Responding to the Charge of Alchemy: Strategies for Evaluating the Reliability and Validity of Costing-Out Research

William Duncombe
Professor of Public Administration
Education Finance and Accountability Program
Syracuse University
(315) 443-4388
duncombe@maxwell.syr.edu

Responding to the Charge of Alchemy:
Strategies Evaluating the Reliability and Validity of Costing-Out Research

Abstract

Reforming school finance systems to support performance standards requires estimating the cost of an adequate education. Cost of adequacy (COA) studies have been done in over 30 states. In several recent papers, Eric Hanushek has challenged the legitimacy of COA research, calling it "alchemy" and "pseudo-science." The objectives of this study are to present reliability and validity criteria for evaluating the scientific merit of COA studies, and to illustrate how they can be tested using cost function estimates for the state of Kansas. Based on reliability and validity estimates for Kansas, Hanushek's blanket dismissal of all COA methods as unscientific seems unwarranted. However, he has raised an important issue; reliability and validity of COA estimates have not generally been presented in published studies. To encourage more systematic evaluation of COA estimates, this research needs to move away from the advocacy environment to the realm of social science research where methods can be tested and evaluated without pressure to produce only one answer. Funding of basic adequacy research should be by neutral parties, such as foundations or the federal government, and not parties with a direct interest in the outcome.

Responding to the Charge of Alchemy:
Strategies Evaluating the Reliability and Validity of Costing-Out Research

After a decade of the standards movement (Olson, 2006) including the passage of the No Child Left Behind Act (NCLB) of 2001, high stakes school accountability systems have become the norm in most states. States, however, have been much slower to change their funding systems to support the new focus on adequate student performance (Olson, 2005). The lack of state government action to reform school finances to support adequacy has lead to litigation in over 20 states since 1989 challenging the constitutionality of state school finance systems (Lukemeyer, 2003). A key component in school aid systems designed to support a performance adequacy standard is an estimate of the cost for each district to provide their students the opportunity to reach this standard. [1] Several different approaches for estimating the cost-of-adequacy (COA) have been developed, and COA studies have been carried out in at least 30 states (Hoff, 2005).

In several recent papers, Hanushek (2005a and 2005b) has challenged the legitimacy of COA research, calling it "alchemy" and "pseudo-science." The gist of Hanushek's critique is that these methods are not scientific, and should instead be viewed as "political documents" (2005a, p. 1), funded by advocacy groups (primarily plaintiffs in adequacy lawsuits) supporting more state spending on education. He points out a number of flaws in COA methods, which he feels undermines their credibility as scientific research.

The objectives of this paper are threefold: 1) to examine systematically Hanushek's criticisms of COA research; 2) to present criteria for evaluating the scientific merit of COA

---

[1] A number of recent studies on the design of state school aid systems (Ladd and Yinger, 1994; Reschovsky, 1994; Duncombe, Ruggiero, and Yinger, 1996; Duncombe and Yinger, 1998; Duncombe, Lukemeyer, and Yinger, 2003; and Reschovsky and Imazeki, 1998 and 2001) have indicated that the best general aid formula to support a performance adequacy standard, is a modified foundation, where the foundation expenditure level should reflect the cost of providing students an opportunity to reach the adequacy standard.

studies, and discuss how they can be tested; and 3) to illustrate reliability and validity tests of cost function estimates for the state of Kansas. My principal conclusion is that Hanushek's blanket dismissal of all COA methods as unscientific is unwarranted, but COA research needs to move away from the advocacy environment to the realm of social science research where methods can be tested and evaluated without pressure to produce only one answer. Criteria for evaluating these methods, such as political acceptability and ease of understanding by legislators, should be replaced by reliability, and validity.[2] Funding of basic adequacy research should be by neutral parties, such as foundations or the federal government, and not parties with a direct interest in the outcome.

The paper is organized into four sections after the introduction. First, I will examine the substance of Hanushek's (2005a) charge of alchemy. I will discuss criteria for examining adequacy studies, and how these criteria can (or cannot) be applied to 3 different adequacy methods. Using the cost function method in Kansas, I will illustrate how the cost function results can be tested on several of these criteria. The paper will conclude with a summary, and recommendations for future research on COA methods.

THE CHARGE OF ALCHEMY

Based on his review of a number of COA studies, Hanushek (2005a) has strongly challenged the scientific basis of COA research.

---

[2] Guthrie and Rothstein (1999) dismiss the cost function approach as being too difficult for legislatures to understand. But as Downes (2003) points out rejecting the cost function method because it not easy to understand "means that other methodologies should be used in place of the cost function methodology, even if the cost function methodology is theoretically sound and is most likely to generate valid estimates of the spending levels needed to meet the standard. Taken to the extreme, this argument implies that, in choosing a method to determine adequate spending levels, one is better off choosing a method that is easy to understand but wrong rather than a method that is difficult to explain but produces the right answers." (p. 8)

The methodologies that have been developed are not just inaccurate. They are generally unscientific. They do provide reliable and unbiased estimates of the necessary cost. In a variety of cases, they cannot be replicated by others. And they obfuscate the fact that they are unlikely to provide a path to the desired outcome results. (Hanushek, 2005a, pp. 35-36)

Hanushek does not formally define what he means by scientific, but his use of the terms seems consistent with tenets of what Little (1991) calls "weak reform naturalism" (p. 232): 1) "empirical testability"; 2) "logical coherence"; 3) peer review; and 4) ideally, identification of causal explanations (p. 223). He argues that these studies do not adhere to a scientific method, because they cannot be replicated by other researchers, and their predictions cannot be tested against actual data. "In other words, the beauty of these methods is that they do not require any basis in the empirical reality of the specific state…The professional judgment panels or state-of-the-art researchers are free to declare anything without worry about being contradicted by the data." (2005a, p. 17).

Second, the logical basis for at least several of methods are flawed, because they are not designed to find the minimum spending (cost) of reaching a given performance target. Third, he argues that the purpose of COA research is not to advance knowledge about the relationship between school resources, non-school resources, and student performance, but to advance the interest of the parties that commission these studies. He implies that all of this research is little more than advocacy analysis supporting the position of a particular client, and that the research has not undergone rigorous peer-review. "They are seldom used as analytic tools to aid in policy deliberations." (p. 1) Finally, Hanushek dismisses the ability of any of these approaches to establish causality between resources and student performance.

Hanushek harshly criticizes the methodologies used by each of the common approaches. For this paper, I am going to focus on three costing out methods—professional judgment,

evidence-based studies, and cost function analysis.[3] Descriptions of these methods have appeared

in several other publications (Downes, 2004; Baker, 2006). The professional judgment (PJ)

method involves the solicitation from professional educators of their judgment on adequate

levels of resources to meet student performance standards.[4] Hanushek (2005a) labels PJ studies

as the "educator's wish list" model, because he feels that the design of PJ studies encourages

panel members to overestimate the required resources (p. 30). He questions the qualifications of

panel members to estimate the required spending, particularly for schools requiring large

increases in performance.

The evidence-based method (EB) has been characterized as a variant of the PJ method,

where instead, of using professional educators to make judgments about the resources and

programs needed to meet standards, the researchers carrying out the study select a set of

educational interventions (and resources) based on their review of the education evaluation

literature. Hanushek (2005a) labels this as the "consultant's choice model", because he asserts

that users of this approach "make little effort to assess the accumulated evidence on different

aspects of schooling." (p. 30). He argues that given the general weakness of the education

evaluation literature, the research basis for the consultant choices is a small set of studies on

particular interventions in particular states, and "there is no reason to believe that they reflect the

empirical reality *anywhere*." (Hanushek, 2005a, p. 19)

---

[3] I did not look at the "successful schools" approach, because it can only provide information on spending in a subset of districts. To estimated costs in other districts this approach has to borrow pupil weights, and regional cost adjustments from other studies.

[4] Panels may be given some information on the performance standard, existing resource use, and what programs have been found to be successful. Panelists determine staffing ratios for prototype schools, and spending for other support services and material. School staffing combined with estimated teacher salaries, and estimates of non-instructional spending are used to estimate the cost of operating the school to meet a particular performance objective. Differences exist across studies in the structure of the panels, the information and instructions provided panels, and in the process for aggregating the panel estimates into COA estimates for each district.

4

The cost function method (CF) method uses multiple regression methods and historical data to estimate the relationship between per pupil spending and student performance, controlling for differences in student characteristics, district enrollment, resource prices (teacher salaries), and district efficiency. Hanushek (2005a) labels this as the "expenditure function" approach, because he feels this method does not remove inefficiency from spending. Hanushek also argues that the finding of a large coefficient on the performance measure in a cost function is consistent with the finding that money doesn't matter in the production function literature, and does not support large increases in school spending. He claims that the education cost functions "bear little relationship to classic cost functions in microeconomic theory" because they don't include prices or capture the nature of the school budget process.

Instead, of recommending how one or several of these approaches can be tested and refined to be made more scientific, Hanushek (2005a) appears to argue that it is not possible to improve the reliability and validity of COA estimates sufficiently to be used in the design of school aid systems. Instead, decisions "on the right balance among different government programs and between public and private spending along with structuring the schools and their incentives is rightfully the province of the democratic appropriations process and not consultants hired by interested parties." (p. 2) In other words, better designs of school finance systems to support adequacy will emerge from the political process, not from research on the cost of adequacy.

