Using Adjusted Performance Measures for Evaluating Resource Use

LEANNA STIEFEL, ROSS RUBENSTEIN, AND AMY ELLEN SCHWARTZ

Public service organizations are looking for ways to improve the evaluation of performance and resource allocation. One of the approaches is to use adjusted performance measures, which attempt to capture factors that affect the organizational performance but are outside of the organization’s control. This article illustrates the construction and use of adjusted performance measures to assess the performance of public schools, and reports findings from a study of school-based budgeting in Chicago that relates adjusted performance measures and patterns of budget allocations.

As budgets tighten and financial resources become increasingly scarce, the notion that public organizations are inefficient and wasteful is enjoying increasing popularity, leading to heightened interest in discovering methods to increase the “bang” received for the public buck spent. One solution to the perceived inefficiency problem lies in the improvement of methods for identifying and rewarding high performance. Thus, policymakers, client groups and funders have begun to require public organizations to gather data that can be used to evaluate performance and resource allocation, in the hope of connecting the two. A useful approach for doing so involves the construction of adjusted performance measures to evaluate the efficiency and efficacy of resource allocation. This article outlines the theory underlying these measures and discusses the issues presented by their practical application. An illustration of these techniques is provided using data on the Chicago Public Schools.

Leanna Stiefel is a Professor of Economics and Director of The Public and Nonprofit Management and Policy Program at the Robert F. Wagner Graduate School of Public Service, New York University, New York, NY 10003. E-mail address: leanna.stiefel@wagner.nyu.edu. Ross Rubenstein is Assistant Professor of Public Administration and Urban Studies in the Andrew Young School of Policy Studies and the College of Education at the Georgia State University, Atlanta, GA 30303. E-mail address: rrubenstein@gsu.edu. Amy Ellen Schwartz is Assistant Professor of Public Administration at the Robert F. Wagner Graduate School of Public Service, New York University, New York, NY 10003. E-mail address: amy.schwartz@wagner.nyu.edu.

We gratefully acknowledge funding from the Andrew W. Mellon Foundation. All the conclusions are ours.

Stiefel, Rubenstein & Schwartz / Adjusted Performance Measures for Evaluating Resource Use 67
The push to enhance the evaluation of performance and resource allocation efforts has spread across a wide range of federal, state and local governmental organizations, as well as nongovernmental public service organizations. At the federal level, the Government Performance and Results Act of 1993 mandates federal agencies to establish performance indicators and to link them to budgetary requests.\(^2\) The Governmental Accounting Standards Board, through its Service Efforts and Accomplishments project, has been studying ways for state and local governments to combine output and outcome measures with financial information.\(^3\) Health care organizations are under pressure to measure health outcomes.\(^4\) Elementary and secondary school systems frequently publish report cards showing how well the schools in their districts are doing.\(^5\) A variety of school districts, including large ones such as New York City and Chicago, are working toward producing school-level resource use data along with performance evaluations.\(^6\) Finally, funders are increasingly requiring nonprofit organizations to justify support with accounts of performance.\(^7\)

This is not the first time the United States has seen widespread concern for performance, and some organizations have considerable experience producing and refining outcome measures. Public schools, hospitals, and public job training providers, in particular, have developed somewhat sophisticated performance measures that adjust for differences in the myriad factors outside the organization's control that may affect an organization's apparent performance.\(^8\) As an example, using a hospital's mortality rate to measure the hospital's performance in treating heart attack patients clearly requires some adjustment and care in interpretation because the rate is a function of both the characteristics of the case mix and the characteristics of the hospital itself.\(^9\)

In the area of budgeting, systems such as PPBS or performance budgeting that emphasize output or outcome measures have been popular at several times during the past half century at the federal, state, and local levels. Performance-based budgeting, in particular, is experiencing renewed popularity at all levels of government. Joyce\(^10\), for example, reviews federal efforts to link budgeting and performance measurement, while Melkers and Willoughby find that all but three state governments have some type of performance-based budgeting requirement.\(^11\) Despite the prevalence of performance budgeting in public organizations, Joyce notes four challenges that may limit the utility of these efforts: lack of agreement about objectives; insufficient data for measuring costs; lack of appropriate output and outcome measures; and public managers' limited control over resource allocation decisions.\(^12\)

Efforts to link budgeting and performance have not been limited to the United States. Schick reviews efforts in Great Britain, Canada, Australia, Denmark and Sweden to link resource allocation decisions to performance.\(^13\) Much like current U.S. efforts, governments in each of these countries attempted to shift the focus of budgeting from inputs to results, although Schick points out that none of these countries established a "tight relationship" between the two.

While there is some experience in producing adjusted performance measures in particular settings, a variety of conceptual and practical issues need to be addressed...
before organizations begin using performance measures to evaluate the effectiveness of existing or future resource allocations. This article illustrates the construction and use of adjusted performance measures to assess the performance of public schools, and reports findings from a study of school-based budgeting in Chicago that relates adjusted performance and patterns of budget allocations. The specific topic is of interest in its own right because school-based budgeting is one of several reforms aimed at improving performance of schools in large cities. Further, no systematic evaluations of school-based budgeting efforts yet exist. More generally, however, the topic is a useful illustration of the conceptual and methodological issues involved in utilizing performance measures to evaluate resource allocations and to guide budgeting decisions.

The first section of this article reviews the construction and interpretation of adjusted performance measures. The second section addresses several difficult and unresolved conceptual issues inherent in constructing such measures. The third section discusses the use of adjusted measures to evaluate performance and resource allocation. Section four considers a specific application of these concepts by examining the relationship between school-level resource allocation patterns and measured performance of schools in Chicago. Section five concludes with a discussion of lessons and implications for constructing and using performance measures in general, and in the evaluation of school resource use in particular.

