

Draft Copy: Please do not cite without permission. Comments are very welcome.

IDENTIFYING LOW-PERFORMANCE SCHOOLS: WITH APPLICATION TO NEW YORK CITY

Robert Bifulco
William Duncombe

Center for Policy Research
The Maxwell School
426 Eggers Hall
Syracuse, NY 13244-1020
(315) 443-3114
Duncombe@maxwell.syr.edu
Rbifulco@maxwell.syr.edu

1998 Annual Conference
Association of Budgeting and Financial Management
Washington, DC
November 5-8, 1998

IDENTIFYING LOW-PERFORMANCE SCHOOLS: WITH APPLICATION TO NEW YORK CITY

The growing emphasis on performance in public organizations has reached public schools. In contrast to past attempts to reform process and government, present efforts to reform American schools have focused more directly on performance. Key features of these performance-based reforms in schools as well as other public organizations include: establishing clear, measurable performance standards; granting local actors the autonomy to find the best means of achieving these standards; providing rewards for local actors that achieve performance goals; and developing remedies for cases when goals are not met (King and Mathers, 1997). While receiving less media attention, an important element in performance-based reforms is the identification and improvement of organizations with the lowest performance.

In 1991, the New York State Board of Regents issued *A Compact for Learning*, a statement of principles intended to guide school reform efforts in New York State. In many ways this document, and recent reform efforts in New York State, exemplify the performance-based reform movement. An integral part of New York State's reform efforts is the New York State Education Department's Registration Review program. This program establishes a process for identifying low-performance schools as well as a series of interventions intended to improve these schools. The process for identifying low-performance schools relies primarily on student performance on state-wide tests. Program regulations require schools placed under Registration Review to participate in a school improvement process and provide for a series of additional measures, including school closure, if school improvement efforts are not successful.

New York's efforts to identify and intervene in low-performance schools are not unique. At least 18 other states have established procedures for identifying low-performance schools or districts. Legislation and regulations in these states specify a number of actions ranging from visits by technical assistance teams

to state-takeover (Advisory Council to New York State Board of Regents, 1994). The 1994 reauthorization of the federal Title I program encouraged more programs of this sort by requiring states and districts that receive Title I money to establish systems, based primarily on student achievement on state-wide assessments, for identifying and improving low-performance schools.

Low-performance school programs, like New York's Registration Review, face two controversial and difficult sets of issues. The first set of issues concerns how to identify which schools should be classified as having low student performance. Although not without critics, the idea that low-performance schools should be identified primarily on the basis of student outcomes is widely accepted. However, several issues concerning how student outcomes can be used to determine school performance remain unresolved. The second challenge facing these programs is to evaluate the causes of low-performance and determine what actions should be taken to improve student achievement in the schools identified. Most attempts to improve low-performance schools have assumed that management failures lie at the heart of the problem, and thus that administrative changes are in order. However, there is little empirical evidence to support the supposition of management failure.

In this paper, we will address the first set of issues--how do state governments identify schools with the lowest student performance? At first glance, the task of identifying low-performance schools would appear quite straight-forward. Simply find the schools with the worst test scores. However, there are several important issues involved in identifying low-performance schools. In the first section of this paper we outline and provide a general discussion of these issues. In the second section, we examine New York's Registration Review process. Using test data for New York City elementary schools, we demonstrate that the method used by New York to identify low-performance schools is sensitive to the cohort of students examined, the tests included in the evaluation and choices concerning the aggregation of multiple sets of test

results. In the third part of the paper, we lay out an approach for identifying low performance schools that is potentially more reliable and stable than those currently in use. We conclude by discussing the implications that our findings have for the design of low-performance school programs and directions for future research.

I. Issues Involved in Identifying and Intervening in Low-Performance Schools

Once it is decided to evaluate school performance based on the outcomes of students in the school, a number of issues need to be resolved. Specifically, officials must answer three questions: (1) what student outcome measures should be used; (2) should identification be based on absolute levels of student achievement or a schools' relative contribution to student learning; and (3) how should student outcome measures be aggregated into school-wide outcome measures?

What outcome measures should be used?

The choice of outcome measures is of paramount importance. Low-performance school programs involve ranking of schools on some student performance measure, and designation as a low-performance school often leads to public or state pressure for improvement. Improvement efforts are likely to be focused on those outcomes that are used to identify low-performance schools. If these outcomes are not important ones, then valuable resources will be misspent. Thus, it is crucial that the student outcome measures used to evaluate schools reflect important school improvement goals (King and Mathers, 1997).

Norm-referenced and statewide criterion-referenced test scores, particularly in reading and math, are commonly used to identify low-performing schools.¹ While the reliability of many norm- and criterion-referenced tests is well established, the use of statewide tests to evaluate school performance has been criticized on other grounds. Standardized statewide tests are not always aligned with curricular goals.

Subjects such as science, social studies and the arts are often not tested, and even in the tested subjects higher order thinking and problem solving skills are often not assessed . There are also concerns that the development of higher order thinking skills is undermined when emphasis on improving test results leads teachers to "teach-to-the-test" (Darling-Hammond, 1991).² The move to student portfolios and other "authentic" assessment devices are attempts to correct the perceived limitations of standardized tests. Unfortunately, more comprehensive assessment instruments often do not provide measures of performance that are comparable across schools.

Besides standardized tests, student attendance and dropout rates are also commonly used in assessing school performance.³ Other outcome measures that are used include grade promotions, suspension/incident rates, college placements and job placements (Schumacker and Brookshire, 1992). Multiple outcome measures provide more information about student performance, but also complicate the identification of low-performance schools.

Should identification be based on absolute levels of achievement or a school's relative contribution?

The answer to this question depends crucially on the objectives of the program, and, ultimately, on the unit of analysis. If this program is directed at children and their future success, then absolute levels of student achievement should be used. Success in college, and to a growing extent in the job market, depends on the acquisition of certain cognitive skills in high school. Acquisition of these skills in high school depends, in turn, on a basic level of achievement in the elementary grades. If a large portion of the students in a school are not developing the skills needed to succeed, then the educational program in that school needs to be improved. Whether or not the employees in that school are doing as well as employees in other schools might do under similar conditions is not the issue. Regardless of the reasons why students are not developing the required skills, action needs to be taken to ensure that these student are given the opportunity

to succeed in high school, college and the job market. Such a program is more accurately called a *low-performance school* program since the performance of students is the focus of the program.

If the objective of the program is to help schools improve their efficiency (regardless of the level of student performance), then student achievement gains are more relevant than absolute levels of student achievement. The commonly used term, *low-performing school* program, more appropriately describes a program with an efficiency focus. If the program is primarily intended to provide incentives to poor performing school officials to change their behavior, then school performance measures should reflect factors under the control of school officials (Meyer, 1996). A program that holds school officials accountable for student performance levels that are unlikely to be achieved in the short-run, regardless of how the school's effectiveness, will not provide much incentive for those officials to change their behavior. Such a program might only serve to demoralize school members or encourage fraudulent reporting practices (Darling-Hammond, 1991; King and Mathers, 1997). Because a school has more control over how much a student's achievement improves while she is at the school than over the absolute level of performance, a measure of student gains is preferred for assessing school performance.

