are mainly for administrative rather than educational purposes. It is the only data available for studies such as this and for implementation of NCLB. If however, the new goal were to emphasize educational purposes, educational-oriented data would be identified, collected, and integrated into the profile system outlined herein. The result would be an educational improvement profile rather than an accountability profile. It is not enough to tell schools "how they are doing," it is more important to clear evidence regarding how they could improve. What a paradigm shift that would be!

References

Bogart, W.T. & Cromwell, B.A.(1997). How much more is a good school district worth? *National Tax Journal, 50*, 215-232.

Figlio, D.N. (2005). Measuring school performance: Promises and pitfalls. In L. Stiefel, A.E. Schwartz, R. Rubenstein, & J. Zabel (Eds.), *Measuring school performance and efficiency: Implications for practice and research* (pp. 119-134). Larchmont, NY: Eye on Education.

Guilford, J.P. (1965). *Fundamental statistics in psychology and education*. New York: McGraw-Hill Book Co.

Hanushek, E.A. (1979). Conceptual and empirical issues in the estimation of education production functions. *Journal of Human Resources, 14*, 351-388.

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

Hanushek, E.A., Rivkin, S.G., & Taylor, L. (1996). Aggregation and the estimated effects of school resources. *The Review of Economics and Statistics*, *78*(4), 611-627.

Kane, T.J. and Douglas O. Staiger. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives 16*(4), 91-114.

Kuhn, T.S. (1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Leibenstein, H. (1966). Allocative efficiency and x-efficiency. *The American Economic Review*, *56*, 392-425.

Levin, H.M. (1997). Raising school productivity: an x-efficiency approach. *Economics of Education Review*, *16*, 303-311.

Monk, D.H. (1990). *Educational finance: an economic approach*. New York: McGraw-Hill Publishing Company.

Mortimore, P., et. al. (1988). *School Matters*. Berkeley, CA: University of California Press.

Murnane, R.J. & Phillips, B. (1981). What do effective teachers of inner-city children have in common? *Social Science Research*, *10*, 83-100.

No Child Left Behind (NCLB) Act of 2001. Public Law 107–110.

Schwartz, A.E., & Zabel, J. (2005). The good, the bad, and the ugly: Measuring school efficiency using school production functions. In L. Stiefel, A.E. Schwartz, R. Rubenstein, & J. Zabel (Eds.), *Measuring school performance and efficiency: Implications for practice and research* (pp. 37-66). Larchmont, NY: Eye on Education.

Stiefel, L., Schwartz, A.E., Hadj, H.B., & Kim, D.Y. (2005). Adjusted measures of school performance: A cross-state perspective. In L. Stiefel, A.E. Schwartz, R. Rubenstein, & J. Zabel (Eds.), *Measuring school performance and efficiency: Implications for practice and research* (pp. 119-134). Larchmont, NY: Eye on Education.

Taylor, J.R. (1982). *An Introduction to error analysis: The study of uncertainties in physical measurement*. Mill Valley, CA.: University Science Books.

U.S. Department of Education. (2005). *No Child Left Behind: A Road Map for State Implementation*. Retrieved September 20, 2006 from http://www.ed.gov./admins/lead/account/roadmap.pdf.

Wooldridge, J.M. (2000). *Introductory econometrics: A modern approach*. Cincinnati, OH: South-Western College Publishing.

Endnotes

[1] The author acknowledges the substantial contribution made by Michael F. Addonizio; however, the analysis and conclusions are attributable exclusively to the author.

[2] The regression equation may include the prior year's test score as an independent variable to estimate the school's "value added," or contribution to student achievement over the past year. For a good discussion of APMs, see Stiefel, Schwartz, Hadj, & Kim (2005).

[3] APMs are generally calculated with school-level data despite evidence that student-level data would yield more accurate estimates of school resource coefficients. Specifically, aggregation may exacerbate problems of omitted variables bias and overestimate the marginal contributions of school resources on student outcomes. See Hanushek, Rivkin, & Taylor, 1996).