**WHAT ARE ADJUSTED PERFORMANCE MEASURES?**

In general, a performance measure is a quantitative estimate of an organization’s output (loosely defined as work products) and/or its outcomes (the consequences of the outputs). As an example, an output measure for a high school might be days of instruction or number of graduates; an outcome measure might be the average income earned by the school’s graduates. Clearly, both the output and outcome of any individual school may depend not only upon the resources, performance and efficiency of the school per se, but also upon a set of variables that are outside of the school’s control. For example, graduation rates may be affected by the extent to which students have adequate familial and financial support while attending high school, and subsequent income may depend upon graduates’ ability to pay for college education. Thus, performance measures need to be adapted in some way if they are to be effectively used to compare performance across organizations operating in different environments and with different client mixes. Adjusted performance measures take account of such variables or correlates (factors) that are outside the individual organization’s control.

The adjustment typically begins with the identification of a set of variables that capture uncontrollable factors affecting an organization’s performance (denoted Z). The next step is to estimate a regression equation in which the performance measure of interest, P, is regressed on this set of variables, pooling data on a set of similar organizations. The adjusted performance measures are then constructed as the “predic-
tion error” for each organization. That is, the adjusted performance measure for the organization, \( M_i \), is the difference between the value of the performance measure predicted by the regression equation for that organization and the actual value realized.\(^{17}\) Constructed in this way, \( M_i \) is meant to measure performance while accounting for the differences in the organization’s work environment or the characteristics of the clients whom the organization serves. Notice that the \( M_i \) can form the basis for a comparison of diverse organizations’ relative performance. At the simplest, \( M_i > 0 \) indicates the organization has produced a level of \( P \) that is greater than is predicted by the regression equation, suggesting better than average performance. \( M_i < 0 \) indicates the organization has produced a level of \( P \) that is less than that predicted by the regression, which suggests lower or poorer performance. (By construction, \( M_i \) has mean zero, which can be interpreted to indicate that the “average” organization has an \( M_i \) equal to zero). Of course, the \( M_i \) can be compared (and subject to greater statistical analysis) to provide more insight into differentiating, for example, high and low performing organizations. (Section three addresses this point in greater detail).

Notice that the regression might also include a set of variables, \( X \), which could be controlled under some circumstances or are controllable by some authority even if not the organization itself. Such variables might typically describe the resources used to produce the output.\(^ {18} \)

If the \( X \)'s are included in this way, however, \( M_i \) must be interpreted differently. Here, the \( M_i \) captures the extent to which the organization’s performance differs from that of other organizations given the differences in the variables in \( X \) across organizations. This approach may be particularly appropriate if these variables are susceptible to policy intervention by some governing body “above” the organization itself. As an example, an individual public school typically has little control over the level of resources it has to perform its duties. Thus, inclusion of resource variables may be appropriate in school performance measurement analyses, even though there are others outside the school—such as the local Board of Education, school district administrator, or government officials—who do control these resources.

While creating adjusted performance measures is appealing and addresses potential shortcomings of the raw performance measures, there are a variety of difficulties and issues that need to be addressed before relying on them for decision making. These include: choosing between alternative raw performance measures; using existing data for performance measures; choosing independent variables; adjusting for initial conditions (or not); and other technical issues. These are discussed in the next section.

**ISSUES IN CONSTRUCTING ADJUSTED PERFORMANCE MEASURES**

*Choosing a Performance Measure*

The first step in constructing an adjusted performance measure is choosing the raw performance measure for analysis. Given the multidimensionality of services provided
by any organization, it is rare that any single measure is immediately identified as capturing all dimensions of performance. Sometimes more than one measure is analyzed and then the results are combined in a summary statistic, or index number. Or, the results of multiple analyses may be presented separately, allowing users to draw conclusions on their own. In practice, these decisions are dictated in no small part by practical considerations such as availability and ease of interpretation. The result is that some performance measures may become "standard" for a particular type of organization or in an industry, even while there is dissatisfaction with them on conceptual grounds.

In elementary and secondary schools, for example, standardized test scores are a commonly used performance measure. This widespread use notwithstanding, much discussion and debate continues around the question of whether test scores are, in general, good enough performance measures for schools. Students study many subjects, so the use of more than one test score will provide a better picture of a school's performance and will encourage schools to resist deploying inordinate resources to achieve one output at the expense of other desirable ones. But, of course, good test scores are not really what schools are supposed to be producing; test scores are simply readily available, generally low cost assessment tools. If schools are rewarded too heavily on the basis of test score performance, then other less-easily-measured learning, as well as other outputs and outcomes such as grade advancement rates, college-going rates, and future employment rates, may be overlooked. Depending on how correlated each type of output is with another, using one type of measure may skew results in unintended directions.

*Using Existing Data for Performance Measurement*

Given that a choice is made on theoretical grounds about an appropriate measure, practical issues still remain. How, exactly, are the data for performance measures to be collected? How shall the various organizations be induced to provide comparable data? How expensive and reliable is the collection likely to be? Even if a conceptually appealing set of performance measures can be identified, will a typical organization have (or be persuaded to collect) the data needed to calculate the measures? These practical considerations are as important as the conceptual ones, and are often the deciding factors in the choice of performance measures.