Using student gains on test scores, rather than absolute levels of performance, only partially addresses this issue. If disadvantaged students are likely to show smaller gains in achievement from year-to-year as well as lower absolute levels of achievement, then even using a measure of student gains to assess school performance will still bias identification toward schools with high percentages of disadvantaged students. This concern can be addressed in a number of ways. Until recently, South Carolina grouped schools in comparison bands based on various predictors of pupil achievement (King and Mathers, 1997; Clotfelter and Ladd, 1996). Thus, schools were compared only to schools facing similar environments. In Texas, schools are assessed based on the performance of students in four separate groups including

economically disadvantaged students. The economically disadvantaged students in a school must show acceptable levels of performance regardless the percentage of disadvantaged students in the school (King and Mathers, 1997). The Dallas Independent School District uses multiple regression analyses to predict the average student gain in each school. The performance of each school in the district is then assessed relative to its predicted score (Webster and Mendro, 1995).⁴

How should student outcome measures be aggregated into school-wide outcome measures?

Whenever individual test scores are aggregated into a measure of school performance, information is lost. Three aggregation issues are of particular importance. First, what part of the student performance distribution should be the focus of the measure? Of particular concern is the fact that mean test scores can mask low achievement by groups of students within the school. The scores of high-achieving students can skew results and create the impression that all students in the school are performing effectively (Darling-Hammond, 1991; King and Mathers, 1997). New York State avoids concerns about using mean test scores by using the percentage of students in a school achieving a standard of minimum competency. This helps to focus attention on the amount of students not receiving the level of service they need to prepare them for future endeavors. However, this measure also masks important information. It provides no information on whether the students in a school are developing higher order thinking skills. In addition, it provides little information for schools with small numbers of disadvantaged students who are, in any case, unlikely to show significant numbers of students falling below minimum competency. Some states have moved beyond single indicators by providing performance measures for different parts of the distribution and/or different socio-economic groups.⁵

A second aggregation issue may arise if there is performance variation across different cohorts of students. In general, we would expect that the performance of third graders in one school in one year should

be similar to third grade performance in the following year, i.e., no cohort effect. However, if there are noticeable performance differences across student cohorts, then a school's performance ranking may be partially determined by the cohort of students included in the evaluation. Under circumstances of significant cohort effects, some method for combining results across several cohorts may need to be developed.

The third aggregation issue is how different types of tests can be combined into one school-wide index. For K-6 elementary schools, New York State uses four different student outcome measures to identify schools for Registration Review. If a school's results on any one of these four measures meet the identification criteria, the school can be identified for Registration Review. Thus, if a school shows a low enough percentage above minimum competency on the Grade 3 Math test, it will be identified for Registration Review regardless of how well its students perform on the other tests. An alternative approach is that used by Dallas, where a set of externally generated weights are used to combine different exams into one school-wide index (Webster, Mendro and Almaguer, 1994).

An argument can be made for New York State's approach. If students are to succeed in high school and beyond they need to develop a certain level of competency in several areas, particularly in math and reading. If a large number of students in a school are not developing adequate math skills, this is a problem that needs to be addressed regardless of how well the school's students are reading. Using an aggregate measure that combines math and reading results can mask problems in one of these areas. However, if a larger number of school performance indicators are used (as many recommend), then New York State's approach becomes more problematic because fewer and fewer schools are identified by each test (for a fixed total number schools to be identified).⁶

Ideally, we would like to develop identification methods that are not overly sensitive to small changes in the evaluation instruments used in the assessment, the particular cohort of students that is

examined or specific decisions about how to combine multiple sets of test results. Student performance is multi-dimensional. Schools with low-student performance in one dimension might not be the same as schools with low-student performance on another dimension. Consequently, methods of identifying low-performance that focus on different outcomes might be expected to identify different schools. However, once a decision is made about which student outcomes are most important, the set of schools that a method identifies should not vary substantially with decisions about what instruments and student cohorts to use in assessing those outcomes. Similarly, once it is agreed to focus on math and reading performance, for instance, specific choices about how to combine various indicators of math and reading performance should not substantially affect the group of schools that are identified as low-performance schools.

II. Identifying Low-Performance Schools in New York State

To provide a concrete example of how a state classifies low-performance schools, we will discuss the Registration Review program in New York. We will then use test data on New York City elementary schools to determine how sensitive the Registration Review identification system is to student cohort effects and to choices about how to combine results from multiple tests. Under the original Registration Review system in New York, a school was placed on the list if it scored below an established criterion on one or more measures of school performance, and student achievement on that measure had not shown improvement over the preceding three years. For elementary schools, four measures of school performance were used. These were the percentage scoring above a minimum competency level on statewide pupil evaluation exams (PEP tests) in third and sixth grade math and reading. If the percentage of students in a school scoring above minimum competency on any one of these four tests was below a certain level and declining over a three-year period, the school was identified.

In 1994, the Regent's Advisory Council on Low-Performing Schools (1994), in a report provocatively titled *Perform or Perish*, recommended revision of the Registration Review identification process to ensure identification of all schools that perform far below standards. In 1996, the Board of Regents adopted revised regulations for the Registration Review program. These regulations require that the State Education Department identify for Registration Review those schools that are "farthest from meeting State standards" rather than those that are merely below a State standard and showing a three-year pattern of decline. The standards are still formulated in terms of the percent of students in the school demonstrating minimum competency on State tests. Thus, the new regulations require the identification of those schools with the lowest percentages of students achieving minimum competency.

By deciding to identify schools that have the lowest percent of students scoring above a minimum competency level on specific state exams, New York State has made decisions on several issues. Particularly, the Board of Regents has decided to focus on: (1) outcomes on statewide, criterion reference tests in math and reading; (2) absolute levels of student performance rather than the contribution of schools to student learning; and (3) the lower end of the student achievement distribution.

Nevertheless, there are questions that still need to be addressed concerning the definition of "farthest from meeting State standards." Specifically, how should year to year changes in measures of math and reading performance be handled and how should the various outcome measures be combined to identify a single set of low performing schools. As a matter of practice, the State tries to avoid questions about aggregating across outcome measures by applying each outcome measure separately. It also tries to ignore questions about how to weight results across years by relying solely on results from the most recent year. However, these practices merely represent implicit decisions about issues that should be addressed

explicitly. In this section we attempt to assess these issues and determine how much difference they make for what schools end up being identified as “farthest from meeting State standards.”

The Data and Analysis

To conduct our study, we used data provided by the New York City Board of Education. The data is drawn from files used by the Board to produce annual citywide school report cards. The files provide school level data on standardized test results for the years 1994, 1995 and 1996. We limited our study to elementary schools. After eliminating schools whose students do not participate in the statewide test program as well as schools with missing observations, our final data set included 603 schools. Of these 603 schools, 274 include a sixth grade. For these schools we have results from the math and reading PEP tests administered in grades 3 and 6. For the other 329 schools, we have results from the grade 3 PEP tests.⁷

Our question is whether decisions about how to combine test results across multiple years and multiple outcome measures make a difference for which schools end up being identified as “farthest from meeting state standards.” To answer this question we constructed several different methods for identifying low-performance schools and compared the list of schools generated by these methods. Each method can be interpreted as identifying the schools that are “farthest from state standards”. Our base-case method is similar to that currently used by the New York State Education Department. Using data from the most recent year only, this method identifies a school if the percentage above minimum competency in 1996 on any of the PEP tests administered in the school ranked among the lowest 10 schools. After generating a list of low-performing schools using this base-case method, we generated several alternative lists using slightly modified methods. A summary of the different methods used is provided in Figure 1.