To argue as Hanushek does that there is no role for technical analysis in the costing out process is akin to arguing that there is no role for technical analysis in forecasting state revenues, because forecasts by different methods and organizations can vary significantly. While state revenue forecasts by state agencies are influenced by politics (Bretschneider et al., 1989), they

are informed by technical forecasts developed using a variety of methods. I would concur with

Guthrie and Rothstein (2001) that "the appropriate point of political intervention should be the

definition and shaping of the polity's expectations for educational outcomes." (p. 104). [5]

Decisions about the objectives of a state education system, including selecting equity standards,

student performance standards, and curriculum requirements should be determined in the

political process in each state. Technical analysis has an important role to play in two parts of the

costing out process: 1) developing assessment instruments that produce valid and reliable

measures of state standards; and 2) developing accurate predictions of the cost to provide

students in a district the opportunity to support state standards. In this paper, I am focusing on

latter type of technical analysis. If the state of art in COA analysis is as inaccurate and biased as

Hanushek contends, then the solution is to try and fix the analytic tools, rather than turn the

process of estimating COA into a political bargaining exercise.

CRITERIA FOR EVALUATING COA RESEARCH

While the accuracy of Hanushek's critique of COA methods can certainly be challenged

and he provides no constructive advice on how to improve the methods,[6] he has raised an

important issue for COA research. How can the scientific merit of COA studies be evaluated,

---

[5] The example that Guthrie and Rothstein (2001) use is setting of the defense budget by Congress (p. 104). Congress and the President decide on the level of security, and then the Pentagon estimates the optimal weapon systems, staffing, etc. to support this level of security.

[6] For the cost function, for example, Hanushek (2005a) either misunderstood or mischaracterized recent education cost function research in several ways. Several recent cost functions have attempted to control for differences in efficiency. Since Hanushek (2005a) does not directly critique the efficiency controls used in these studies, it is not possible to evaluate his criticisms of these controls. Regarding the coefficient on performance, it is not clear how he defines a large coefficient, but the coefficients on outcome measures in recent cost models indicate that a one percent increase in performance is associated with a 1 to 2 percent increase in spending per pupil. In addition, in most studies these coefficients are statistically significant from zero, which is inconsistent with his characterization of a lack of a statistically significant relationship between school inputs and outcomes. Despite Hanushek's claims to the contrary, education cost functions are a modified form of a classic cost functions, and most include teacher salaries as an input price. In recognition of the interactive nature of school budget processes, several cost function studies have treated performance measures and teacher salaries as endogenous variables (Duncombe and Yinger, 2000, 2005b; Reschovsky and Imazeki, 1998, 2001).

and used to improve the methods? In this section I will discuss several criteria, and how they

can be applied to the evaluation of the PJ, EB, and CF methods. I begin with the premise that

COA studies can have two major purposes: 1) to provide reasonably reliable and accurate

estimates of the cost of reaching student performance targets that can be used in education aid

formulas; and 2) to identify a specific set of resources and educational interventions that can

produce a given level of student performance. Presumably all of COA methods share the first

objective, but not all methods attempt to accomplish the second objective, which implies

establishing a causal connection between certain resources and programs, and student

performance. In selecting criteria, I have focused initially on criteria that are consistent with the

first objective--reliability, statistical conclusion validity, construct validity, and predictive

validity (Shadish, Cook, and Campbell, 2001). I will then turn to the criteria for establishing

causality (internal validity).

Reliability

An important criteria for evaluating COA estimates is reliability, which can be defined as

the "consistency and repeatability" of a measure (Trochim, 2001, p. 88). If COA estimates have

very high measurement errors, then they do not have sufficient precision to be used in policy

decisions. Reliability is typically estimated by comparing consistency of measures of the same

phenomenon by different raters (inter-rater reliability), at different times (test-retest reliability),

or using different items measuring the same phenomenon (internal consistency). Of the three

types of reliability, inter-rater reliability appears to be the one reliability test that could be

applied to the three COA approaches.

For the PJ method, inter-rater reliability could be assessed by randomly assigning

potential panel members to several alternative panels. The panels could be given the same

instructions, and asked to estimate staffing and other instructional costs for the same prototype schools. Simple measures of variation (e.g., coefficient of variation) could be used to evaluate how large the differences were across the different panels. I am not aware of any PJ studies that have randomly assigned participants to different panels to evaluate the inter-coder reliability.[7] In an innovative application of the PJ method, Rose, Sonstelie, and Richardson (2004) survey 45 principals individually about how they would allocate resources across different prototype schools. They found significant variation across principals in how they would allocate a fixed budget, and in the level of performance they thought students would achieve.

Inter-rater reliability could be assessed for EB studies by selecting different researchers to put together the package of interventions and resources to produce the performance standards. Besides selecting interventions and resource requirements, the raters could indicate what educational evaluations were the most influential in designing their proposal, and for which resources or programs they feel the evaluation research is weak or the results mixed. For this test to have credibility at least five consultants would need to be asked, and they should produce their results without consultation with each other. Besides providing evidence on the reliability of EB estimates, this type of process could shed light on what types of programs researchers feel have strong support in the evaluation literature, and identify gaps in the evaluation literature.

Given the statistical methodology of cost function, *inter-rater reliability* could be tested by asking different researchers to estimate cost functions for the same state over the same time period. A recent example of this are the two sets of cost function studies done for the state of Texas (Imazeki and Reschovsky, 2004a, 2004b; Gronberg, et al., 2004). These studies produced significantly different COA estimates even with the same standard, which suggests is that it is

---

[7] Guthrie and Rothstein (2001) mention that there were 2 panels in Wyoming that did estimates 6 months apart and got similar results. However, 2 panels is too small a sample to judge reliability.

important for cost function researchers to examine the sensitivity of their results to different assumptions. One of the strengths of the cost function approach is that sensitivity analysis and replication of studies can be done at relatively low cost. *Test-retest reliability* tests of CF results are also feasible estimating the same cost function at several periods of time.

Statistical Conclusion Validity

The term validity refers to "the approximate truth of an inference. When we say something is valid, we make a judgment about the extent to which relevant evidence supports that inference as being true or correct." (Shadish, Cook, and Campbell, 2001, p. 34). Statistical conclusion validity refers to whether there is a statistical relationship between variables in a potential cause and effect relationship. If the statistical relationship between these variables is weak, this casts doubt on the validity of the cause-effect relationship. For COA research, one simple test of statistical conclusion validity is comparing the relationship between student performance in a district and the funding gap, defined as the difference between projected costs to meet an adequacy standard and actual spending. We would expect a negative relationship— the higher the funding gap, the worse the performance. Hanushek (2005a) examined this using the results of a PJ study for the state of North Dakota, instead, finds that funding gaps are positively related to student performance. Baker (2006) and Chambers, Levin, and Parish (in press) makes this type of comparison for several other states, and finds the expected negative relationship. While statistical conclusion validity is a fairly weak validity test, it does provide an initial assessment of whether COA estimates fit expected relationships.

Construct Validity

Construct validity refers to how well the measures developed in a study reflect the theoretical constructs they are designed to measure (Trochim, 2001). Construct validity is the

key criteria for evaluating the quality of individual measures, such as student assessment instruments.[8] Assessing construct validity of COA studies involves examining the accuracy of the basic information used to construct the estimates.

For CF studies, the accuracy and reliability of the historical data used in the statistical analysis should be assessed, by examining stability of the data across time, and correlating similar measures from different sources. For example, the most common measure of student poverty used in CF research is the percent of students receiving free or subsidized lunch. Cost function studies should examine the stability of this measure across time, and examine the correlation with the Census child poverty rate. Similar assessments of data quality should be done for other measures used in the study. For EB studies, the basic "data" are the results from education program evaluations that are used to support a certain package of interventions. The quality of these evaluations should be assessed, and reported as part of the EB study. Ideally, only high quality evaluations would be used to support the recommended interventions.

As a complex survey method, PJ studies may be susceptible to several construct validity problems: 1) can respondents answer accurately, 2) will they answer accurately, and 3) are responses influenced by the question order?[9] Hanushek (2005a) has challenged the ability of professional educators to link resources and performance in PJ panels. Instead, he argues that panel members will act strategically in their responses, which will inflate the COA estimates. Rose, Sonstelie, and Richardson (2004) argue that the use of panels without budget constraints

---

[8] Construct validity can be tested by examining how closely the content of the measure matches the underlying measure (translation validity), and by checking the performance of the measure against other criteria (criterion-related validity). (Trochim, 2001)

[9] PJ studies are comparable in many respects with another type of interactive survey commonly used in environmental economics called the contingent valuation method (CVM). Similar to a PJ study, CVM surveys provide respondents with a detailed scenario, and then ask them to estimate values based on this information. However, unlike PJ research, the reliability and validity of CVM has been evaluated extensively in the academic literature (Carson, Flores, and Meade, 2000; Mitchell and Carson, 1989; Portney, 1994; Diamond and Hausman, 1994; Hannemann, 1994), and refinements made to the surveys based on these evaluations.

could lead to overestimates of the cost of adequacy through log-rolling where everyone gets to keep their favorite proposal. The construct validity of PJ studies could be examined by systematically varying the scenario, and the constraints given panel members to see how this affects their COA estimates. For example, three different randomly selected PJ panels could be asked to estimate adequacy in two prototype schools with poverty rates of 10% and 40% (panel 1), 10% and 60% (panel 2), and 80% (panel 3). If the increase in the cost of adequacy between low poverty and higher poverty schools is about the same across the three panels, this suggests that panelists may have difficulty accurately estimating the effects of poverty on student performance. The strategic response of panel members could be tested by comparing the results with and without a budget constraint, or examining whether answers change significantly if the estimates are done individually rather than as a group.