In the context of the schools, these questions may be stated more specifically: while agreement might be reached about the use of test scores as performance measures, are the specific test scores generated routinely by school systems good enough? Or, are different tests or testing procedures necessary? Although most school districts have some test score data available, the data are often not ideal for constructing good performance indicators. At least four different problems may arise. First, tests vary as to whether they are measuring basic, average or high-end competency. If tests are meant to measure basic (or low level competency), then they will not yield much
information on schools that routinely score well (or vice versa). Second, some tests are norm-based and others are criterion-based. Norm-based tests are most familiar to the public; they are scored relative to the performance of all test takers, with percentile scores indicating a school’s position on a distribution of all test takers. Criterion-based tests are constructed to reveal if students know specific information; every school could do well if students in the schools were able to answer questions about specific subject areas correctly. Not all districts use both kinds of tests and one may be more appropriate for one district’s goals than another. Third, tests in specific subject matter are often given only in selected grades so that it will be unfeasible to obtain a good picture of student progress in the selected areas from one year to the next. Fourth, the particular test used in a district may be changed between years, or the test company may norm each year’s test differently. All of these problems could be resolved at some cost. Using low-cost, readily-available data, however, will usually mean encountering some of these problems.21

Finally, individual outputs need to be translated into organization-level performance measures. To return to the schools’ example, individual student test scores can be aggregated to the school level in many different ways. The most common ways are to compute the average score (arithmetic mean) or the percent reaching a pre-determined level. Other possibilities are to use the median score or the score at another percentile (the 90th for example). Each method may produce different results for individual schools and if the measures are used for resource allocation decisions, they will present the schools with different kinds of incentives. Using a mean provides the incentive for schools to pay extra attention to students who can be induced to produce the highest scores or the highest gains. Using medians provides the incentive to focus attention on those students who are performing in the vicinity of the median—ignoring anyone well above or below the 50th percentile.

Notice that the same sorts of issues arise in many organizational settings or industries. Using average length of stay at a hospital provides incentive to focus attention on the relatively small number of lengthy stays. Using the median will focus attention on those near the median, and draw attention away from those with lengthy stays. Clearly, there is no one right answer. Ultimately, these choices must be made in light of the purpose for which the measures will be put.

Choosing Independent Variables—Uncontrollable Factors

Given some decisions about performance measures, the next question is: what variables should be included in Z? As described thus far, these should be variables describing the particular characteristics of the organizations and its clients, which are typically outside the control of the organization. These independent variables are included to capture the relative difficulty (or ease) that the organization has in “earning” high levels of measured performance, compared to other organizations in the group. Again the choice of variables will be dictated—at least in part—by cost and availability, but
there is likely to be some consensus among organizations about what measured factors make their jobs harder (or easier) to do well.22

Assuming, then, that uncontrollable factors will be included, which ones should they be? In analyses of schools, socioeconomic ones are usual, although data availability may limit the choices. Income level or poverty may be captured by a variable based on the percentage of the students who are eligible for free or reduced price school lunch since these data are likely to be collected to fulfill federal funding requirements. Similarly, data on the fraction of the students with limited English proficiency may be available and used to capture higher associated costs. Race and ethnicity variables are commonly included, although the use and interpretation of the estimated coefficients is not straightforward. Negative relationships between performance and the representation of some racial or ethnic group may be viewed as implying lower performance expectations for these groups.23 Alternately, this information could be viewed as providing important insight into sub-populations that might benefit from differential treatment where needed.24

Choosing Independent Variables—Controllable Factors

As described above, variables may be included that are either controllable in principle or could be controlled by some governing authority other than the organization itself. The typical candidates for inclusion are variables capturing the resources used by the organization. In school analyses, the most commonly used resource variables are measures of class size, pupil-teacher ratios, or various types of spending per pupil.

Under what circumstances should resource variables be included? If they are not, the coefficients of the included socioeconomic variables will reflect not only their direct impact on performance but also an indirect effect through their impact on resources. That is, if resources are omitted, then the coefficient on an included variable capturing limited English proficiency (LEP) will reflect both its direct effect on performance and the indirect effect caused by the additional resources that might follow students who are not native speakers of English. If resources are included, then these effects will be somewhat disentangled. Unfortunately, to the extent that higher performing schools are also better able to attract funding (perhaps by attracting private sector support or through more effective utilization of public funding programs), the inclusion of resource variables may introduce endogeneity into the equations and, therefore, bias the coefficient in an unknown direction.25

A pragmatic approach to resolving this issue is to estimate the equations with and without the resource variables to assess the difference their inclusion makes. Somewhat more appealing is to evaluate whether resource variables should be included on conceptual grounds. The decision should depend upon the use to which the resulting adjusted performance will be put and on an assessment of the extent to which endogeneity is likely to be a problem. For the purpose of evaluating the impact of budget allocations in schools where overall resources are not easily controlled by the

Stiefel, Rubenstein & Schwartz / Adjusted Performance Measures for Evaluating Resource Use 73
schools, including resource variables in estimating performance equations is warranted
and allows the disentangling of the impacts of patterns of allocations from the impact of levels of resources.

Choosing Independent Variables—Controlling for “Initial Conditions”

An important issue revolves around whether (or how) to include information on measured performance from a previous period. If the prior year’s output, P, is included as an explanatory variable, it is typically a powerful predictor of the subsequent year’s performance, and improves the “fit” of the equations dramatically. Simply put, the best predictor of this year’s performance is often last year’s performance. An alternative method of adjusting for previous performance is to use changes between periods as the performance measure. The resulting measures are sometimes referred to as “value-added” measures. These alternatives will give the same results only if the regression coefficient on prior year performance is one—that is, if every one unit increase in last year’s output translates into an expected one unit increase in this year’s performance.26

Now, which of these—if any—is appropriate to measure performance? If performance measures are used to allocate any kind of scarce resource, then it makes sense to look at how the organization has done over the period of consideration. For example, if a school started the year with very low performance on the part of its students but brought those students up to a high level, while a second school began with high performance but added little, the two schools should not be considered to have performed similarly. Including pre-tests may ameliorate some of the problems associated with including information on uncontrollable factors since it is not clear whether schools with high percentages of at-risk students will do better or worse on gains.27 On the other hand, for reporting purposes, absolute levels are important as well. A change from 10 percent to 20 percent of children reading at grade level is a good change, but might still result in a level of performance that the community finds unacceptable. It is also possible to use levels of indicators, but to set targets for the next period’s levels that challenge the organization to keep improving.28