New York State’s approach to aggregating across multiple outcome measures (our base-case method) has two important elements. The first is the set of criteria, or cut-points, that are chosen for each

test. In the base-case, ranking among the lowest 10 schools is the criterion chosen for each test (see first row, first column of Figure 1). The choice of this cut-point depends partially on the total number of schools to be identified. However, there is no clear reason why the cut-point needs to be the same for each test. Thus, one of the ways we modified the base-case method is by changing the criterion for each test. Particularly, we applied a method that identifies a school if the percentage above minimum competency on either of the third grade PEP tests ranked among the lowest 15 schools, or the percentage above minimum competency on either of the sixth grade tests ranked among the lowest 5 (second row, first column). In a more substantial modification we eliminated use of the sixth grade test results altogether, and identified schools that were among the bottom 20 schools on either of the third grade PEP tests (third row, first column). Comparing the lists generated by these modified methods to the list generated by the base-case method indicates how sensitive New York State's approach to aggregating across multiple outcome measures is to the choice of cut-points.

The second important element of New York State's approach is the rule used to combine the sets of schools identified by each criterion. In our base-case, the combination rule is the logical operator "or". A school is identified as a low-performance school if it is identified by any one of the criteria. Another possible combination rule is the logical operator "and". This combination rule stipulates that a school is identified as a low-performance school only if it is identified by each and every criterion. We applied a modified identification method based on this alternative combination rule (fourth row, first column). Comparing the schools identified by this modified method to the base-case list indicates how sensitive New York State's aggregation method is to the choice of combination rule.

An alternative to New York State's approach to aggregating across multiple outcome measures is an averaging approach. One version of this approach that we applied uses the simple, unweighted average

of the percentages of students below basic competency on each test as an aggregate measure of school performance, and then identifies those schools that rank among the lowest on this aggregate measure (fifth row, first column). One question concerning this approach is how different is the list of schools it identifies from the list of schools identified by New York State's method. A second question is how sensitive is the averaging approach to the weights chosen for each outcome. To examine this second question, we applied methods that made use of a weighted average of the test results that weight reading results more heavily than the results on the mathematics tests (sixth and seventh rows, first column).

Finally, for each identification method we applied, we used a version that made use of 1996 test results only and an alternative version that used a simple average of the test results from the last three years (second column). Examining the affect that using a three-year average has on the lists of schools identified provides information on the importance of cohort effects.

Findings

Each of the 14 methods we applied generated a list of between 29 and 32 schools, approximately 5 percent of the sample. A total of 85 schools were identified by at least one method. Although, this indicates a significant amount of overlap among the 14 lists generated, it also indicates that there are significant differences among the lists. As Table 1 shows, only 1 school was identified by all 14 methods and only 27 percent of the schools identified were on more than half of the lists (column 4). Over 20 percent of the schools identified were only identified by one of the methods (column 3), and over half of the 85 schools identified were identified by less than a third of the methods.

Table 2 provides some indication of the sensitivity of the identification methods to cohort effects. We applied each cross-test aggregation method using 1996 test results only and using the average of 1994 to 1996 test results. The second column in Table 2 counts, for each cross-test aggregation method, the

number of schools in the union of the set identified using 1996 test results with the set identified using three-year averages. These numbers represent the total number of schools identified by either using 1996 results or using 1994 to 1996 averages. The third column in Table 2 counts the number of schools identified by either using 1996 test results only or by using 1994 to 1996 averages, but not by both. This represents the number of schools which would be identified by one of methods by not by the other. For these schools, the selection of which cohorts to use in the identification process determines whether or not they are identified. The fourth column simply expresses the number in the third column as a percentage of the number in the first column.

The first method listed in Table 2 identifies a school if the percentage above minimum competency on any of the PEP tests administered in the school ranked among the lowest 10 schools. Whether one chooses to apply this method using the 1996 cohorts only or the average of the 1994, 1995 and 1996 cohorts makes a difference for 27 schools, i.e. 27 schools would be identified under one alternative but not the other. This represents 61.4 percent of the total number of schools identified by either alternative. These figures are similar for the other cross-test aggregation methods, and indicate that differences in performance across cohorts in a school significantly effect which schools are identified as low-performing. This suggests that relying on single year of test data might provide misleading measures of the typical level of performance at a school.

Tables 3A and 3B present evidence concerning the sensitivity of New York State's approach to aggregating multiple outcome measures to the choice of cut-points and combination rules. Our base-case method identifies a school if the percentage above minimum competency in 1996 on any of the PEP tests administered in the school ranked among the lowest 10 schools. The first row of Table 3A compares the list of schools identified by this method to the list of schools whose percentage above minimum competency

on either of the third grade PEP tests ranked among the lowest 15 schools, or the percentage above minimum competency on either of the sixth grade tests ranked among the lowest 5. The second row compares our base-case method to a method that identifies those schools that are ranked among the bottom 20 on either reading test. The small change in cut-points examined in the first row only makes a difference for 10 schools. Each method individually identifies 29 schools and together they only identify 34 schools, indicating substantial overlap. This suggests that New York State’s aggregation approach might not be overly sensitive to small changes in cut-points. However, when more substantial changes in identification criteria are made, such as eliminating the use of Grade 6 tests altogether, the resulting lists of schools identified are considerably different from one another. For 27 schools, the decision whether or not to use Grade 6 test results determines whether or not it is identified as a low-performance school.

The third row in Table 3A examines the effect of changing the rule for combining sets of schools identified by various criteria into a single set of low-performance schools. More specifically, the third row compares the list of schools identified by the “or” combination rule to the list of schools identified by applying the “and” combination rule. This change has a substantial impact. Nearly, three-quarters of the schools identified by either combination rule are only identified by one of the rules. This indicates that the choice of combination rule is important.

Table 3B makes the same comparisons as Table 3A for those methods that make use of average test results over three years rather than merely the most recent year’s test results. The pattern of results in Table 3B is similar to that in Table 3A. The most notable difference between the two tables is that the aggregation methods are less sensitive to changes in cut-points and combination rules when multiple years of test results are used. This provides additional reason for New York State to use multiple years of test results in identifying schools for Registration Review.

Tables 4A and 4B examine the averaging approach to aggregating multiple outcome measures. First we compare our base-case method to a method that uses the simple, unweighted average of the percentages of students below basic competency on each test to rank and identify schools. The first row of Table 4A indicates that the lists of schools identified by these two approaches are significantly different. More than half of the schools identified by either approach are identified by only one of the approaches. The second and third row of Table 4A indicate that the averaging approach to aggregating multiple outcome measures can be quite sensitive to the weights chosen for each measures. If we increase the weights on Grade 3 and Grade 6 reading tests from 0.25 to 0.375 and reduce the weights on Grade 3 and Grade 6 math tests from 0.25 to .125, the list of schools identified as low-performing changes almost entirely. Only 10 of the 50 schools identified by either method, are identified by both methods.