[4] Bogart and Cromwell (1997) use revealed preferences to infer the value of public school districts from sale prices of houses in neighborhoods that are served by the same city but different school districts. The authors decompose the difference in mean house value across neighborhoods into a part due to differences in observable characteristics and an unobservable part due to differences in public services. Under a variety of assumptions about the degree of tax and service capitalization, the authors find that high-quality school districts provide services valued in excess of the higher taxes that they levy. The analysis, however, does not address school district impact on measured student achievement.

[5] Considerable controversy exists as to whether educational phenomena can be adequately represented in a strict production function framework. For an overview of the debate about the existence of an educational production function, see Monk, 1990, especially chapter 11.

[6] Individual student scores on Minnesota's reading and mathematics assessments are based on a scale ranging from a minimum of approximately 50 to a maximum of approximately 2,500. The minimum and maximum scores vary slightly from year to year according to the performance of students at the extremes of the achievement range.

[7] Each of these component variables was found to be statistically significant in regressions of student achievement for each of the four years. Each component variable was then assigned a weight inversely proportional to its variance averaged over the four years. With this weighting method, each component variable contributes approximately the same amount of variance to the total variance of the composite SES variable. The SES index is an inverse measure of socioeconomic status. That is, a higher index score reflects lower socioeconomic status. For a complete discussion of the construction of composite measures, see Guilford,1965, pp. 416-426).

[8] A set of regressions was also estimated by weighted least squares (WLS), with each observation (school) weighted by the square root of the school's enrollment. WLS is an appropriate estimation technique when one suspects that the error terms are not of equal variance for each observation (heteroskedasticity). The most common instance of heteroskedasticity is with aggregate data, such as the school-level data examined here, where the dependent variable is a mean value for the individuals in the observational unit.The accuracy of the dependent variable will be a function of the number of individuals in the aggregate; that is, observations for more populous units (e.g., schools) are presumably more accurate and should exhibit less variation about the true value than data drawn from smaller units. This leads to different values of the error term variance for each observation, the heteroskedasticity problem. In this analysis, this problem appears negligible. The unweighted regressions yielded slightly lower coefficients of determination in 14 of 16 equations as compared with the weighted regressions. The average difference was a mere .028, indicating nearly equal explanatory power across the two sets of regressions.

[9] Such school and district level variables may also systematically influence the classroom practice of individual teachers, although such practice also undoubtedly varies idiosyncratically across classrooms.

[10] See Kane & Staiger (2002). School rankings based on annual difference scores, however, are unstable due to measurement error. Tests have large stochastic components and results may be particularly volatile from year to year as different cohorts are tested (Figlio, 2005).

[11] The coefficient matrix is given in Appendix B.

Appendix A
Data (school-level)

READ3:	Mean student achievement in grade 3 in reading
READ5:	Mean student achievement in grade 5 in reading
MATH3:	Mean student achievement in grade 3 mathematics
MATH5:	Mean student achievement in grade 5 mathematics
SES:	An index of family and peer characteristics
RLADMIN:	Licensed administrators per 1,000 students
RLSUPPORT:	Licensed support staff per 1,000 students
RLINSTRUCT:	Licensed instructional staff per 1,000 students (Teachers)
RNLINSTRUCT:	Non-licensed instructional staff per 1,000 students (Aides)
Tch_yrs:	Teachers' average years of teaching experience
Tch_sal:	Average teacher salary
Tch_age	Average teacher age
Pct_mas:	Percent of teachers with a master's degree
Tot_adm:	Average daily membership
Total PP:	Total operating expenditures per pupil

Appendix B
Pearson Correlations Between School Difference Scores, SES, and Resource Indices

Test Scores by Subject and Grade	Socioeconomic Status	Resources		
		Instructional Materials	Staff Quantity	Staff Quality
MATH3	.08	-.15	-.04	-.05
MATH5	-.04	.06	<.01	.06
READ3	-.04	.06	<.01	.06
READ5	-.11	.08	.04	.13