Additional Issues

There are, as usual, a series of additional concerns and decisions that need to be made. Thus far, we have only discussed estimating fairly straightforward linear equations, although the “theory” of performance adjustment puts no particular bound on the form of this function. Why should not some (or all) of the variables be specified as logarithms, allowing the estimation of elasticities rather than marginal contributions? What about interaction variables? What about heteroskedasticity, especially if organizations differ in size and the performance measure is a mean value? The list could be much longer. In general, our intention has been to favor the simplest functional form and estimating methodology in order to make this already-complicated adjustment proce-
dure most accessible and understandable to policymakers and the public, and therefore, more likely to be used to guide decision making. 29

ISSUES IN USING ADJUSTED PERFORMANCE MEASURES TO EVALUATE PERFORMANCE

Adjusted performance measures can be estimated to serve a variety of purposes. One important application is to help gain insight into the efficacy of policy alternatives for improving performance. Thus, police station performance measures may be estimated to evaluate the effectiveness of different policing strategies being considered for dissemination. As illustrated below, another important application is the use of performance measures to evaluate the relative merits of resource allocation patterns. More simply, policymakers are interested in determining the extent to which altering the allocation of resources, holding constant overall resource levels, can improve performance. 30

The first step in conducting such an analysis for a school district is to devise a method of using the adjusted performance measures to distinguish between higher and lower performing schools. While it might be tempting to use the adjusted performance measures directly, caution is warranted. These measures are estimates and therefore subject to the usual statistical properties. That is, small differences in measured performance are unlikely to be indicative of statistically significant differences in true performance. Even “average” schools are unlikely to have estimated adjusted performance measures that are identically zero. How we proceed depends upon the purpose of the analysis.

Given that most organizations’ observed performance levels will differ from their predicted levels, a threshold or cut-off must be chosen to identify units whose residuals are large enough to confidently label them as organizations whose performance is above or below expected levels. A statistical test for identifying outliers is desirable, but can raise difficult problems depending upon the application. For example, one might wish to determine whether a particular organization has a residual that is significantly different from zero or not. In this case, a confidence interval can be constructed around the estimated regression line, with the actual value of that organization’s output measure falling inside or outside the confidence interval. This method, well established for one prediction, does not solve the problem of determining lower and higher performance for subsets of organizations. 31

For the purpose of understanding resource allocations, we are interested in examining aggregate (rather than individual) patterns across subsets of “high” and “low” performing units. These high and low subsets could be selected by identifying groups of units whose mean residuals differ significantly from zero, or preferably, differ significantly from each other. This latter procedure would be analogous to a t-test for the difference of means between two samples, one with high residuals and one with low residuals. However, these subsets are constructed based on residuals and are
therefore not samples randomly drawn from a population based on an exogenous characteristic. Instead, they are selected based on their outcome (dependent variable) values. Therefore, a standard statistical test cannot be used to identify significantly different mean performance levels of these subsets.

Rather than turning to more arcane statistical tests, we turn to a more intuitive approach to select the high and low performing groups and examine differences in resource allocation patterns. Thresholds can be chosen based on various criteria, such as the size of the residual (for example, schools whose standardized residuals have an absolute value greater than two), or the distribution of the residuals (for example, quartiles). Notice that the choice of a threshold will determine the number of schools in each group and should be selected within the framework of the policy application. For example, if the analyses are intended to identify schools for sanctions or rewards, it may not be feasible or productive to select all schools in the upper and lower quartiles; doing so would identify one-quarter of all schools as high performers and one-quarter as low performers by construction. This is likely to be too many to be useful. Alternatively, if the purpose of the analysis is to identify groups of schools in order to examine similarities within the groups and differences between them, lower thresholds will produce larger subsamples for subsequent analyses. A trade-off between sample size and precision is inevitable, however; lower threshold levels produce larger groups, but also less confidence that the observed values in each group are truly outliers.

An alternative approach would involve the estimation of a second regression in which the performance measures are regressed on the variables describing the allocation patterns. This is, in fact, quite similar to including the allocation patterns in the performance regressions themselves. The two-stage approach adopted here, however, allows us to create multi-purpose performance measures and also to focus on the allocation relationship in isolation. The details of this approach are explained below.

RESOURCE ALLOCATION PATTERNS AND ADJUSTED PERFORMANCE MEASURES IN THE CHICAGO PUBLIC SCHOOLS: AN APPLICATION

In this section, output and budget data from the Chicago Public Schools (CPS) for the 1994–95 school year are used to illustrate an empirical application of the concepts described above. Since 1989, the CPS has been engaged in a wide-ranging decentralization process, shifting control over school management and budgeting from the district office to local schools. With more than 400,000 students in over 550 schools, Chicago is the largest district in the United States to implement school-based budgeting system-wide. As such, it presents an important opportunity to begin to examine the largely untested hypothesis that providing schools and their communities with greater discretion over resource allocation decisions can improve school productivity and performance. The availability of detailed school-level budget and test score data in Chicago facilitates analysis of school-level performance and spending patterns.
Using Existing Data for Performance Measurement

Output data come from the 1994–95 Illinois Goals Assessment Program (IGAP), a series of norm-referenced, state-administered tests given annually in selected grades in reading, mathematics, writing, social studies, and science. Each test is scored on a scale from 0 to 500, with a score of 250 representing average performance. State and district officials use mean school-level IGAP test results to measure school performance, and the scores are widely reported in the local media.

Choosing Independent Variables—Uncontrollable Factors

The analysis includes five independent variables to adjust for “uncontrollable” factors that may affect school-level test scores. The five are: student mobility (MOBILITY), percentage of students from low-income households (LOWINC), percentage of students with limited English proficiency (LEPSCH), total school enrollment (SCENROLL), and parent involvement (PARINV). It is well known that schools serving a more mobile student population, large numbers of students from low-income families or large numbers of students with limited English proficiency are likely to have lower test scores. The equations also include school enrollment to capture any (dis)advantages larger schools have at fostering high levels of student achievement, ceteris paribus. Parent involvement is included because research suggests that schools with high levels of parental involvement may have advantages in promoting student involvement and achievement as compared to schools with limited parental participation. Of course, including the parent involvement variable in this way assumes that schools have little control over the level of involvement among students’ families.