Table 4B indicates that if multiple years of test results are used rather than merely the most recent years results, then the variation between the New York State aggregation approach and the averaging approach, as well as the sensitivity of the averaging approach to the choice of outcome weights, is substantially reduced. The reduction in sensitivity to the choice of outcome weights is quite marked. This indicates that making use of multiple years of test results can go a long way towards stabilizing the list of low-performing schools.

In summary, New York State officials have made explicit decisions concerning several important performance measurement and school identification issues, including what outcome measures to use, whether to focus on the absolute levels of student performance or the relative contribution of schools, and what part of the student distribution to target. However, there remain issues that need more careful consideration. Particularly, we have shown that student performance varies among different cohorts in the same school, and that the decision to rely on a single year of test results needs to be reconsidered. We have

also shown that New York State’s approach to aggregating across multiple measures of student outcomes can be sensitive to choices concerning identification cut-points and combination rules. This suggests that these aggregation issues need to be considered more explicitly.

III. Building a Reliable and Stable Identification Method for Low-Performance Schools

The results of the analysis on New York City elementary schools indicate that methods for identifying low-performance schools (or other public organizations) may be sensitive to small changes in the identification method. Specifically, different choices about what cohorts, subjects, grade levels, and aggregation methods to use in the evaluation can each change significantly which schools are identified as low performance. The lack of consistency is troubling because of the consequences the low-performance designation may have for school personnel. Despite the fact that school personnel may be performing well above average under the circumstances they face, the low-performance label carries a stigma that may affect staff morale, and may even lead to replacement of school administrators and staff.

Ideally, the identification method should reliably capture the underlying performance of students on multiple dimensions, and should be stable across different cohorts of students and small changes in the choice and weighting of assessment instruments. In this section of the paper we discuss two sets of tools that might help to improve the stability of New York State’s current process for identifying low-performance schools as well as other methods of evaluating public organizations. The first set of tools is drawn from the literature on techniques for short-term forecasting, and can be used to help remove cohort effects from the identification process. The second set of tools is drawn from the growing field of “fuzzy logic,” and has the potential to provide more robust means of aggregating multiple performance measures.

Removing Cohort Effects

One of the striking results from our analysis of New York City schools, is sensitivity of the identification process to the year of data used. The variation in relative scores across years can be accounted for by either changes in the programs and staff in the school (school effect), or differences in the preparation, ability, or motivation of students in a particular cohort (cohort effects). If the years compared are close together and no major reforms or staff changes have occurred in the school, then the school effects should be minimal. When cohort effects are significant, using test scores from the previous year to evaluate student performance in a school this year may be inaccurate. In addition, relying on only one cohort to identify schools can lead to unstable rankings across years.

We need a method that will smooth out these cohort effects, and allow us to accurately predict the performance of a typical cohort. A potential set of tools for this task are the forecasting techniques used for short-term predictions of regular phenomena, such as cash flow or revenues (Bretschneider, 1985). The goal of such forecasts is not to explain the variable being forecast, but to provide an accurate forecast over the next several time periods. The principal methods employed, commonly called univariate time-series methods, require only historical information of the phenomenon being forecast. For the case of identification of low-performance schools, such methods allow us to predict the performance of a typical cohort next year based on the performance of previous cohorts.

Most univariate forecasting methods are a variant of a moving average. In its simplest form, a moving average takes information for a set number of previous years, takes the average, and then applies this average as the forecast for next year. Each year, the oldest year is dropped off the average, and the most recent year's result is added. This simple average assumes that each year of historical data should be given equal weight in the forecast. If variation across years is entirely due to cohort effects, and distribution of

performance across cohorts is relatively random, then a simple moving average may be reasonably accurate. If school effects exist, or there are systematic changes in cohorts over time, then placing higher weights on recent years should provide a more accurate forecast. Another forecasting issue is whether test scores are trending up, or down or staying the same over time. In the case of a trend, additional steps must be taken to ensure an accurate forecast.⁸

Another approach to projecting aggregate student performance of the typical cohort is to use an “exponential smoothing” method. A simple exponential smoothing forecast without trend can be represented as:

$$F_{t+1} = \alpha X_t + (1-\alpha) F_t$$

where F_{t+1} is next year’s forecast, F_t is this year’s forecast, X_t is the actual value for this year, and α is a number between 0 and 1 which serves as the adjustment parameter. This parameter determines the weights placed on results from past years. The closer the parameter is to one, the more weight is placed on recent years. For example, if $\alpha=0.8$, then X_t has a weight of 0.80, X_{t-1} has a weight of 0.16, X_{t-2} has a weight of 0.032, etc. In contrast, if $\alpha=0.2$, then the weights for the first 3 years are 0.20, 0.16, 0.128, respectively. Commonly, forecasters try different values of α and select the one which provides the most accurate forecast.⁹

To illustrate the use of a simple univariate forecasting methods for removing cohort effects, we use norm-referenced exams in math for New York City elementary schools. Consistent data exists from 1992 until 1997.¹⁰ We will use data from 1992 through 1996 to forecast performance in 1997 using three methods, test performance for the previous year (commonly called a naive forecast), single exponential smoothing, and double exponential smoothing. Single exponential smoothing assumes there is no underlying trend, while double exponential smoothing assumes that there is a linear trend in the data.¹¹ We present forecasts

for several different α values to examine how much weight should be put on last year versus past years. Forecasting accuracy is estimated for 1997 and the average of 1993 until 1997 using both the average absolute error as a percent of the actual value (MAPE), and the square root of the average squared error (RMSE). We look at the forecast for all elementary schools in New York City, and those schools with performance in the bottom quartile in 1997.

Results of this comparison of forecasting methods are presented in Table 5. Looking first at forecasting accuracy for 1997 (second panel), we can see that the naive forecast does a fairly good job of predicting scores (column 1). For all schools, the average percent error (MAPE) is 2.1 percent and for schools in the first quartile the error is only 0.5 percent. Single exponential smoothing forecasts with an alpha level of 0.7 or higher are more accurate, with the errors with an alpha level of 0.9 almost zero (column 5). By contrast, the double exponential smoothing methods do not perform very well at any alpha level (columns 6 to 9). When accuracy is averaged over period of 1993 to 1997, the advantages of the single exponential smoothing method become even more clear. The MAPE for all schools with an alpha level of 0.9 is 3.8 percent compared to 8.8 percent for the naive forecast. The double exponential smoothing method with an alpha level of 0.9 performs about the same as the naive method.

While these results are presented purely for illustrative purposes, they do indicate that using the test score for the previous year to predict future years may not be accurate. Instead, some univariate forecasting method that develops a weighted average of several years should provide more accurate forecasts. In summary, when there is significant variation in student performance across cohorts, an average of performance across multiple years will help to remove cohort effects, and the weights on each year can be determined using established time-series forecasting methods.

Developing Stable Aggregate Performance Measures

In Section II we demonstrated the sensitivity of New York State’s approach to aggregating multiple outcomes measures to changes in cut-points and combination rules. The sensitivity of New York State’s method derives from the crisp nature of the cut-points used for each outcome measure. Either a school is identified as having low-performance on a given measure or not. There is no way to deal with schools that are close to the cut-point. Thus, although the school ranked 9th lowest on Grade 3 reading might be more similar to the school ranked 11th lowest, than the school rank 2nd lowest, it is grouped with latter.