Rose, Sonstelie, and Richardson (2004) provide the only test of strategic behavior I am aware for a PJ-type study. They asked principals to allocate resources and estimate student performance for 3 different budget levels for schools. They also asked the same principals to estimate the required resources to reach a given performance level without a budget constraint. Their finding suggests that estimates without budget constraints may be significantly larger than estimates with budget constraints in some situations.[10]

Predictive Validity

Predictive validity indicates how well a measure predicts "something it should theoretically be able to predict" (Trochim, 2001, p. 68). Predictive validity is closely related to the concept of forecasting accuracy, which measures how well a forecast of a phenomenon fits the actual values. In many respects, predictive validity is the most appropriate validity measure

---

[10] Estimates without budget constraints are 34% larger than budget constrained estimates for elementary school, but the differences for middle schools and high schools was not significant. They are careful to indicate that these results should be viewed with caution, because of differences in how the two sets of estimates were determined.

for COA studies, because COA estimates are forecasts of the minimum spending required in a district to meet a particular performance standard. The predictive validity criteria focuses on the accuracy of the bottom-line cost estimate, thus, identifying "successful" education strategies is not required to score highly on this measure. A couple cautions about using forecasting accuracy measures to judge COA estimates are in order. First, if performance standards are well beyond actual performance levels in most districts, then COA estimates are forecasts outside the sample, which can significantly increase forecasting error. Moreover, the underlying assumption of all time-series forecasts is that the past is a good predictor of the future. If the education environment is very volatile, or part of the objective of an education reform is to significantly change the environment (improve efficiency), then high forecasting accuracy based on historical data may not accurately predict the future accuracy of the forecast.[11]

Testing forecasting accuracy of CF estimates is relatively simple as long as consistent data is available for at least 5 years. Unfortunately, in a number of states, frequent changes in the assessment instruments have made it difficult to collect a time series of this length. Assuming that 5 years of data exist, forecasting accuracy can be tested by first estimating the cost model for the first three years, and using model coefficients to predict spending for the 5th year. Comparisons of predicted and actual spending can be made, and the size of the forecast error and bias can be estimated, as demonstrated in the next section.

While it should be possible to also examine the predictive validity associated with the PJ and EB methods, the nature of these methods makes testing predictive validity difficult. The performance standards that are considered by PJ panels or in EB studies may not be expressed in

---

[11]A similar criticism of using forecasting accuracy as a predictive validity test is that it is only testing the ability to predict spending and not minimum spending or cost. To some degree this is true since costs cannot be observed and predictive accuracy can only be assessed using actual data. However, if the efficiency control variables included in the cost model are adequate to remove most differences in relative efficiency, then measures of forecasting accuracy can be used to evaluate predictive validity.

terms of numeric targets.[12] To develop estimates of predictive accuracy, performance measures

need to be measured quantitatively. Testing forecasting accuracy requires that predictions can be

compared to actual values, which is difficult when performance levels for the prototype schools

are set well above present performance in most schools. It may be possible to develop estimates

of predictive accuracy of PJ estimates, if performance levels or budgets are set at lower levels.

For example, budget levels could be set at realistic levels within the state (see Rose, Sonstelie,

and Richardson, 2004), and panelists are asked to forecast student performance. Once

performance levels (and budgets) are extrapolated to all districts (based on panel estimates for

prototype schools), it would be possible to examine for districts with spending in the range of the

budget scenarios the difference between the predicted performance level and actual performance

levels.[13] Assessing the predictive validity of the EB method is difficult, precisely because these

are typically hypothetical packages of interventions. If no schools have actually used this

combination of interventions, then it is not possible to assess differences between actual and

predicted student performance.

Internal Validity

Internal validity captures the success of a study in isolating cause-effect relationships,

and is the principal criterion used to judge the quality of program evaluations. Given the

difficulty in establishing causality, particularly in non-experimental studies, this is a high bar for

any social science study to pass over (Barrow and Rouse, 2005). Not surprisingly, all three

methods face significant challenges in establishing the internal validity of the results. While

---

[12] The lack of numeric standards is viewed by some as a strength of the PJ method (Guthrie and Rothstein, 2001) since it allows a broader set of performance measures. However, this may come with the cost of less reliable and valid estimates since panel members may have a different conception of what these non-numeric performance standards mean. Imprecise standards are likely to produce imprecise estimates.

[13] Ideally, the spending and performance data used to assess the forecasting accuracy should be from different years and/or districts than those used by panels in development of their estimates.

directly testing internal validity will be difficult, COA studies can take steps to eliminate potential biases. It should be noted that if the objective of the COA method is only to accurately forecast the cost-performance relationship for school districts, establishing causality between a particular set of interventions and student performance is not necessary.[14]

For CF improving internal validity involves removing potential biases in the estimates of the cost model. Of particular concern, in statistical studies of this type are biases caused by omitted variables and simultaneous relationships between independent variables and the dependent variable. An important potential set of omitted variables are controls for efficiency. While studies have attempted to control for efficiency, if these controls are inadequate, the results could be biased.

For EB studies to establish causality between a particular package of interventions, and resources and student performance levels would ideally require a well designed program evaluation of this package of interventions. Instead, EB studies draw from the existing education program evaluation literature on individual education interventions. More information needs to be provided in these studies on which evaluations they are base their judgments, and their critique of the strengths and weaknesses of these studies.[15] The authors need to present more explicitly how they use this research to estimate the link between resources and student performance to convince readers that there is a sound basis for these judgments.[16] For PJ studies,

---

[14] Using the example of state revenue forecasting, a revenue forecast can be accurate, even if there is not a direct causal relationship between measures used in the forecast and state revenue. The extreme example of this is a univariate forecast that only using past information on the variable to be forecasted.

[15] For example, developers of the EB method started out by recommending a set of whole school reforms as the primary educational interventions. While there have been a number of evaluations done of these programs, the evidence of their effectiveness in independent evaluations is mixed (Cook et al., 1999; Cook, Murphy and Hunt, 2000; Bifulco, Duncombe, and Yinger, 2005; Bloom et. Al, 2001; Jones, Gottfredson, and Gottfredson, 1997).

[16] For example, if the study recommends use of a comprehensive school reform, such as Success for All, reduced class sizes, and pull-out tutoring for students behind grade level, they need to provide estimates of the effect of this package of interventions on different groups of students in a particular prototype school, and explain how they arrived at this estimate.

explicit tests of potential biases in the data collection process, and steps taken to eliminate bias would strengthen their credibility. For example, Hanushek (2005a) argues that panel members in a PJ study have an incentive to overestimate the required resources. To address this potential bias, PJ studies would need to estimate the size of this potential bias, and demonstrate that changes in study protocol have addressed this bias.

EXAMINING THE RELIABILITY AND VALIDITY OF COST FUNCTION ESTIMATES

One of the advantages of the cost function approach is that tests for reliability and validity can be undertaken at relatively low cost. To illustrate the implementation of several of these tests, I will use the results of a cost function analyses done for the state of Kansas. Kansas has consistent data used in the cost function for a five-year period. Before presenting results for several reliability and validity tests, I will provide a brief introduction to the cost function method, and the data and measures used in the cost model.[17]

Introduction to Cost Function Method

Applied to education, the term cost represent the minimum spending required to provide students in a district with the opportunity to reach a particular student performance level. Minimum spending can also be interpreted as the spending associated with current best practices for supporting student performance. Spending can be higher than costs because some districts may not use resources efficiently, that is, they may not use current best practices. Because we have data on spending, not costs, the cost function approach must control for school district efficiency (more on this below).   Education policy and finance scholars have established that the cost of producing educational outcomes depends not only on the cost of inputs, such as teachers,

---

[17] For a more detailed discussion of data sources, measures, and application of the cost function in Kansas see Duncombe and Yinger (2005b).

but also on the environment in which education must be provided, including the characteristics

of the student body, and size of the school district (Downes and Pogue, 1994; Duncombe,

Ruggiero, and Yinger, 1996; Andrews, Duncombe, and Yinger, 2001).

To model the relationship between spending, student performance, and other important

characteristics of school districts, a number of education researchers have employed one of the

tools of production theory in microeconomics, cost functions.[18] A cost function for school

districts relates five factors to spending per pupil: 1) student performance; 2) the price districts

pay for important resources, such as teacher salaries; 3) the enrollment size of the district; 4)

student characteristics that affect their educational performance, such as poverty; and 5) other

school district characteristics, such as the level of inefficiency. In other words, a cost function

measures how much a given change in teacher salaries, student characteristics, or district size

affects the cost of providing students the opportunity to achieve a particular level of performance

controlling for efficiency differences.

Data Sources and Measures

The cost function estimates in this paper are based on five years of data (1999-2000 to

2003-2004) for approximately 300 school districts.[19] The Kansas State Department of Education

(KSDE) is the primary source of data. The cost function study presented in this section was

---

[18]The cost function methodology has been refined over several decades of empirical application, and cost function studies have been undertaken for New York (Duncombe and Yinger, 1998, 2000, 2005a; Duncombe, Lukemeyer, and Yinger, 2003), Arizona (Downes and Pogue, 1994), Illinois (Imazeki, 2001), Texas (Imazeki and Reschovsky, 2004a, 2004b; Gronberg, et al., 2004), and Wisconsin (Reschovsky and Imazeki, 1998).
[19] Three sets of districts consolidated during this time period. To assure consistency, the information in the pre-consolidated districts is combined so that data is only available for the consolidated district. There are 1500 potential observations (300 districts x 5 years). We used only 1468 observations in estimating the cost function, because test scores are not available for 6 districts. We are able to develop cost indices for all districts, because outcomes are set at the state average.

carried out for the Kansas Legislative Division of Post Audit (LPA).[20] This section is organized

by major type of variables used in the cost model, and summary statistics are reported in Table 1.