Choosing Independent Variables—Controllable Factors

The CPS school-level budget data report all publicly provided resources available to schools. Total per-pupil spending is included in the adjusted performance measures in order to disentangle spending levels from subsequent analyses of spending patterns. Levels of spending are largely determined outside each school, by the district, state, and to some degree, the federal government. Alternative specifications of the model include pupil-teacher ratios in place of per-pupil expenditures and produce similar results.

Choosing Independent Variables—Controlling for Initial Conditions

While achievement gains cannot be directly measured with the available data (which are for one year), a “pre-test” approach can be approximated using data from the grade below the one in which the output test is administered. The schedule for administering
the IGAP tests raises several difficulties for assessing performance using a pre-test approach. First, same subject-area tests are administered at intervals of two or three grade levels. Therefore, it is not possible to examine differences in student performance in specific subject areas for the same cohort of students from one year to the next. In principle, achievement gains could be measured across several years (for example, between third and sixth grade, or between sixth and eighth grade). Unfortunately, only school-wide averages are available in Chicago, not student-level performance data. High student mobility in Chicago makes changes in test scores across several years an unreliable indicator of school performance.43

To approximate a pre-test approach, third grade scores can be used as independent variables in models to predict fourth grade scores from the same year. This approach is not ideal, but to the extent that the current third grade strongly resembles the previous third grade (the current year's fourth grade) in characteristics and performance, prior student performance will be controlled for. Of course, if the school is steadily improving (deteriorating) over time at all grades, this method will systematically underestimate (overestimate) performance gains.

Chicago schools administer reading and mathematics tests—the most commonly used indicators of student performance—in the years prior to administration of the social studies and science examinations. Therefore, social studies and science scores are the only scores available for use as dependent variables in models with previous grade scores, and they are used here.

**Performance Equations And Results**

Ordinary Least Squares (OLS) regressions with linear functional forms are estimated using fourth grade social studies IGAP scores to measure performance. Table 1 lists means and standard deviations for the model variables. Although, there are 434 schools reporting fourth grade scores in Chicago, only 426 report both third and fourth grade scores.

Table 2 shows the parameter estimates from four specifications of the adjusted performance models. Columns (1) and (3) use per-pupil expenditures as resource variables; columns (2) and (4) use pupil-teacher ratios. Columns (1) and (2) report results using third grade "pre-test" scores. In all specifications, the percentage of students from low-income families has a significant, negative effect on school outputs. As expected, third grade scores have a significant positive impact on fourth grade scores. Student mobility negatively affects test scores while higher student enrollments produce lower scores in all but one model. Note that the models are insensitive to the measure of resource use employed.

The choice between these models for further analysis should depend upon the intended application. For our purposes, models including pre-test measures are most attractive because they show the greatest predictive power (highest $R^2$) and will therefore be more accurate in identifying schools producing higher than expected gains in
TABLE 1
Descriptive Statistics of Regression Variables
Chicago Public Elementary Schools

<table>
<thead>
<tr>
<th>Variable</th>
<th>With Pre-tests</th>
<th>Without Pre-tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>(SD)</td>
<td>(SD)</td>
</tr>
<tr>
<td>Limited English Proficient (%)</td>
<td>13.82</td>
<td>13.77</td>
</tr>
<tr>
<td>Students (%)</td>
<td>(17.51)</td>
<td>(17.53)</td>
</tr>
<tr>
<td>Low-Income Students (%)</td>
<td>80.80</td>
<td>80.96</td>
</tr>
<tr>
<td>LOWINC (%)</td>
<td>(20.09)</td>
<td>(19.96)</td>
</tr>
<tr>
<td>Student Mobility (%)</td>
<td>30.40</td>
<td>30.40</td>
</tr>
<tr>
<td>MOBILITY (%)</td>
<td>(16.54)</td>
<td>(16.46)</td>
</tr>
<tr>
<td>Parent Involvement</td>
<td>93.49</td>
<td>93.43</td>
</tr>
<tr>
<td>PARINV (%)</td>
<td>(13.36)</td>
<td>(13.43)</td>
</tr>
<tr>
<td>School Enrollment (number)</td>
<td>664.62</td>
<td>662.58</td>
</tr>
<tr>
<td>SCHENROLL (%)</td>
<td>(275.07)</td>
<td>(274.39)</td>
</tr>
<tr>
<td>Total Budget Per Pupil (BUDPUP)</td>
<td>5,271.78</td>
<td>5,288.42</td>
</tr>
<tr>
<td>(1,215.03)</td>
<td>(1,122.16)</td>
<td></td>
</tr>
<tr>
<td>Pupil-Teacher Ratio</td>
<td>16.60</td>
<td>16.58</td>
</tr>
<tr>
<td>PUPTCH (%)</td>
<td>(2.51)</td>
<td>(2.51)</td>
</tr>
<tr>
<td>Fourth Grade Social Studies Scores (SOSC4SC)</td>
<td>174.21</td>
<td>173.68</td>
</tr>
<tr>
<td>Third Grade Math Score (MATH3SC)</td>
<td>206.74</td>
<td>(54.47)</td>
</tr>
<tr>
<td>Third Grade Reading Score (READ3SC)</td>
<td>179.69</td>
<td>(54.32)</td>
</tr>
<tr>
<td></td>
<td>(53.55)</td>
<td></td>
</tr>
</tbody>
</table>

student achievement. We choose model (1) for further analysis because of the appeal of including the expenditure variable.

To proceed, standardized residuals are calculated to measure the difference between each school’s actual output and that predicted by the model. By standardizing the unit of measure, analysts can compare and aggregate residuals across multiple output measures with different scales.