An alternative to New York State’s approach to aggregation is the averaging approach examined in Section II. However, as we have seen, this approach can be sensitive to the choice of weights for various outcomes. “Fuzzy logic” presents another alternative.

New York State’s aggregation method is based on classic set theory (Aristotelean logic) in which an outcome is either in or out of a set. Partial membership in several sets is not allowed. For example, schools can’t be partly a member of the lowest performing school group, and the group of schools with moderately below average student performance. “Fuzzy logic” is an alternative form of logic that has been developed over the last 30 years to remove some of the constraints imposed by classic set theory. The key component of “fuzzy logic” is fuzzy set theory that allows partial membership in several sets. A school, for example, could be classified as a 0.7 member of the “low-performance” school set and a 0.4 member of the “moderately below average” set. By allowing partial membership, “fuzzy logic” based methods are much less sensitive to changes in the cut-points that define a given set. These methods attach numeric sets to imprecise terms, such as “few”, “often”, and “low”. While it is beyond the scope of this paper to provide a detailed description of “fuzzy logic” (see Treadwell, 1995, Durkin, 1994, and Ammar and Wright, 1997), we will highlight the key steps in using this methodology to develop aggregate performance measures.

1. Identify the components (test scores or other assessment measures) used in evaluating schools, and the output (overall performance) you want produced. Essentially, you need to identify the inputs and outputs of the process.
2. For each input, determine the descriptive categories which are relevant. For test scores, categories might include “above average”, “average”, “below average”, “low.” Categories also need to be developed for the aggregate output that is being created. In this case, similar categories are probably relevant.
3. Determine the cutoff points where a category begins, reaches its peak, and ends. Unlike typical classification systems, these cutoff points can overlap. Assume, for example, that for a certain criterion referenced exam that 10 percent of the students, on average, did not achieve basic competence, and that in the worst school 40 percent did not meet the standard. You might define “below average” performance as beginning with 8 percent, ending with 20 percent, and with a peak of 14 percent. The category “low” performance could be set as beginning with 15 percent and reaching full membership at 25 percent. Thus, these categories overlap from 15 percent to 20 percent. Selection of the appropriate cutoffs is typically based on actual distribution of the data and professional judgement.
4. Select membership functions for each input and the output. Essentially, these functions determine the degree of membership in a particular category. An example of triangular membership functions for the criterion referenced exam example presented above are illustrated in Figure 2. There are four categories for this exam: “above average,” “average,” “below average,” and “low”. If a particular school has 17 percent of students not passing this exam, then it has a membership values of 0.14 in “low” performance, and 0.5 in “below average” for this exam. Since the membership functions

could take on several functional forms, researchers may try several different types to see how sensitive the results are.

5. Determine the rules that should be applied in combining the inputs into an aggregate measure of output. These rules should reflect the priorities given to different inputs and the required combinations of input categories to produce particular output categories. To illustrate this process, assume that results on four tests will be used assess performance: criterion referenced and norm-referenced math and reading tests. A decision is made that the highest priority is to bring students up to basic competence in reading, and then math. However, very few students achieving grade level in both math and reading is also an indicator of poor performance. These priorities could be translated into the following series of rules about when a school is classified as “low-performance”.

(1) If performance on the criterion reading score is “low ”, and the other exam scores are “average” or lower, then the school is classified as “low-performance”.

(2) If performance on criterion reading score is below average, the performance on the criterion math exam is “low”, and the performance on the norm-referenced exams is “average” or below, then the school is classified as “low-performance.”

(3) if performance on the criterion exams are below average, and performance on the norm-referenced exams are “low”, then the school is classified as “low-performance.”

An example of the resulting matrix that summarizes the rules is presented in Figure 3. The three rules discussed above determine which schools are classified as low-performance. The rules can be readjusted to reduced or expand the number of schools classified as low performance.

6. Using the membership functions and fuzzy rules identified above, the aggregated output is generated using properties of fuzzy set theory with regard to combining various categories.¹² The results of the

aggregation process could lead to the representation of overall school performance illustrated in Figure 4. Both schools 1 and 2 have their highest membership in the category, “low” performance, but they also have partial membership in the category “below average” performance. This is a more realistic representation of a school’s performance, which may fall partially into several different categories.

7. Since a decision will still have to be made as to which schools to treat as low performance, it may still be necessary to produce an aggregate classification of a school. The process of coming up with one number or category from a fuzzy set is called “defuzzification.” While there are several different “defuzzification” methods, the most popular is to find the median or centroid of the area under each category. In Figure 4, both schools would be classified as low-performance using the centroid method, since more than 50 percent of the area falls in this category.

Because it provides for partial membership and overlapping sets “fuzzy logic” can be expected to avoid the problems associated with choosing arbitrary cut-points to distinguish different levels of performance. Studies in which “fuzzy logic” was been applied to measure organizational performance have indeed found that evaluation results are robust with respect to the choice of cut-points and different specifications of membership functions (Ammar and Wright, 1997). Of course, application of fuzzy logic involves numerous other choices including the selection of combination rules and “defuzzification” methods. Whether the “fuzzy logic” approach is robust with respect to these choices is an open question. Of course, no method of aggregating across multiple outcome measures can render value judgements about the relative importance of various outcomes irrelevant.

Whether or not “fuzzy logic” provides a more robust means of identifying low-performance, it does help to identify the various choices involved in aggregating multiple outcome measures and provides a framework for making those choice more deliberately. In addition, it can provide a richer characterization of performance than is provided by simple labels such as low-performing. For these reasons, in addition to the potential it holds for providing more stable approach to aggregating outcomes, “fuzzy logic” provides a useful tool for addressing identification issues.

IV. Conclusions

Our analysis suggests that decisions about how to aggregate individual student outcome measures into measures of school performance make a significant difference for what schools are identified as low-performance schools. Particularly, how results across multiple years are used and how different student outcome measures are aggregated is important. Even if important decisions about what outcome measures to use and about whether to focus on low-performance or low-performing schools are made, different methods of aggregating student outcomes can lead to the identification of substantially different sets of schools. This suggests the need for policy makers to explicitly address what might seem like technical identification issues.

Research can help resolve some of these issues. For instance, forecasting methods, such as “exponential smoothing,” can be used to determine schemes for weighting results from different years. If sufficient prior year data were available, models that most accurately predict test results in a given year from test results in past years could be used to identify weights. The weighted average of several years could then be interpreted as a best approximation of the true underlying level of performance in a school.

Decisions about what tests to use in the performance evaluation, what statistical measures to use from each test (what part of the student distribution), must ultimately be based on value judgements. Thus, efforts to identify voter and parent values and professional judgement about the importance of certain subjects, for example, are needed. Empirical studies that link student achievement in various areas with later life outcomes can also provide important information for this debate. For instance, studies linking performance on various high school assessments with college and job market outcomes could provide useful information. Also, studies that identify which elementary school outcomes determine success in high school are needed. Finally, evaluations of identification systems that incorporate different parts of the performance distribution (minimum competency, mastery, etc.), or that focus on performance of different socio-economic groups, such as those used in Kentucky and Texas (King and Mathers, 1997), would also provide important information.