<Table 1 about here>

District Expenditures. The dependent variable used in the cost function is district expenditures

per pupil. To broadly reflect resources used in the production of educational outcomes in Kansas

school districts, the spending measure included expenditures for six functional areas: instruction,

student support, instructional support, school administration, general administration, operations

and maintenance, and other.[21]

Student Performance. Most of the performance measures in the Quality Performance and

Accreditation (QPA) standards adopted by the Kansas State Board of Education are used in the

cost function. The key test measures in the QPA are proficiency rates for criterion-referenced

exams in math and reading in three grades (grades 4, 7, 10 for math, and grades 5, 8, and 11 for

reading). Also included in the QPA is a proxy measure for the cohort graduation rate.[22] To

construct an overall measure of student performance, we took a simple average of these seven

measures. The overall performance index varies from 0 to 100.

Student Enrollment Measures. The enrollment measure used in the study, fulltime equivalent

students (FTE), which equals total enrollment from 1st to 12th grades, and half of total enrollment

in kindergarten and pre-kindergarten programs. Enrollment measures in cost functions have

been specified as either a quadratic function (the natural logarithm of enrollment and its square),

---

[20] Most of the data used in this analysis was assembled by the staff at LPA.

[21]Spending on special education, transportation, vocational education, food service, and school facilities are excluded. The major source of spending data is the School District Budget mainframe data files maintained by KSDE.

[22] This graduation rate is equal to the number of graduates in a given year divided by total graduates plus dropouts in this year and the 3 previous years. The accountability system also includes an attendance rate and participation rate. We did not include these since both measures are very close to 100 percent and do not vary significantly across districts.

or by using a set of enrollment classes. The base model in this study uses enrollment classes, but I examine the sensitivity of the results using a quadratic specification instead.

The poverty measure used in the analysis, percent of students that are eligible for free lunch, is the at-risk student measure in the General State Aid formula in Kansas.[23] Nationally, there is some descriptive evidence suggesting that student performance in high poverty inner city schools is significantly worse than high poverty rural schools (Olson and Jerald, 1998). To examine whether this may be the case in Kansas, I have created an additional poverty variable, which is the percent free lunch students multiplied by pupil density (pupils per square mile). The higher the pupil density, the more urbanized we would expect the school district to be. Another student characteristic that can affect the cost of bringing students up to a performance level is their fluency with English. The source of the bilingual education data is the bilingual headcount districts report to KSDE. Due to potential under-reporting of bilingual headcount, it was supplemented using the percent of students, who live in a household where English is not spoken well at home from the *2000 Census of Population*.[24]

Teacher salaries. Teacher salary is the most important resource price affecting school district spending. In addition, teacher salaries are typically highly correlated with salaries of other certified staff, so that teacher salaries serve as a proxy for salaries of all certified staff. To develop a comparable salary measure across districts, data on individual teachers is used to predict what teacher salaries would be in each district if the teacher experience in the district

---

[23]Enrollment (FTE) and free lunch count are collected by KSDE from school districts using the Superintendent's Organization Report. Another measure of child poverty is the child poverty rate produced by the Census Bureau every ten years as part of the *Census of Population*. The share of free lunch students, and the Census child poverty rate for Kansas districts are strongly related (correlation = 0.7).

[24] The share of bilingual students is predicted using the Census measure of poor English spoken at home by regressed the share of bilingual students on the Census measure, and restricting the intercept to be zero to assure only positive predictions (adjusted R-square = 0.44). If the district reported it had bilingual students, we used the actual share of bilingual headcount; otherwise we used the predicted bilingual share.

equaled the state average (of teachers), and the district had the state average share of teachers with a masters, doctorate or law degrees.[25]

Efficiency-Related Measures. Costs are defined as the minimum spending, but data is only available on actual spending. Some school districts may have higher spending relative to their level of student achievement not because of higher costs, but because of inefficient use of resources. In addition, some districts may choose to focus on other subject areas (e.g., art, music, and athletics) that may not be directly related to improving test score performance in math and reading or improving the graduation rate. Controlling for relative efficiency differences across districts is an important step in estimating education cost functions.[26]

Unfortunately, directly measuring efficiency is very difficult. The approach used in this study is to include in the cost model variables that have been found to be related to efficiency in previous research: fiscal capacity, and factors affecting voter involvement in monitoring local government (Leibenstein, 1966; Niskanen, 1971; Wyckoff, 1990). Research on New York school districts indicates that taxpayers in districts with high fiscal capacity (property wealth, income and state aid) may have less incentive to put pressure on district officials to be efficient (Duncombe, Miner, and Ruggiero, 1997; Duncombe and Yinger, 2000).[27] In addition, voters might have more incentive and capacity to monitor operations in school districts with relatively

---

[25] The natural logarithm of a teacher's salary is regressed on the logarithm of their total experience and indicator variables (0-1) for whether they had a masters, doctorate, or law degree. The fit of this regression was fairly high (adjusted R-square = 0.56). We did not find that the model fit significantly improved when measures of teacher assignment (e.g., math teacher), or when measures of the teacher performance on certification exams are added to the model. There are a few districts with missing observations for salaries in a few years. We used information on predicted salaries in adjacent years and statewide trends in average salaries to impute missing salary information.

[26] If all districts are inefficient, then measures of relative inefficiency will underestimate the absolute level of inefficiency. It is not possible to estimate or control for absolute inefficiency differences, because it cannot be observed, at least in the short-run.

[27] Although aid per pupil might appear to be an appropriate way to measure the amount of aid a district receives, the underlying theory behind the use of fiscal capacity variables indicates that the appropriate measure of aid is actually per pupil aid divided by per pupil income (Ladd and Yinger, 2001). The measure used in the cost model is per pupil total aid (state general and supplemental aid plus federal aid) divided by per pupil adjusted gross income.

19

more college educated adults, more elderly residents, a larger share of households that own their own homes, or where the typical voter pays a larger share of school taxes.[28]

Cost Function Estimates

The cost function for school districts in Kansas is estimated using linear multiple regression techniques. Because spending, performance, and salaries may be set simultaneously in the budgeting process, an instrumental variable method (two-stage least squares) is used with student performance and teacher salaries treated as endogenous variables.[29] Table 2 presents the cost function results with per pupil operating expenditures as the dependent variable. Most of the independent variables are expressed in relative terms (either per pupil or as a percent).[30]

<Table 2 about here>

In general, the relationships between the different variables and per pupil spending fit expectations, and are statistically significant from zero at conventional levels. A one percent increase in teacher's salaries is associated with a 1.02 percent increase in per pupil expenditures, and a one percent increase in outcomes (as measured by reading and math test scores and the graduation rate) is associated with a 0.83 percent increase in per pupil expenditures. As expected, the cost of operating a school district is higher in small school districts. School districts with 100 or fewer students are almost 30% more expensive to operate than districts with 150 to 300 students, 45% more expensive than districts between 500 and 1000 students, and 55% more

---

[28] The latter concept is commonly referred to as a local tax share, and is measured as the median housing price divided by per pupil property values. In communities with little commercial and industrial property, the typical homeowner bears a larger share of school taxes (higher tax share) than in communities with significant non-residential property. See Ladd and Yinger (1991), and Rubinfeld (1985) for a discussion of the tax share measure used in median voter models of local public service demand. The source of most of these variables is the *2000 Census of Population*.

[29] Instruments are based on values for performance, salaries, and other socio-economic characteristics in districts in neighboring counties. Instruments are tested with an overidentification test (Wooldridge, 2003) and weak instruments test (Bound, Jaeger, and Baker, 1995).

[30] Per pupil spending, the performance index, teacher salaries, pupil density, per pupil income, and per pupil property values are expressed as natural logarithms.

expensive than districts with 1,700 or more students. As discussed above, we have included two

measures of disadvantage: 1) percent of FTE receiving free meals (child poverty measure); and

2) an adjusted measure of the share of bilingual students. We have also multiplied the share of

free lunch students by pupil density to capture any concentrated urban poverty effect. The

coefficients on all three measures are positive and statistically significant at the 10% level. The

pupil weights for the free lunch share range from 0.65 to 1.15 depending on the concentration of

poverty in the district and pupil density. The weight for the bilingual share is quite low, 0.14. It

is possible that the weight on free lunch is partially capturing the higher costs associated with

bilingual students, if many bilingual students are also eligible for free lunch. The efficiency-

related variables have the expected relationship with spending.[31]

Reliability Estimates

Test-Retest Reliability. The concept of test-retest reliability is typically used to assess how

consistent measures are when they are calculated at two reasonably close periods of time. In the

cost function case, test-retest reliability could be assessed by looking at how sensitive the cost

function results are to the time period used to estimate the model. Table 2 presents regression

estimates for 3 sub-periods: 2000-2002, 2000-2001, and 2003-04. Given that NCLB was

implemented in 2002, we might expect some variation between pre- and post-NCLB estimates.

While there are some differences in regression coefficients for the performance measure and cost

variables, the coefficients are relatively stable. However, fewer observations are statistically

significant with subsets of years.

      To evaluate the impact of these different estimates on the results of the cost model, we

estimated cost indices using regression results for the 4 time periods. A cost index indicates how

---

[31] To control for possible short-term adjustment costs among recently-consolidated districts an indicator variable for whether the district had consolidated in the last five years is also included in the cost model.

much more or less a particular district needs to spend compared to a district with average

characteristics to provide its students an opportunity to reach the same performance level. For

example, a cost index of 120 indicates that a district will require 20% more spending than the

average district to reach any student outcome level.[32] The correlations indicate a very strong

relationship between cost indices for 2004 calculated from the 4 different regression estimates

(Table 3). Comparing average cost indices by Census district type also indicates a high degree of

reliability. [33] One exception is the estimate of cost indices for large cities, which are lower when

data from 2000-01 is used to estimate the model.