There is no single correct way to analyze the residuals to identify high and low performing groups of schools. As discussed previously, a threshold must be chosen to determine which schools will form each of these groups. The purpose of the analyses presented here is not to assess individual school performance, but rather to examine differences in resource allocation patterns across groups of high and low performing schools. Therefore, thresholds should be selected to ensure that the subsamples are
### TABLE 2
Results of Adjusted Performance Measurement OLS Regressions

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>With Pre-tests (st. error in parentheses)</th>
<th>Without Pre-tests (st. error in parentheses)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Constant</td>
<td>109.24** (17.94)</td>
<td>102.80** (18.10)</td>
</tr>
<tr>
<td>LEPSCH</td>
<td>.100 (.091)</td>
<td>.103 (.090)</td>
</tr>
<tr>
<td>LOWINC</td>
<td>-.527** (.093)</td>
<td>-.525** (.093)</td>
</tr>
<tr>
<td>MOBILITY</td>
<td>-.148 (.094)</td>
<td>-.145 (.094)</td>
</tr>
<tr>
<td>PARINV</td>
<td>-.048 (.103)</td>
<td>-.048 (.103)</td>
</tr>
<tr>
<td>SCENROLL</td>
<td>-.015** (.006)</td>
<td>-.015** (.006)</td>
</tr>
<tr>
<td>BUDPUP</td>
<td>.000 (.001)</td>
<td>-.001 (.001)</td>
</tr>
<tr>
<td>PUPTCH</td>
<td>.223 (.593)</td>
<td>1.015 (.804)</td>
</tr>
<tr>
<td>MATH3SC</td>
<td>.123* (.054)</td>
<td>.124* (.054)</td>
</tr>
<tr>
<td>READ3SC</td>
<td>.569** (.066)</td>
<td>.568** (.066)</td>
</tr>
<tr>
<td>N</td>
<td>426</td>
<td>426</td>
</tr>
<tr>
<td>F</td>
<td>160.92**</td>
<td>160.92**</td>
</tr>
<tr>
<td>R²</td>
<td>.755</td>
<td>.755</td>
</tr>
</tbody>
</table>

** Significant at the .01 level.
* Significant at the .05 level.

The results indicate that the regression model is effective in predicting performance outcomes. The coefficients for various independent variables, such as LEPSCH, LOWINC, MOBILITY, and PARINV, show significant relationships, suggesting that they are important factors in determining performance. The R² values indicate a moderate to strong explanatory power of the model.

### Analysis

Large enough to facilitate the analysis of allocation patterns. We classify all schools with standardized residuals higher than 1 as high performing schools, and those with negative residuals larger than -1 as low performing schools, producing subsamples of forty-nine and forty-six schools respectively.

Table 3 lists the mean values for the groups of high and low performing schools on each of the independent and dependent variables in the performance measurement equations, and for several resource allocation variables. As expected, few differences...
### TABLE 3
**Difference of Means—High and Low Performing Schools**
(Standardized Residual > |1|)

<table>
<thead>
<tr>
<th>Variables in Regression</th>
<th>Low</th>
<th>High</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N = 46</td>
<td>N = 49</td>
<td></td>
</tr>
<tr>
<td>Pct. Limited English Proficient Students (LEPSCH)</td>
<td>10.75</td>
<td>10.89</td>
<td>-0.14</td>
</tr>
<tr>
<td>Pct. Low-Income Students (LOWINC)</td>
<td>73.01</td>
<td>78.10</td>
<td>-5.09</td>
</tr>
<tr>
<td>Student Mobility (pct.) (MOBILITY)</td>
<td>28.69</td>
<td>30.76</td>
<td>-2.07</td>
</tr>
<tr>
<td>Pct. Parent Involvement (PARINV)</td>
<td>92.07</td>
<td>93.06</td>
<td>-0.99</td>
</tr>
<tr>
<td>School Enrollment (SCENROLL)</td>
<td>620</td>
<td>597</td>
<td>23</td>
</tr>
<tr>
<td>Total Budget Per Pupil (BUDPUP)</td>
<td>5794</td>
<td>5527</td>
<td>267</td>
</tr>
<tr>
<td>Pupil-Teacher Ratio (PUPTCH)</td>
<td>15.68</td>
<td>16.19</td>
<td>-0.51</td>
</tr>
<tr>
<td>Third Grade Reading Score (READ3SC)</td>
<td>194.35</td>
<td>195.61</td>
<td>-1.26</td>
</tr>
<tr>
<td>Third Grade Math Score (MATH3SC)</td>
<td>220.43</td>
<td>219.61</td>
<td>.82</td>
</tr>
<tr>
<td>Fourth Grade Social Studies Score (SOSC4SC)</td>
<td>143.89</td>
<td>234.33</td>
<td>-90.44</td>
</tr>
</tbody>
</table>

**Patterns of Total Expenditures**

| Pct. Spent on Instruction | 60.73 | 62.01 | -1.28 |
| Pct. Spent on Instructional Support | 12.77 | 10.27 | 2.50  |
| Pct. Spent on Administration | 6.11  | 6.30  | -0.19 |
| Pct. Spent on Operations    | 18.66 | 19.74 | -1.08 |

**Patterns of Chapter One Expenditures**

| Pct. Chapter One on Instruction | 61.97 | 66.93 | -4.96 |
| Pct. Chapter One on Instructional Support | 20.10 | 16.95 | 3.16  |
| Pct. Chapter One Administration | 14.52 | 12.84 | 1.68  |
| Pct. Chapter One on Operations   | 2.41  | 2.77  | -0.36 |

In the independent variables emerge between the groups. Surprisingly, the high performing schools have a higher average rate of student poverty and higher student mobility. Low performing schools receive higher average levels of funding. While the two groups have virtually identical mean scores on the “pre-tests” (third grade reading
and math), the mean fourth grade score is ninety points (63%) higher in the high performing than the low performing group. That is, high performing schools produced substantially larger achievement gains between third and fourth grades.