While the decisions about how to weight various outcomes in the construction of an aggregate output are also subjective, the tools of “fuzzy logic” might be able to provide more stable aggregation methods. By allowing partial membership in several performance categories, these methods provide a more realistic classification scheme for schools which is insensitive to small changes in performance measures used. The stability of these classifications is particularly important when the results of the process are used to publicly classify schools as low-performance, and to intervene in the operation of the school. While no classification system is perfectly reliable, a system that uses multiple performance measures from different subjects and types of tests, and combines these measures with a robust aggregation method, should provide a solid foundation for identifying low-performance schools.

Endnotes

- * This paper has benefited significantly from the comments of Hampton Lankford and John Yinger, and the assistance of Henry Solomon with the data on New York City. However, we are solely responsible for any errors or omissions.
- 1. In a study of accountability systems in four states, King and Mathers (1997) found that South Carolina and Indiana use both norm- and criterion-referenced tests, Texas uses criterion-referenced tests, and Kentucky makes use of performance based assessments that include criterion-referenced elements. New York uses results on statewide criterion-referenced tests in math and reading administered in Grades 3, 6 and 8, as well as criterion-referenced tests at the high-school level.
- 2. In surveys conducted by Schumacker and Brookshire (1992), superintendents in the State of Texas rated state mandated test results less important than nine other school quality indicators. This provides evidence for the claim that "educators are unconvinced that group-administered tests provide an accurate indication of what a student knows or has learned." (Sanders and Horn, 1995)
- 3. Each of the states studied by King and Mathers (1997) use student attendance and three of the four use dropout rates. New York State uses a dropout rate standard as one of its criteria for identifying schools for Registration Review.
- 4. A related question is how changes in performance over time should be evaluated. In Kentucky performance assessments are based on how much the average level of achievement in the school increases in comparison with past years. If the rate of increase is not sufficient, the school is targeted for improvement. If the average level of achievement shows a significant decline, the school is identified as a school "in crisis". Thus, Kentucky's accountability system focuses on the trend in school performance over time rather than the absolute level of performance. This allows the state to set forth extremely high standards, and hold every school to reasonable expectations related to how quickly it is moving towards those high standards (King and Mathers, 1997). Before recent changes were made in the Registration Review process, New York State similarly focused on schools that were below a certain standard of performance and declining. By identifying schools that had shown a recent decline, the State hoped to focus assistance efforts on those schools where problems were just starting to emerge and were perhaps more manageable. The problem with this identification criteria is that schools that languish at very low levels of performance without showing a downward trend are not addressed.
- 5. The school accountability program in Texas uses school ratings that depend on the overall percentage of students meeting an achievement standard on state tests, and the average year-to-year gains on those tests made by four student groups: African-American, Hispanic, White and Economically Disadvantaged. In Kentucky students are categorized into one of four achievement levels: distinguished, proficient, apprentice and novice. Each school that fails to make a specified amount of progress towards the goal of having students performing on average at a proficient

level is required to participate in school improvement activities. If the proportion of students performing at a proficient level declines by a certain percentage the state declares the school to be "in crisis". Kentucky also provides rewards for schools that move at least 10 percent of the students previously at the "novice" level to a higher performance level (King and Mathers, 1997).

6. Consider, for instance, a state that determined it only had resources to assist 20 schools and that used 20 different indicators of school performance. Should the state, in this hypothetical example, identify only those schools that are ranked the absolute lowest on one of the 20 indicators? If some of the indicators are more important than others this approach might not be appropriate. In any case, the state would be forced to consider how much weight each measure should carry and how the several indicators should be aggregated to provide the most meaningful indices of school performance.
7. Our analysis was limited by our data to New York City schools. Registration Review is a statewide program, and does identify schools in other parts of the state. However, identification standards are formulated in terms of absolute levels of student performance rather than student gains, and no attempt is made to adjust measures of school performance based on the environment faced by the school. As a result, all the schools identified have substantial numbers of disadvantaged students, and the overwhelming majority of schools identified are in New York City.
8. Forecast with trends generally involve using two stages to construct the forecast. One stage captures the underlying stable part of the time series, and the other the trend. Another element that may exist in some time-series is regular fluctuations in the data, such as seasonality. Additional steps can be used to capture the seasonal part of a time-series. Since there is no seasonality in annual data, making seasonal adjustments is not an issue. See Makridakis, et al, 1983, for a good discussion of a number of univariate forecasting methods.
9. One of the decisions that must be made when using exponential smoothing methods, is what should be the starting value for the forecast. In the second year, a forecast must be selected to start the process. For simplicity, many forecasters use the actual value from the first year as the starting forecast. If the time-series is fairly long, or α is high, then the starting value is not important. See Makridakis, et al., 1983, for more on starting values.
10. One of the problems faced in forecasting future performance levels, is that tests are often changed in content and scoring. It is difficult to get long time-series on test scores that are consistent. One of the advantages of the exponential smoothing method, is that it does not require many years of data to produce a forecast, as long as accurate starting values can be determined.
11. The actual double-exponential method which is used is the Brown's One-Parameter Method, presented in Makridakis, et al., 1983, pp. 93-97. Another method which is more general in form is the Holt's Two-Parameter Method. The starting values for the exponential smoothing models was derived from a regression of the actual values from 1992 until 1996. The predicted value for 1992 was treated as the starting value. Makridakis, et al., 1983, recommend this in cases where the time series is not very long.

12. The process of combining fuzzy inputs into a fuzzy output set involves using several properties of fuzzy set theory. For example, the rule “or” in fuzzy set theory involves taking the maximum of the membership values for the inputs being combined. The rule “and” takes the minimum of the membership values. For example, assume a school has 0.8 membership in low performance for one exam and 0.3 membership in low performance for another. If the rule is that a school is classified as low performance if it has low performance on the first exam “or” the second exam, then the “or” function will take the maximum membership value for each exam, in this case 0.8. The school will be classified as having 0.8 membership in the category of low performance for aggregate output.

References

- Advisory Council to the New York State Board of Regents Subcommittee on Low Performing Schools. 1994). *Perform or Perish: Recommendations of the Advisory Council to the New York State Board of Regents Subcommittee on Low Performing Schools*. Albany: New York State Board of Regents.
- Ammar, Salwa and Ronald Wright. 1997. "Fuzzy Logic: A Case Study in Performance Measurement," In *Uncertainty Analysis in Engineering and Sciences: Fuzzy Logic, Statistics, and Neural Network Approach*, B. Ayyub and M Gupta (eds.). Boston: Kluwer Academic Publishers.
- Bretschneider, Stuart. 1985. "Forecasting: Some New Realities," *Metropolitan Studies Program Occasional Papers*, No. 99.
- Clotfelter, Charles and Helen Ladd. (1996). "Recognizing and Rewarding Success in Public Schools." In H.F. Ladd, ed., *Holding Schools Accountable: Performance-Based Reform in Education*, Washington, D.C.: The Brookings Institution.
- Darling-Hammond, Linda. (1991). "Accountability Mechanisms in Big City School Systems." *ERIC/CUE Digest No. 71*.
- Durkin, John. 1994. *Expert Systems: Design and Development*. New York: Macmillan Publishing Company.
- King, Richard and Judith Mathers. (1997). "Improving Schools Through Performance-Based Accountability and Financial Rewards." *Journal of Education Finance* 23: 147-176.
- Makridakis, Spyros, Steven Wheelwright, and Victor McGee. 1983. *Forecasting: Methods and Applications*. New York: John Wiley & Sons.
- Sanders, William L. and Sandra P. Horn. (1995). "Educational Assessment Reassessed: The Usefulness of Standardized and Alternative Measures of Student Achievement as Indicators of the Assessment of Educational Outcomes," *Educational Policy Analysis Archives* 3.
- Schumaker, Randall E. and William K. Brookshire. (1992). "Defining Quality Indicators for Secondary Schools," *Educational Research Quarterly* 15: 5-10.
- Treadwell, William. 1995. "Fuzzy Set Theory Movement in the Social Sciences." *Public Administration Review*. 55: 91-97.
- Webster, William and Robert Mendro. (1995). "Evaluation for Improved School Level Decision-Making and Productivity." *Studies in Educational Evaluation* 21:361-399.