<center><Table 3 about here></center>

Inter-rater reliability. Inter-rater reliability involves comparing the consistency of measures

developed using a similar method by different sets of raters. For CF studies, raters could be

viewed as different researchers making choices on how to specify the cost model. Four key

differences in model specification have appeared in the literature. First, researchers can use

different controls for efficiency in the model. Typically, this involves including different

efficiency-related variables in the model to control for omitted variable bias.[34] Besides the types

of variables included in this analysis, studies have included measures of public school

---

[32] For each variable a district can influence (outcome measure, and efficiency-related variables), the estimated coefficient of the cost model is multiplied by some constant, typically the state average for that variable. For each cost factor outside of district control, the estimated coefficient from the cost model is multiplied by the actual values for the district. The sum of the products for factors outside and within district control is used to predict costs in a district with average outcomes and efficiency. Predicted costs are also calculated for a hypothetical average district, which has average values for all variables in the cost model. Predicted spending in each district is divided by spending in this average district (and multiplied by 100) to get the overall cost index.

[33] The U.S. Census Bureau classifies Kansas City and Topeka as medium cities; however, they were reclassified for this analysis as large central cities, because they have similar socio-economic characteristics as Wichita.

[34] Downes and Pogue (1994) control for omitted variables using a panel data method (district fixed effects). In general, consistent time series are too short to used fixed effect cost models, because there is not enough within district variation for key cost factors (e.g., enrollment and poverty). Another approach is to use "stochastic frontier" regression to estimate relative efficiency for each district (Gronberg, et al., 2004). This approach involves carving the error term into random error and inefficiency. Since this approach is essentially equivalent to shifting the intercept of the regression (Ondrich and Ruggiero, 2001), it does not correct for omitted variable bias in the regression. John Ruggiero has developed a multi-stage method using data envelopment analysis (DEA) to estimate a cost index, costs to reach adequacy, and relative inefficiency, which was applied in Minnesota (Haveman, 2004)

competition (Imazeki and Reschovsky, 2004b), [35] and an efficiency index derived using a linear programming method (Duncombe, Ruggiero, and Yinger, 1996). [36] Second, there are differences in how enrollment is specified. This study uses enrollment classes, but a number of studies use a quadratic or cubic specification for enrollment. Third, there can be differences in the student need measures included in the cost model. Most studies include the percent of free lunch students (or free and reduced price lunch students) as the measure of poverty, and some measure of limited English proficiency. Some studies have included special education variables, and Imazeki and Reschovsky (2004b) also included measures of racial subgroups in the model. Fourth, studies may use different functional forms for the cost model. Most studies have used a modified version of a Cobb-Douglas (constant elasticity) function, which imposes restrictions on production technology. Gronberg et al. (2004) used a more flexible cost function, the translog cost function, which is similar to the Cobb-Douglas function, but includes a number of quadratic terms, and interaction terms.

In evaluating inter-rater reliability, I have tried to explore some of the alternative choices in model specification that I think are reasonable. For example, a cost model can be estimated without the interaction term between poverty and pupil density (Model 1). Using a quadratic enrollment specification instead of enrollment classes is another alternative (Model 2). In evaluating alternative functional forms for an education cost function, I have attempted to add flexibility to the model, without overwhelming the model with additional variables. Many of the additional quadratic and interaction terms in Gronberg et al. (2004) are statistically insignificant, which is probably due to high multicollinearity among independent variables. I included

---

[35] The measure of public school competition is a Herfindahl index capturing the degree of concentration of students within a few school districts within a county.
[36] The efficiency measure was developed using data envelopment analysis (DEA), which attempts to find the cost frontier for a set of outcome measures. This is essentially the highest student performance found among school districts at a particular spending level.

quadratic terms to allow for non-linear relationships for student performance, poverty, teacher

salaries, bilingual headcount, and various efficiency-related variables. Only the quadratic terms

on some of the efficiency variables were significant, thus I kept these additional efficiency terms

in the model (Model 4). For comparability with the base model, the three alternative models are

estimated with data from 2000-2002, and regression results are reported in Table 4.

<center><Table 4 about here></center>

Table 5 reports correlations between cost indices for the base model and the three

alternative models. All of the correlations are above 0.90, and the correlations of the alternative

models with the base model are above 0.945. While there is some variation in average cost

indices derived from different cost models by Census district type, in most cases the indices are

within 5 percentage points. One exception is Model 3 (with a quadratic enrollment specification),

which has lower cost indices for medium cities and large towns.

<center><Table 5 about here></center>

Validity Estimates

Of particular importance in evaluating COA estimates is whether they appear to be

capturing accurately the costs required to meet student performance standards. While it is

impossible to establish with certainty that any forecast of the future is reasonably accurate,

several validity tests can be undertaken to assess COA estimates. Statistical approaches, such as

cost functions, are particularly well suited to validity testing, because adjustments can be made

relatively easily, and their effects on validity assessed. For this study I am going to focus on two

types of validity assessments: statistical conclusion validity and predictive validity.

Statistical conclusion validity. Identifying whether the cost estimates appear to correlate with

other measures that they should be associated with is a simple validity test. I will use a similar

<center>24</center>

approach as Hanushek (2005a) and Baker (2006) by comparing the funding gap based on cost

function estimates to student performance. The funding gap is the percent difference between

predicted costs to meet the Kansas adequacy standard in 2006 and actual spending in 2004.[37] As

expected this relationship is negative, and the correlation between them is -0.40 (Figure 1).

While there is significant variation in this relationship, it certainly suggests that the cost function

COA estimates are on average targeting the right districts.

<Figure 1 about here>

Predictive validity. Evaluating predictive validity (forecasting accuracy) is a particularly

appropriate validity test for COA studies, because COA estimates are forecasts. In selecting a

forecasting method, it is important to consider the timeframe and objectives of the forecast

(Bretschneider, 1985; Armstrong, 2001). COA estimates are typically made for the medium

term (1 to 5 years), and the estimates need to be able to adjust to changes in student performance

and the education environment (e.g., poverty, enrollment size). Bretschneider (1985) labels these

types of forecasts as "prediction" forecasts (p. 6), and indicates that multivariate statistical

models are the typical forecasting method used for this type of forecast.

Forecasts are principally evaluated on accuracy and bias (Makridakis, Wheelwright and

McGee, 1983). The process for estimating forecasting error is to first build the forecast model

using data for one period, use it to forecast for years not in the sample, and then compare

forecasted and actual values. In this analysis I use data from 2000 to 2002 to estimate the cost

function, and use the cost function coefficients to forecast spending in 2004. Percent error is the

difference between forecasted and actual values, divided by actual values. Bias can be measured

by the *mean percent error (MPE),* which is the average percent error. If the MPE is primarily

---

[37] The adequacy standard in Kansas is set by the State Board of Education. To be consistent with actual spending, the cost function estimates for the 2006 standard are in 2004 dollars.

negative (positive), then the forecast is underestimating (overestimating) actual values.

Forecasting accuracy can be assessed by taking the absolute value of the percent error for each observation and calculating the *mean absolute percent error (MAPE).* Forecasting bias and accuracy is usually determined by comparing MPE and MAPE for a forecast with MPE and MAPE for a simple (naïve) forecast. Since COA forecasts have to adjust with changes in student performance standards, a naïve forecast would be based on a simple model of spending regressed on the student performance index.[38]

The first two columns of Table 6 compare forecasting bias and accuracy for the base model to the naïve forecast. Forecasting errors for the base model range from -31% to 39%, but there appears to be relatively little bias. The MPE is 1.1% and median percent error is 0.2%. The high errors appear to be for a small number of districts. Ninety-five percent of the districts have errors of less than 23%, and 75% have errors under 12%. By contrast, the naïve forecast has errors ranging from -51% to 28%, and the MPE indicate a systematic underestimate of spending (MPE = -7.1%). Forecasting accuracy is also significantly lower with 95% of districts with errors of 35% or less, and 75% of districts with errors of 19% or less.

<Table 6 about here>

While overall the spending estimates from the base cost model do not appear to be biased, there is still significant room for improvement in the forecasting accuracy. Using percent error as the dependent variable it is possible to assess whether there are factors that are systematically associated with forecasting error. Table 7 presents the results when percent error is regressed on the variables in the cost model. Percent error is positively associated with the performance index indicating that the cost function may be overestimating the effect of

---

[38] Both are measured in natural logarithms. The regression line is (standard error in parentheses): Y = 7.96 (0.1586) + 0.1842 (0.03827) X.

performance on costs. Percent free lunch, the interaction of free lunch and pupil density, and enrollment are also positively related to forecasting error. On the other hand, the percent of adults that are college educated and percent of population 65 and older are negatively related to percent error suggesting that the model may underestimate the impacts of these efficiency variables on spending. Since several cost function variables are systematically related to forecasting error, modifications could potentially be made to the cost function to reduce error. I included in the cost model squared terms for enrollment, poverty, and performance, and the efficiency variables. The squared terms were significantly only for enrollment and some of the efficiency variables. I also tried using other functional forms for the performance variable, but did not find that any of these improved forecasting accuracy.

<Table 7 about here>

As discussed previously 3 alternative specifications of the cost model are estimated and their forecasting bias and accuracy statistics are reported in Table 6. Removing the interaction of free lunch share and pupil density (Model 2) appears to increase forecasting error, and led more frequently to overestimates of spending. Using a different functional form for enrollment (Model 3) and also including squared efficiency variables (Model 4) appears to marginally improve forecasting accuracy for about a quarter of the districts.