Comparison of Spending Patterns

Using the CPS budget data, six functional categories are created to explore school-level spending patterns.\textsuperscript{46} Table 3 shows the average proportion of total school spending, and of Illinois State Chapter 1 spending, is allocated to each of the four largest functions across high and low performing schools.\textsuperscript{47} For total spending, the differences are small. Low performing schools spend an average of 2.5 percentage points more on instructional support while high performing schools spend slightly more in the other three areas.

These relatively minor spending differences may reflect the limited discretion that even schools in Chicago have over their full budgets. Focusing on discretionary money may yield more insight. State Chapter 1 funding is the largest source of such “discretionary” money under the control of school-level personnel, representing approximately 12 percent of the average elementary school’s budget. The bottom panel of Table 3 shows greater differences in the distribution of discretionary spending across functions than are shown in total spending. This is consistent with the belief of many proponents of school-based budgeting that fiscal decentralization will result in more diverse spending patterns as schools target their resources to meet individual school needs.

Notice also that high performing schools average almost five percentage points more discretionary spending on instruction and less on instructional support and administration. Although the cross-sectional nature of the data limit causal inferences, the results do provide some indication that schools producing higher student outputs tend to target a larger share of their discretionary resources toward direct instructional expenditures. Of course, without more detailed analysis of individual school needs in relation to their spending patterns, we should not conclude that low performing schools budget their resources inappropriately.\textsuperscript{48}

CONCLUSIONS

Despite the difficulties inherent in measuring the performance of government and nonprofit organizations, such analyses hold great promise for improving accountability, budgeting, incentive programs, and resource allocation decisions. Since performance assessment in education has advanced beyond similar efforts in many other areas, lessons learned and problems encountered in the measurement of education performance should be of interest to those involved in other programs.

What are some of the lessons?

First, as our analysis of public schools shows, it is not necessary to be satisfied with input measures of performance, such as days of instruction or miles of highway...
plowed, due to unwarranted fear that only input measures are under the control of public managers. Instead, risk adjusted output measures allow the use of metrics closer to the real goals of programs and can still afford reasonable accountability of managers by controlling for factors outside the organization’s influence. Those who are responsible for a given organization can then meaningfully evaluate its performance on education tests, lives saved due to highway safety rules, employment of clients, health outcomes and so forth, if these indicators are adjusted for factors that the organization cannot be expected to influence.

Second, there are a variety of ways to present the results of statistically adjusted performance measures so that the informed public can understand them. As we show in this article, organizations can be divided into high and low performing groups (or into quartiles or any other equal groupings) based on “behind-the-scenes” analyses of residuals. Alternatively, the unadjusted and the predicted performance measures can be presented for a particular organization, with the factors used to predict listed below the measures. The concept of adjusted measures is intuitive and the results of the analyses need not be presented in equation form.

What are some of the incompletely resolved issues surrounding the use of adjusted performance measures?

First, there are a variety of ways to link resource usage to adjusted performance measures. In this article, we have illustrated a two-step process that involves calculating the adjusted measures and then computing averages for resource use by level of performance. This two-step process is accessible to a wide audience, but it is less elegant than one that combines the resource usage information with the adjusted performance measure calculation. A combined method, however, puts us right in the middle of the current literature on production functions in public organizations, a literature that is decidedly inconclusive about the right way to formulate such production functions. Additional work in the area of combining adjusted measures and budget information is clearly warranted.

Second, a “one-size-fits-all” approach is not appropriate for analyzing organizational performance. Analysts face numerous decisions regarding both the choice of performance indicators and the choice of risk factors that statistics alone cannot resolve. Instead, the users themselves will have to agree upon those measures they consider meaningful. Furthermore, since risk factors may be controllable at one level of an organization but uncontrollable at another, these choices will also need to be addressed.

In addition, we have illustrated the use of adjusted performance measures in resource allocation evaluations only. Other uses of these measures, such as to inform individuals choosing an organization to serve their particular needs, or to guide legislators or executives in distributing monetary rewards for performance, present additional challenges. As an example, these will require more attention to the particular outputs, more sophisticated ways to identify the appropriate risk factors for each organization, and experimentation with the sensitivity of results to alternative specifications.
The use of performance measurement is sweeping public and nonprofit organizations both in the United States and abroad. Finding ways to use measures that track outputs instead of inputs and methods to adjust for factors outside an organization’s control is desirable. Linking adjusted measures to resource use is essential if performance budgeting is to have its intended impact. As we illustrate in this article, the progress being made in K–12 public education provides helpful insights for the future of performance measurement in other program areas.

NOTES

1. Other explanations have captured the public interest at different times. For example, during the 1980s there was considerable discussion about the productivity slowdown in the private sector, a slowdown that was blamed on insufficient investment in infrastructure, R&D, and human capital among American workers. The solutions to the productivity problem focused on increasing government spending on infrastructure, R&D and education.


6. The intention to gather fiscal data to match with outcome data for performance analyses is often explicit, as illustrated by the following excerpt from New York City’s School Based Budget Reports, Fiscal Year 1995–96 (Brooklyn: Board of Education of the City of New York, November 1996): In a preface no note, Chancellor Rudolph Crew asserts: “The result of a year of hard work is in these School Based Budget Reports, which give us a framework to begin assessing how resource decisions relate to student outcomes.”


15. Average scores on standardized tests such as the Scholastic Assessment Test (SAT) might be viewed as an “intermediate outcome.”