Table 1
Comparison of Alternative Identification Methods

No. of Methods	No. of Schools Identified by this No. of Methods	Pct of Total No. of Schools Identified	Cumulative Pct of Schools Identified
14	1	1.18%	1.18%
11 to 13	13	15.29%	16.47%
8 to 10	9	10.59%	27.06%
5 to 7	8	9.41%	36.47%
4	8	9.41%	45.88%
3	14	16.47%	62.35%
2	14	16.47%	78.82%
1	18	21.18%	100.00%
At Least 1	85		

Table 2
Effect of Decision About How to Handle Multiple Years of Test Results

	No. of Schools Identified Using 1996 Results Only or 1994-96 Averages	No. of Schools Identified by Either Method, But Not by Both Methods	Pct. of Schools Identified by Either Method, But Not by Both Methods
“Or” combination rule with equal cutpoints	44	27	61.4%
“Or” combination rule with unequal cutpoints	43	25	58.1%
“Or” combination rule with Grade 3 tests only	41	22	53.7%
“And” combination rule	46	31	67.4%
Cross-test Average	48	36	75.0%
Weighted Cross-Test Average	40	20	50.0%
Average of Reading Tests Only	39	18	46.2%

Table 3A
Sensitivity of Aggregation Methods to Cut-points and Combination Rules
(When 1996 Tests Results Only Are Used)

	No. of Schools Identified by Either Method	No. of Schools Identified by Either Method, But Not by Both Methods	Pct. of Schools Identified by Either Method, But Not by Both Methods
“Or” rule w/equal cutpoints vs. “Or” rule w/unequal cutpoints	34	10	29.4%
“Or” rule w/equal cutpoints vs. “Or” rule w/Gr. 3 tests only	43	27	62.8%
“Or” rule w/equal cutpoints vs. “And” combination rule	48	35	72.9%

Table 3B
Sensitivity of Aggregation Methods to Cut-points and Combination Rules
(When Averaged 1994-96 Test Results Are Used)

	No. of Schools Identified by Either Method	No. of Schools Identified by Either Method, But Not by Both Methods	Pct. of Schools Identified by Either Method, But Not by Both Methods
“Or” rule w/equal cutpoints vs. “Or” rule w/unequal cutpoints	35	6	17.1%
“Or” rule w/equal cutpoints vs. “Or” rule w/Gr. 3 tests only	36	11	30.6%
“Or” rule w/equal cutpoints vs. “And” combination rule	40	19	47.5%

Table 4A
Sensitivity of Aggregation Methods to Outcome Weightings
(When 1996 Tests Results Only Are Used)

	No. of Schools Identified by Either Method	No. of Schools Identified by Either Method, But Not by Both Methods	Pct. of Schools Identified by Either Method, But Not by Both Methods
“Or” rule w/equal cutpoints vs. Cross-test Average	41	23	56.1%
Cross-test Average vs. Weighted Cross-Test Average	50	40	80.0%
Cross-test Average vs. Average of Reading Tests	54	48	88.9%

Table 4B
Sensitivity of Aggregation Methods to Outcome Weightings
(When Averaged 1994-96 Test Results Are Used)

	No. of Schools Identified by Either Method	No. of Schools Identified by Either Method, But Not by Both Methods	Pct. of Schools Identified by Either Method, But Not by Both Methods
“Or” rule w/equal cutpoints vs. Cross-test Average	37	12	32.4%
Cross-test Average vs. Weighted Cross-Test Average	32	4	12.5%
Cross-test Average vs. Average of Reading Tests	36	12	33.3%

Table 5
Comparison of Forecasting Accuracy for Several Forecasting Methods
Percent of Students Achieving Grade Level in Elementary Math Exams in New York City

	Naive Forecast	Single Exponential Smoothing Alpha Level				Double Exponential Smoothing Alpha Level			
		0.10	0.40	0.70	0.90	0.10	0.40	0.70	0.90
AVERAGE (1993-97)									
All Schools:									
RMSE ^a	3.90%	9.93%	5.30%	2.62%	1.52%	7.87%	4.32%	3.73%	3.45%
MAPE ^b	8.84%	21.28%	11.80%	6.16%	3.78%	17.25%	10.18%	8.85%	8.25%
School in First Quartile: ^c									
RMSE ^a	3.25%	8.09%	4.39%	2.26%	1.40%	6.44%	3.93%	3.54%	3.62%
MAPE ^b	8.59%	20.44%	11.47%	6.18%	3.99%	16.54%	10.58%	9.57%	9.74%
1997 ONLY:									
All Schools:									
RMSE ^a	1.10%	13.75%	4.68%	1.03%	0.17%	9.27%	0.71%	4.07%	4.46%
MAPE ^b	2.12%	26.55%	9.04%	1.99%	0.33%	17.90%	1.36%	7.86%	8.61%
School in First Quartile: ^c									
RMSE ^a	0.22%	10.42%	3.28%	0.55%	0.03%	6.74%	1.50%	4.55%	5.19%
MAPE ^b	0.51%	24.15%	7.61%	1.28%	0.08%	15.62%	3.48%	10.54%	12.03%

^aThe square root of the squared deviation between actual and forecasted value.

^bThe average percent error where all errors are expressed in absolute values. Percent error is the deviation between predicted value and actual value as a percent of actual value.

^cBased on the lowest quartile of outcomes for schools in 1997.