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

The growth of adequacy standards and state accountability systems have lead to an interest in assessing the costs of reaching student performance standards. Cost-of-adequacy estimates and their affiliated measures (geographic cost indices, and pupil weights) are the building blocks of a school aid system focused on performance adequacy standards. The rapid

increase in COA estimates prepared for state governments, and litigants in school finance

lawsuits and the perceived large differences across studies (Hoff, 2005) has lead to a backlash

against this type of research. Hanushek (2005a) has argued strenuously that COA research is not

"scientific" because it "cannot be replicated by others" and there is a lack of "consistency across

applications." (p. 36).

The objective of this study is to systematically examine the criteria for assessing the

scientific merit of COA estimates and explore how estimates by three of the COA approaches

can be tested. Four criteria seem especially appropriate for assessing COA studies; inter-rater

reliability, statistical conclusion validity, construct validity, and predictive validity. All of the

approaches could employ some type of inter-rater reliability to examine the variation in estimates

across different participants in the process. For PJ studies this could involve comparing results

for different panels (or individual educators) given the same scenarios. For EB and CF studies

different experts could be asked to develop COA estimates for the same state and time period.

Turning to validity, all COA estimates can be evaluated on statistical conclusion validity

by comparing funding gaps (predicted spending to meet a standard minus actual spending) with

actual student performance. Hanushek (2005a), Baker (2006), and Chambers, Levin, and Parish

(in press) have evaluated PJ and CF estimates on this criteria. Construct validity can be assessed

by evaluating the quality of the data (CF) or evaluations (EB) used to construct COA estimates.

To evaluate construct validity for PJ studies, research on sophisticated surveys can be used to

examine whether panel members have the knowledge to provide meaningful answers, whether

there are any strategic biases in their answers, and how sensitive answers are to the starting point

of the estimating process. Rose, Sonstelie, and Richardson (2004) have taken a first step in this

28

direction by examining how COA estimates are affected by whether a budget constraint is imposed or not. Further research of this type can help refine the PJ method.

A stronger validity test is examining the how accurately COA methods forecast spending associated with a particular performance level. The literature on forecasting provides a number of "standards and practices" for designing accurate and unbiased forecasts (Armstrong, 2001). For the CF method examining forecasting accuracy is straight-forward as long as at least five years of consistent data are available. Developing tests of predictive accuracy for EB and PJ estimates are more difficult, although it conceivably could be done for a modified PJ study.

I illustrate the use of a couple of reliability and validity tests for cost function COA estimates for Kansas school districts. Examining both test-retest reliability and inter-rater reliability, I find that cost indices developed from different cost models are highly related to each other. While significant changes in the specification of the cost function could lead to large differences in estimates, more modest changes in model specification and time period may not dramatically affect COA results. I find that the gap between required spending and actual spending is negatively related to student performance suggesting that increased funding is targeting the right districts. The forecasting accuracy estimates indicates that the cost function is more accurate and less biased than a simple naïve forecast. For the vast majority of districts the forecasting errors are less than 20%. An assessment was made of the determinants of forecasting error and adjustments made to the cost model to improve accuracy. While most of the changes I tried for this study did not improve forecasting accuracy, the process illustrates how forecasting accuracy statistics can be used to systematically modify COA estimates to improve accuracy.

In conclusion, Hanushek's blanket dismissal of all COA methods as unscientific is unwarranted. The tests of cost function estimates for Kansas school districts indicate that they

are reliable and that forecasting accuracy for most districts is fairly good. However, he has raised an important issue; reliability and validity of COA estimates have not generally been presented in published studies. To encourage more systematic evaluation of COA estimates, this research needs to move away from the advocacy environment to the realm of social science research where methods can be tested and evaluated without pressure to produce only one answer. I have tried to highlight several reliability and validity criteria that can be used to evaluate COA estimates.

COA research should emulate the example set by of research on the contingent valuation method (CVM) in environmental economics. CVM studies also attempts to develop forecasts for a complex phenomenon (e.g., willingness to pay for environmental quality), and the results of this research have been used as evidence in high-profile litigation. However, CVM researchers have moved beyond the spotlight of public policy debates, and have devoted significant time to evaluating and refining their methods (Carson, Flores, and Meade, 2000). For COA research to respond to the charge of alchemy, similar steps need to be taken.

References

Andrews, M., W. Duncombe , J. Yinger. 2002. "Revisiting Economies of Size in Education: Are We Any Closer to a Consensus?" *Economics of Education Review* 21(3): 245-262.

Armstrong, J. 2001. "Standards and Practices for Forecasting." In J. Armstrong, ed., *Principles of Forecasting: Handbook for Researchers and Practitioners.* Norwell, MA: Kluwer Academic Publishers.

Barrow, L., and C. Rouse. 2005. "Causality, Causality, Causality: The View of Education Inputs and Outputs from Economics." *Federal Reserve Bank of Chicago Working Paper*, WP 2005-15, Chicago, IL: Federal Reserve Bank of Chicago.

Baker, B. 2006. "Evaluating the Reliability, Validity and Usefulness of Education Cost Studies." Paper prepared for the O'Leary Symposium, Chicago, IL., February 17.

Bifulco, R., W. Duncombe, and J. Yinger. 2005. "Does Whole-School Reform Boost Student Performance? The Case of New York City." *Journal of Policy Analysis and Management* 24(1): 47-72.

Bloom, H., S. Kagehiro., S. Melton, J. O'Brien, J. Rock, and F. Doolittle. 2001. *Evaluating the Accelerated Schools Program: A Look at Its Early Implementation and Impact on Student Achievement in Eight Schools.* New York: MDRC.

Bound, J., D. Jaeger, and R. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variables is Weak." *Journal of the American Statistical Association* 90 (June): 443-450.

Bretschneider, S. 1985. "Forecasting: Some New Realities." *Occasional Paper*, No. 99, Syracuse University, Metropolitan Studies Program.

Carson, R., N. Flores, and N. Meade. 2000. "Controversies and Evidence." Unpublished paper.

Chambers, J., J. Levin, and T. Parrish. In press. "Examining the Relationship Between Educational Outcomes and Gaps in Funding: An Extension of the New York Adequacy Study." *Peabody Journal of Education.*

Cook, T., and D. Campbell. 1979. *Quasi-Experimentation*. Boston: Houghton Mifflin Company.

Cook, T., R. Murphy, and H. Hunt. 2000. "Comer's School Development Program in Chicago: A Theory-Based Evaluation." *American Educational Research Journal* 37(2): 535-597.

Cook, T., H. Farah-naez, M. Phillips, R. Settersten, S. Shagle, and S. Degirmencioglu. 1999. "Comer's School Development Program in Prince George's County, Maryland: A Theory-Based Evaluation." *American Educational Research Journal* 36(3): 543-597.

Downes, T. & T. Pogue. 1994. "Adjusting School Aid Formulas for the Higher Cost of Educating Disadvantaged Students." *National Tax Journal,* 67(March): 89-110.

Downes, T. 2004. "What is Adequate? Operationalizing the Concept of Adequacy in New York." Symposium Paper for New York Education Finance Research Symposium. Albany, NY: Education Finance Research Consortium.

Diamond, P., and J. Hausman. 1994. "Contingent Valuation: Is Some Number Better Than No Number?" *Journal of Economic Perspectives*. 8: 45-64.

Duncombe, W., A. Lukemeyer, and J. Yinger. 2003. "Financing An Adequate Education: A Case Study of New York. In W. Fowler, ed., *Developments In School Finance: 2001-02*. Washington, DC: U. S. Department of Education, National Center for Education Statistics.

Duncombe, W., J. Miner, and J. Ruggiero. 1997. "Empirical Evaluation of Bureaucratic Models of Inefficiency." *Public Choice* 93: 1-18.

Duncombe, W., J. Ruggiero, and J. Yinger. 1996. "Alternative Approaches to Measuring the Cost of Education." In: H.F. Ladd, ed., *Holding Schools Accountable: Performance-Based Reform in Education*. Washington, DC: The Brookings Institution.

Duncombe, W. and J.Yinger, J. 1998. "School Finance Reform: Aid Formulas and Equity Objectives." *National Tax Journal* 51(2): 239-262.

Duncombe, W. and J. Yinger. 2000. "Financing Higher Student Performance Standards: The Case of New York State." *Economics of Education Review* 19(5): 363-386.

Duncombe, W. and J. Yinger. 2005a. "How Much Does a Disadvantaged Student Cost?" *Economics of Education Review* 24(5): 513-532.

Duncombe, W. and J. Yinger. 2005b. *Estimating the Cost of Meeting Student Performance Outcomes Adopted by the Kansas State Board of Education*. A study prepared for the Kansas Division of Legislative Post Audit.

Gronberg, T, W. Jansen, L. Taylor, and K. Booker. 2004. "School Outcomes and School Costs: The Cost Function Approach." Available at: http://www.schoolfunding.info/states/tx/march4%20cost%20study.pdf

Guthrie, J. & R. Rothstein. 2001. "A New Millennium and a Likely New Era of Education Finance," *Education Finance in the New Millennium*, Larchmont, NY: Eye on Education.

Hannemann, M. 1994. "Valuing the Environment Through Contingent Valuation." *Journal of Economic Perspectives*. 8: 21.

Hanushek, E. 2005a. "The Alchemy of "Costing Out" and Adequate Education." Paper presented at the conference *Adequate Lawsuits: Their Growing Impact on American Education.* Harvard University, Cambridge, MA, October.

Hanushek, E. 2005b. "Pseudo-Science and a Sound Basic Education, Voodoo Statistics in New York." *Education Next* (Fall): 67-73.