17. More specifically, if \( P_i \) represents the performance measure for organization \( i \) and \( Z_i \) is the set of uncontrollable variables affecting \( i \)’s performance, then estimate:

\[
P_i = \beta_0 \beta Z_i + \varepsilon_i
\]

such that \( i = 1, 2, \ldots, N \).

where \( N \) is the number of organizations in the data set, \( \beta_0 \) is the intercept of the equation, \( \beta \) is the vector of coefficients capturing the marginal impacts of changes in the \( Z_i \)’s on \( P_i \), and \( \varepsilon_i \) is an error term with the usual properties. Denoting estimates of \( \beta_0 \) and \( \beta \) as \( \hat{\beta}_0 \) and \( \hat{\beta} \), respectively, the adjusted performance measure for organization \( i \), \( M_i \), is found as:

\[
M_i = P_i - \hat{\beta}_0 - \hat{\beta} Z_i
\]

for all \( i = 1, 2, \ldots, N \). Thus, \( M_i \) is the estimated residual (or the prediction error) from the estimation of (1).

18. Letting \( \gamma \) represent the vector of coefficients on \( X \), (1) then would be specified as:

\[
P_i = \hat{\beta}_0 + \hat{\beta} Z_i + \gamma X_i + \varepsilon_i
\]

and, letting \( g \) represent the estimate of \( \gamma \), \( M_i \) can be measured as:

\[
M_i = P_i - \hat{\beta}_0 - \hat{\beta} Z_i - g X_i
\]


21. See Hatry et. al., *Service Efforts and Accomplishments*, for illustrations of data development needs in eleven areas.
22. Although this paper focuses on adjusted measures, there is a case for excluding factors that are not controllable under prevailing circumstances, because their inclusion may lead to lower standards for organizations serving vulnerable populations. This difficulty surfaces in education over the possibility that schools having larger percentages of at-risk children will be held to lower standards. Use of growth rates or changes in performance measures, rather than the levels themselves, may ameliorate this problem, as discussed in greater detail below.

23. To see this, note that if the coefficient on a variable like “percent non-white students” is negative, other variables equal, expected performance will be lower for a school with a higher percentage of non-white students.


25. This will matter less, however, when using value-added performance measures, described below, and/or the resources change little between the years.

26. To see how these are related, re-write (3) and (4) adding a subscript ‘t’ to indicate time period and add \( P_{it} \) to the list of explanatory variables with \( \delta \) representing its coefficient.

\( P_{it} = \beta_0 + \beta_Z + \gamma Y_{it} + \delta P_{it-1} + \epsilon_i \)

and, letting \( d \) represent the estimates of \( \delta \), \( M_t \) can be measured as:

\( M_t = P_{it-1} + b \cdot Z_{it} - g X_t - d \cdot P_{it-1} \)

The alternative “value-added” specification can be written:

\( \Delta P_{it} = \beta_0 + \beta Z + \gamma Y_{it} + \epsilon_i \)

where \( \Delta P_{it} \) is \( P_{it} - P_{it-1} \), \( M_t \) then becomes:

\( M_t = \Delta P_{it} - b \cdot Z_{it} - g X_t \)

Notice that estimated coefficients will be the same under (5) and (7) if \( \delta \) is constrained to be equal to one.


30. If the performance measures are meant to motivate performance, then some way of motivating all units should be fashioned. This may involve foregoing a declaration that, based on the size and sign of their residuals, some are high performing and some are low performing. See Behn, “Linking Measurement and Motivation,” for further discussion of these issues. If the purpose of the measures is to report to stakeholders, then a distribution of residuals from high to low may work.


32. Assuming a normal distribution, standardized residuals will have a mean of approximately zero and a standard deviation of approximately one. Therefore, selecting a threshold standardized residual of \( |z| \), for example, can be expected to eliminate approximately 95 percent of the units from the sample.


34. Students take the reading, mathematics, and writing examinations in third, sixth, eighth and tenth grades. Students take the social studies and science examinations in fourth, seventh, and eleventh grades.
35. The writing exam is scored on a scale ranging from six to thirty-two.
37. Student mobility is measured as the number of students moving into and out of the school during the year, divided by the October enrollment.
38. Parent involvement is measured as the percentage of students whose family made at least one contact with the school during the year.
39. As public elementary schools, the schools have little opportunity to select their students. Instead, students generally choose the schools they will attend or are assigned to the schools on the basis of family residential location.
41. This belief was fundamental in shaping Chicago School Reform, which mandated a parental role in schools through the Local School Councils. See Ann Henderson, ed., The Evidence Continues to Grow: Parent Involvement Improves Student Achievement (Columbia, MD: National Committee for Citizens in Education, 1987) for a review of research on the effects of parent involvement.
42. The total per-pupil budget includes allocations from local, state, and federal sources.
43. As David Kerbow, Pervasive Student Mobility: A Moving Target for School Improvement (Chicago: Chicago Panel on School Policy, 1995) shows, high student mobility in Chicago results in classes that may have fewer than fifty percent of the same students from one year to the next, and even fewer across multiple years.
44. The residual values are relatively insensitive to the specification of the model. The Pearson correlation between residuals from the model with pre-tests and the model without pre-tests is .73 (p < .01).
45. While the residuals’ units of measure may be standardized in other ways (for example, by studentizing), the intuitive appeal and relatively simple calculation of the standardized residual (the simple residual divided by its estimated standard deviation) broadens its appeal for policy applications.
46. The six functional categories are: instruction (expenditures, including teacher salaries and benefits, associated with direct instruction of students in classrooms); instructional support (expenditures for providing teacher support, such as professional development, student support, such as guidance and health services, and program support, such as library and media services); administration (including expenditures for principals, assistant principals, attendance services and security); operations (non-instructional expenditures associated with maintenance of the school building and lunchroom services); community services (primarily community and parent outreach activities); and other (a small catch-all category that primarily includes expenditures for debt service).
47. Illinois’ Chapter 1 program, like the federal Title I program, distributes “compensatory” aid to schools (through their districts) based upon enrollment of students from low-income families.
48. Low performing schools may, for example, tend to have greater needs in non-instructional areas due to greater student or physical plant requirements.

Stiefel, Rubenstein & Schwartz / Adjusted Performance Measures for Evaluating Resource Use 87