Figure 1: Alternative Methods for Identify Low-Performing Schools

	Uses Most Recent Year Only	Uses Average Scores Over the Last Three Years
“Or” combination rule with equal cutpoints	1996 Gr. 3 Rdg Rnk ≤ 10 or 1996 Gr. 3 Mth Rnk ≤ 10 or 1996 Gr. 6 Rdg Rnk ≤ 10 or 1996 Gr. 6 Mth Rnk ≤ 10	1994-96 Avg Gr. 3 Rdg Rank ≤ 10 or 1994-96 Avg Gr. 3 Mth Rnk ≤ 10 or 1994-96 Avg Gr. 6 Rdg ≤ 10 or 1994-96 Avg Gr. 6 Mth ≤ 10
“Or” combination rule with unequal cutpoints	1996 Gr. 3 Rdg Rnk ≤ 15 or 1996 Gr. 3 Mth Rnk ≤ 15 or 1996 Gr. 6 Rdg ≤ 5 or 1996 Gr. 6 Mth ≤ 5	1994-96 Avg Gr. 3 Rdg Rank ≤ 15 or 1994-96 Avg Gr. 3 Mth Rnk ≤ 15 or 1994-96 Avg Gr. 6 Rdg ≤ 5 or 1994-96 Avg Gr. 6 Mth ≤ 5
“Or” combination rule with Grade 3 tests only	1996 Gr. 3 Rdg Rank ≤ 20 or 1996 Gr. 3 Mth Rnk ≤ 20	1994-96 Avg Gr. 3 Rdg Rank ≤ 20 or 1994-96 Avg Gr. 3 Mth Rnk ≤ 20
“And” combination rule	1996 Gr. 3 Rdg Rnk ≤ 100 and 1996 Gr. 3 Mth Rnk ≤ 100 and 1996 Gr. 6 Rdg Rnk ≤ 100 and 1996 Gr. 6 Mth Rnk ≤ 100	1994-96 Avg Gr. 3 Rdg Rank ≤ 50 and 1994-96 Avg Gr. 3 Mth Rnk ≤ 50 and 1994-96 Avg Gr. 6 Rdg ≤ 50 and 1994-96 Avg Gr. 6 Mth ≤ 50
Cross-test Average	Rank on $\{.25(96 \text{ Gr } 3 \text{ Rdg}) + .25(96 \text{ Gr } 3 \text{ Mth}) + .25(96 \text{ Gr } 3 \text{ Rdg}) + .25(96 \text{ Gr } 6 \text{ Mth})\} \leq 30$	Rank on $\{.25(94-96 \text{ Gr } 3 \text{ Rdg}) + .25(94-96 \text{ Gr } 3 \text{ Mth}) + .25(94-96 \text{ Gr } 3 \text{ Rdg}) + .25(94-96 \text{ Gr } 6 \text{ Mth})\} \leq 30$
Weighted Cross-Test Average	Rank on $\{.375(96 \text{ Gr } 3 \text{ Rdg}) + .375(96 \text{ Gr } 3 \text{ Mth}) + .125(96 \text{ Gr } 3 \text{ Rdg}) + .125(96 \text{ Gr } 6 \text{ Mth})\} \leq 30$	Rank on $\{.375(94-96 \text{ Gr } 3 \text{ Rdg}) + .375(94-96 \text{ Gr } 3 \text{ Mth}) + .125(94-96 \text{ Gr } 3 \text{ Rdg}) + .125(94-96 \text{ Gr } 6 \text{ Mth})\} \leq 30$
Average of Reading Tests Only	Rank on $\{.5(1996 \text{ Gr } 3 \text{ Rdg}) + .5(1996 \text{ Gr } 6 \text{ Rdg})\} \leq 30$	Rank on $\{.5(94-96 \text{ Gr } 3 \text{ Rdg}) + .5(94-96 \text{ Gr } 3 \text{ Mth})\} \leq 30$

Figure 2: An Example of Membership Functions for School Performance

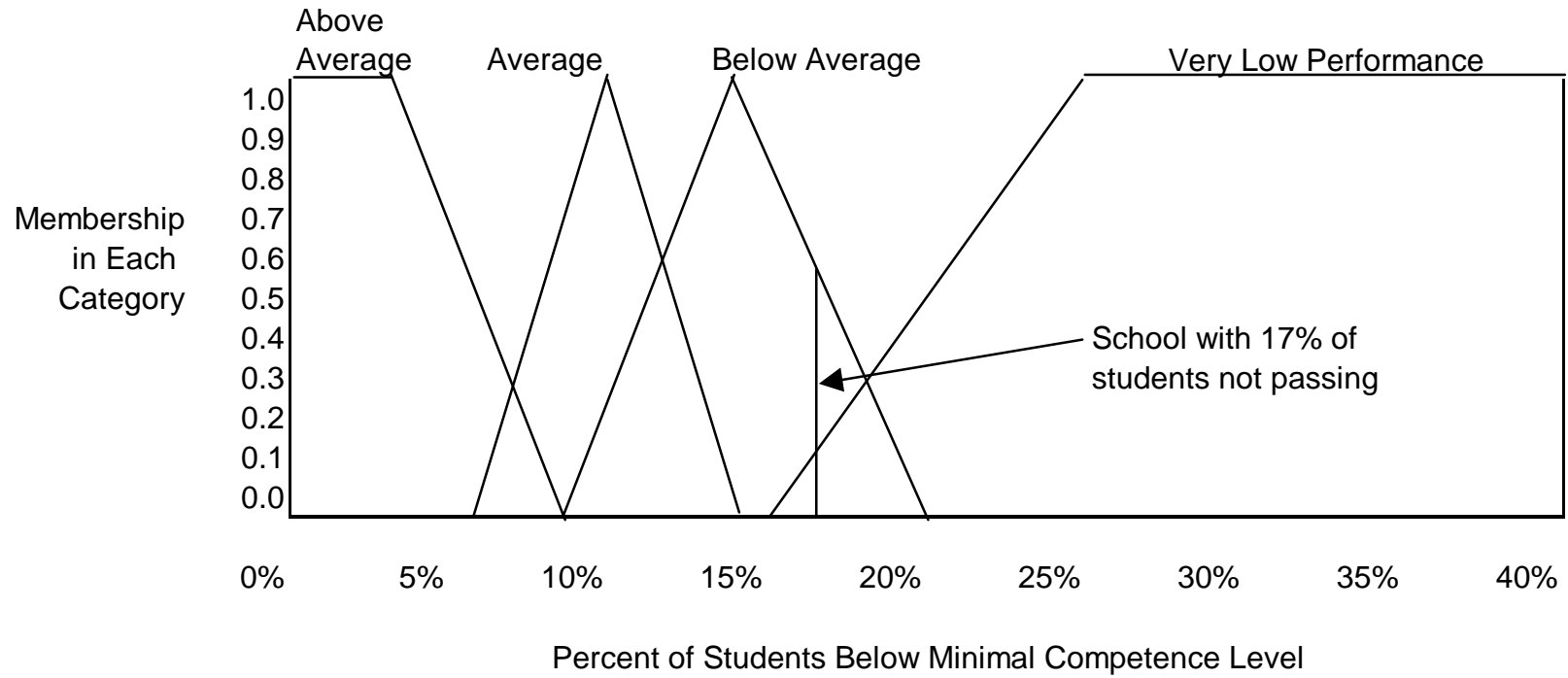


Figure 3: Example of Matrix for Classification of School Performance

Criterion-referenced Reading Exam

		Low-performance				Below Average				Average				Above Average			
		Norm-referenced exams				Norm-referenced exams				Norm-referenced exams				Norm-referenced exams			
		Low Perf.	Below Avg.	Avg.	Above Avg.	Low Perf.	Below Avg.	Avg.	Above Avg.	Low Perf.	Below Avg.	Avg.	Above Avg.	Low Perf.	Below Avg.	Avg.	Above Avg.
Criterion-Referenced Math Exam	Low Perf.	4	4	4	4	4	4	3	3	3	3	2	2	2	2	2	2
	Below Average	4	4	4	4	4	3	3	3	3	2	2	2	2	2	2	1
	Average	4	4	4	3	3	3	2	2	2	2	2	2	2	2	1	1
	Above Average	3	3	3	3	3	3	2	2	2	2	2	2	2	2	1	1

1 = Above average

2 = Average

3 = Below average

4 = Low performance

Figure 4: Example of the Results of the Fuzzy Aggregation Process