Haveman, M. 2004. *Determining the Cost of an Adequate Education in Minnesota: Implications for the Minnesota Education Finance System.* Minneapolis, MN: Minnesota Center for Public Finance Research.

Hoff, D. 2005. "The Bottom Line." *Quality Counts 2005*, Bethesda, MD, Education Week.

Imazeki, J. 2001. "Grade-Dependent Costs of Education: Evidence from Illinois." Draft paper, San Diego State University.

Imazeki, J., and A. Reschovsky. 2004a. School Finance Reform in Texas: A Never Ending Story. In: J. Yinger, ed., *Helping Children Left Behind: State Aid and the Pursuit of Educational Equity*. Cambridge, MA: MIT Press

Imazeki, J., and A. Reschovsky. 2004b."Estimating the Costs of Meeting the Texas Educational Accountability Standards." Available at: http://www.investintexasschools.org/schoolfinancelibrary/studies/files/2005/january/reschovsky_coststudy.doc

Jones, E., G. Gottfredson, and D. Gottfredson. 1997. "Success for Some: An Evaluation of the Success for All Program." *Evaluation Review* 21(6): 599-607.

Ladd, H. and J. Yinger. 1991. *America's Ailing Cities.* Baltimore, MD: The Johns Hopkins University Press.

Ladd, H., and J. Yinger. 1994. "The Case for Equalizing Aid." *National Tax Journal* 47(1): 211-224.

Leibenstein, H. 1966. "Allocative Efficiency vs. X-Efficiency." *American Economic Review* 56: 392-415.

Little, D. 1991. *Varieties of Social Explanation*. Boulder, CO: Westview Press.

Lukemeyer, A. 2003. *Courts As Policymakers*. New York: LFB Scholarly Publishing.

Mitchell, R., and R. Carson. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Washington, DC: Resources for the Future, 1989.

Makridakis, S., S. Wheelwright, and V. McGee. 1983. *Forecasting: Methods and Applications*. New York: John Wiley & Sons.

Niskanen, W. 1971. *Bureaucracy and Representative Government*. Chicago: Aldine-Atherton.

Olson, L., and C. Jerald. 1998. "Barriers to Success." *Education Week* (January 8).

Olson, L. 2005. "The Financial Evolution." *Education Week*, (January 6).

Olson, L. 2006. "A Decade of Effort." .*Education Week*, (January 5).

Ondrich, J. & Ruggiero, J. 2001. "Efficiency Measurement in the Stochastic Frontier Model." *European Journal of Operations Research*. 129: 434-442.

Portney, P., "The Contingent Valuation Debate: Why Economists Should Care." *Journal of Economic Perspectives* 8 (1994): 3-17.

Reschovsky, A. 1994. "Fiscal Equalization and School Finance." *National Tax Journal* 47(1): 185-198.

Reschovsky, A. & Imazeki, J. 1998. "The Development of School Finance Formulas to Guarantee the Provision of Adequate Education to Low-Income Students." In: W. Fowler, ed., *Developments in School Finance, 1997: Does Money Matter?* Washington, D.C.: U.S. Department of Education, National Center for Educational Statistics.

Reschovsky, A. & Imazeki, J. 2001. "Achieving Education Adequacy Through School Finance Reform." *Journal of Education Finance*. 26 (Spring): 373-396.

Rose, H., Sonstelie, J., & Richardson, P. 2004. *School Budgets and Student Achievement in California: The Principal's Perspective*. San Francisco, CA: Public Policy Institute of California.

Rubinfeld, D. 1985. The Economics of the Local Public Sector. In: A. Auerbach, and M. Feldstein, eds., *Handbook of Public Economics*, Amsterdam: North-Holland.

Shadish, W., T. Cook, and D. Cambell, D. 2001. *Experimental and Quasi-Experimental Designs*. Boston: Houghton Mifflin Company.

Trochim, W. 2001. *The Research Methods Knowledge Base*, Second Edition. Cincinnati, OH: Atomic Dog Publishing Company.

Woolridge, J. 2003. *Introductory Econometrics: A Modern Approach*, 2nd Edition. South-Western, Independence, Mason, Ohio Thomson Learning.

Wyckoff, P. 1990. The Simple Analytics of Slack-Maximizing Bureaucracy. *Public Choice* 67: 35-67.
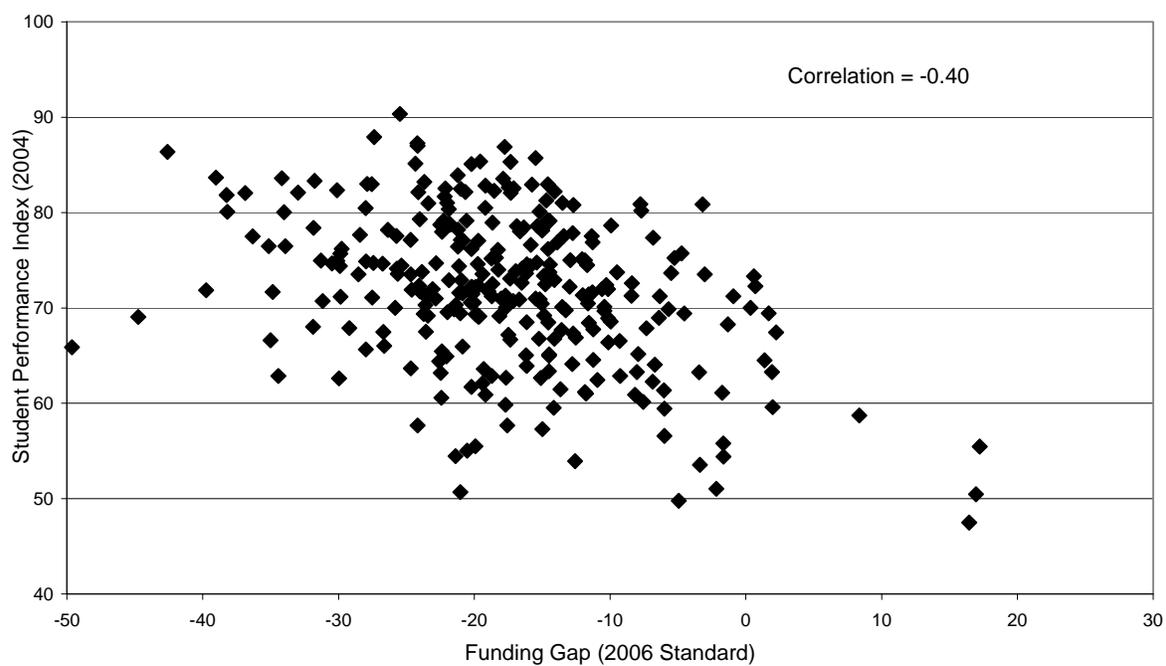
Figure 1: Comparison of Funding Gap with Student Performance Index, Kansas School Districts

Table 1. Cost Model Variables -- Descriptive Statistics (2004)

Variables	Observations	Mean	Standard Deviation	Minimum	Maximum
Per pupil expenditures	300	$6,887	$1,312	$4,915	$12,684
Combined outcome measure	294	71.4	7.9	47.5	90.3
Cost variables:					
Teacher salaries	300	$39,322	$2,949	$28,796	$49,659
Percent free lunch students	300	26.7	11.2	1.7	67.6
Free lunch share multiplied by pupil density	300	5.1	22.5	0.0	222.8
Adjusted percent bilingual headcount	300	4.2	7.4	0.0	53.5
Enrollment	300	1481.9	3828.4	60.5	45507.8
Enrollment categories (=1 if enrollment in this category):					
Under 100 students	300	0.013	0.115	0	1
100 to 150 students	300	0.040	0.196	0	1
150 to 300 students	300	0.183	0.388	0	1
300 to 500 students	300	0.230	0.422	0	1
500 to 750 students	300	0.157	0.364	0	1
750 to 1,000 students	300	0.097	0.296	0	1
1,000 to 1,700 students	300	0.103	0.305	0	1
1,700 to 2,500 students	300	0.070	0.256	0	1
2,500 to 5,000 students	300	0.060	0.238	0	1
5,000 students and above	300	0.047	0.211	0	1
Efficiency-related variables:					
Consolidated districts (=1 if consolidated last five years)	300	0.010	0.100	0	1
Per pupil property values	300	$48,588	$43,556	$721	$470,365
Per pupil income	300	$82,930	$30,972	$4,390	$312,999
Total aid/income ratio	300	0.08	0.10	0.00	1.78
Local tax share	300	1.37	0.88	0.00	4.58
Percent of adults that are college educated (2000)	300	17.97	6.74	5.78	64.44
Percent of population 65 or over (2000)	300	16.87	5.49	0.61	29.33
Percent of housing units that are owner occupied (2000)	300	88.56	5.67	70.00	97.92

Note: Comparable teacher salaries are estimated using state average teacher experience and average education. Adjusted percent bilingual headcount is calculated by first regressing the share of bilingual headcount on the Census measure of poor English (with no intercept). The predicted value from this regression is used as the estimate of the share of bilingual headcount, except in those districts where the share of bilingual headcount is greater than zero.

Table 2. Cost Function Estimates for Different Years for Kansas School Districts

Variables	2000-2004		2000-2002		2000-2001		2003-2004	
Intercept	-6.84027		-7.48244		-1.09761		-7.38553	
Performance measure	0.83013	*	0.83964	*	0.52375		1.16898	
Cost variables:								
Teacher salaries	1.01765	*	1.02435	**	0.53731		0.99824	
Percent free lunch students	0.00636	*	0.00707	*	0.00494		0.00739	
Free lunch multiplied by pupil density	0.00065	**	0.00083	**	0.00051		0.00071	
Adjusted percent bilingual headcount	0.00139	**	0.00168	**	0.00150	**	0.00140	**
Enrollment categories:								
100 to 150 students	-0.12987	**	-0.04699		-0.08882		-0.30470	*
150 to 300 students	-0.29443	*	-0.21412	*	-0.22596	*	-0.46311	*
300 to 500 students	-0.38580	*	-0.28923	*	-0.29834	*	-0.56716	*
500 to 750 students	-0.44523	*	-0.33661	*	-0.34116	*	-0.65338	*
750 to 1,000 students	-0.45612	*	-0.34639	*	-0.35010	**	-0.65368	*
1,000 to 1,700 students	-0.52671	*	-0.41565	*	-0.43104	*	-0.73383	*
1,700 to 2,500 students	-0.57252	*	-0.46577	*	-0.49239	*	-0.75921	*
2,500 to 5,000 students	-0.56802	*	-0.45470	*	-0.47583	*	-0.78139	*
5,000 students and above	-0.55366	*	-0.45975	*	-0.46418	**	-0.74616	**
Efficiency-related variables:								
Consolidated districts	0.14780	*	0.14539	*	0.14942		0.16214	**
Per pupil income	0.13097	*	0.16950	*	0.14501	*	0.08552	
Per pupil property values	0.05341	*	0.05392	**	0.08154	**	0.05188	
Total aid/income ratio	0.80593	*	1.10645	*	0.98858	*	0.32826	*
Local tax share	-0.02102		-0.02198		0.00197		-0.00767	
Percent of adults that are college educated (2000)	-0.00666	*	-0.00694	*	-0.00471		-0.00745	*
Percent of population 65 or older (2000)	-0.00347	*	-0.00460	*	-0.00279		-0.00365	
Percent of housing units that are owner occupied (2000)	-0.00218	**	-0.00240		-0.00252		-0.00243	
Year indicator variables:								
2001	-0.02209		-0.02532		-0.00065			*
2002	-0.01666		-0.02215			*		*
2003	-0.08637			*		*		*
2004	-0.13924	**		*		*	-0.07353	**
Sample Size	1468		880		585		585	

Note:Estimated with linear 2SLS regression with the log of per pupil operating spending as the dependent variable. Performance and teacher salaries are treated as endogenous with instruments based on variables for adjacent counties. Robust standard errors are used for hypothesis testing. The performance index, teacher salaries, per pupil income, per pupil property values and local tax share are logged. * indicates statistically from zero at 5% level. ** indicates statistically significant from zero at 10% level.

Table 3. Comparisons Between Cost Indices for Different Years for
Kansas School Districts

	2000-2004	2000-2002	2000-2001	2003-2004
Correlations:				
2000-2004	1			
2000-2002	0.985	1		
2000-2001	0.954	0.984	1	
2003-2004	0.947	0.984	0.969	1
Averages by Census Region:				
Large central cities	124.1	131.2	115.0	129.3
Medium cities	92.3	93.6	98.8	91.4
Urban fringe of large cities	87.3	87.7	86.5	85.9
Urban fringe of medium cities	98.2	91.5	92.2	91.5
Large town	101.2	103.4	98.0	101.7
Small town	95.7	97.2	95.3	94.8
Rural metro	105.2	104.6	105.6	106.7
Rural non-metro	94.3	94.1	95.2	93.0

Table 4. Cost Function Estimates for Different Years,
Kansas School Districts[a]

Variables	Model 2		Model 3		Model 4	
Intercept	-11.78550		-1.86848		-2.25176	
Performance measure	0.99501		0.74754	**	0.76397	*
Cost variables:						
Teacher salaries	1.33459		0.65221		0.92349	
Percent free lunch students	0.00885		0.00701	**	0.00661	*
Free lunch multiplied by pupil density			0.00111	**	0.00116	*
Adjusted percent bilingual headcount	0.00136		0.00192	**	0.00177	
Enrollment categories:						
100 to 150 students	-0.04327	*				
150 to 300 students	-0.22589	*				
300 to 500 students	-0.29484	*				
500 to 750 students	-0.34478	*				
750 to 1,000 students	-0.35751	*				
1,000 to 1,700 students	-0.41948	*				
1,700 to 2,500 students	-0.45986	*				
2,500 to 5,000 students	-0.44092	*				
5,000 students and above	-0.42882					
Enrollment			-0.36778	*	-0.25088	**
Square of log of enrollment			0.01883	*	0.01329	**
Efficiency-related variables:						
Consolidated districts	0.15376	*	0.16069	*	0.11827	
Per pupil income	0.20631		0.15238	*	0.55125	*
Per pupil property values	0.05085	*	0.05850		-1.47490	*
Total aid/income ratio	1.33477		0.86027	*	6.31573	*
Local tax share	-0.02756	**	-0.00003		0.01599	
Percent of adults that are college educated (2000)	-0.00879		-0.00469	**	-0.00724	**
Percent of population 65 or older (2000)	-0.00541		-0.00515	**	-0.00649	
Percent of housing units that are owner occupied (2000)	-0.00203		-0.00198	**	-0.00086	
Square of efficiency related variables:						
Per pupil income					0.00000	*
Per pupil property values					0.07427	*
Total aid/income ratio					-9.58861	*
Local tax share					-0.00906	
Percent of adults that are college educated (2000)					0.00005	
Percent of population 65 or older (2000)					0.00009	
Year indicator variables:						
2001	-0.04149		-0.01283		-0.03215	
2002	-0.04934	*	0.00298		-0.03771	

Note:Estimated with linear 2SLS regression with the log of per pupil operating spending as the dependent variable. Performance and teacher salaries are treated as endogenous with instruments based on variables for adjacent counties. Robust standard errors are used for hypothesis testing. The performance index, teacher salaries, per pupil income, per pupil property values and local tax share are logged. * indicates statistically from zero at 5% level. ** indicates statistically significant from zero at 10% level.

Table 5. Comparisons Between Cost Indices for Different Models for Kansas School Districts

	Base Model	Model 2	Model 3	Model 4
Correlations:				
Base Model	1			
Model 2	98.8	1		
Model 3	95.4	92.1	1	
Model 4	94.5	95.6	90.74	1
Averages by Census Region:				
Large central cities	131.2	128.9	126.3	138.6
Medium cities	93.6	95.9	85.0	98.6
Urban fringe of large cities	87.7	88.1	86.1	92.9
Urban fringe of medium cities	91.5	90.8	91.9	93.8
Large town	103.4	106.7	97.4	105.6
Small town	97.2	98.5	95.7	99.0
Rural metro	104.6	104.0	106.6	102.9
Rural non-metro	94.1	93.4	92.7	93.9

Table 6. Estimates of Forecasting Error
(Difference Between Predicted and Actual as a Percent of Actual)

Distribution	Naïve Forecast	Base Model	Model 2	Model 3	Model 4
Bias (percent error)					
Mean (MPE)	-7.1	1.1	4.8	0.4	1.2
Median	-6.8	0.2	4.3	-0.6	1.2
Minimum	-50.7	-31.3	-31.6	-27.5	-34.8
5th percentile	-35.2	-17.4	-16.7	-15.7	-16.5
10th percentile	-26.5	-11.3	-9.1	-11.6	-9.9
25th percentile	-16.3	-5.9	-2.9	-6.1	-5.0
75th percentile	2.0	7.6	12.0	6.7	7.6
90th percentile	12.5	16.1	21.5	14.2	15.7
95th percentile	18.7	20.3	28.0	18.0	17.8
Maximum	27.9	39.5	46.7	35.2	34.8
Accuracy (absolute percent error)					
Mean (MAPE)	13.1	8.5	10.4	7.9	7.9
Median	10.6	6.7	7.7	6.4	6.2
Minimum	0.0	0.1	0.0	0.0	0.0
5th percentile	0.9	0.6	0.6	0.8	0.7
10th percentile	1.6	1.4	1.2	1.4	1.2
25th percentile	4.6	3.1	3.5	2.9	2.6
75th percentile	19.1	11.6	14.9	11.6	11.1
90th percentile	27.4	19.5	23.9	17.6	17.6
95th percentile	35.2	22.8	28.3	19.3	20.9
Maximum	50.7	39.5	46.7	35.2	34.8

Note: Naïve forecast is base on log of per pupil base spending regressed on log of performance index.

Table 7. Factors Associated with Forecasting Bias for
Kansas School Districts

Variables	Regression 1		Regression 2	
Intercept	-362.469	*	-414.056	*
Performance measure	70.937	*	71.194	*
Cost variables:				
Teacher salaries	1.111		-1.955	
Percent free lunch students	0.363	*	0.370	*
Free lunch multiplied by pupil density	0.053	**	0.080	*
Adjusted percent bilingual headcount	0.126		0.156	
Enrollment categories:				
Enrollment	3.960	*	15.388	*
Square of enrollment			-0.812	**
Efficiency-related variables:				
Consolidated districts	-5.444		-5.997	
Per pupil income	4.333		7.128	**
Per pupil property values	-1.140		-0.227	
Total aid/income ratio	10.736		46.472	
Local tax share	0.648		0.237	
Percent of adults that are college educated (2000)	-0.645	*	-0.573	*
Percent of population 65 or older (2000)	-0.387	*	-0.418	*
Percent of housing units that are owner occupied (2000)	-0.077		-0.092	

Note: Estimated with OLS. Data is for 1999-2000 to 2003-02.The performance index,
teacher salaries, enrollment, per pupil income, per pupil property values, and local tax
share are expressed as natural logarithm. * indicates statistically from zero at 5% level. **
indicates statistically significant from zero at 10% level.