
Forecasting Statewide Test Performance and Adequate Yearly Progress from District Assessments



Creating
Technology to
Promote Learning

John Richard Bergan, Ph.D.
John Robert Bergan, Ph.D.

Assessment Technology, Incorporated
6700 E. Speedway Blvd.
Tucson, Arizona 85710
Phone: 520/323-9033
Fax: 520/323-9139

TABLE OF CONTENTS

Forecasting Statewide Test Performance and Adequate Yearly Progress from District Assessments

*By John Richard Bergan and John Robert Bergan
Assessment Technology Incorporated*

Table of Contents	i
Acknowledgements	ii
Introduction	1
I. Forecasting Statewide Test Performance	2
A. Minimizing Forecasting Errors	2
B. Forecasting Errors in Student Subgroups	3
C. Using Forecasting Information to Guide Instruction	3
II. Forecasting AYP Classifications Using Classification Analysis	4
A. Classification Errors	4
B. Errors in Classifications Based on Statewide Test Performance	5
III. Illustration of Forecasting Techniques	5
A. Forecasting Statewide Test Performance with Single and Multiple Predictors	7
B. Classification Analysis	10
IV. Forecasting and Instruction	15
A. Instruction and Forecasting Statewide Test Performance	16
B. Instructional Goals and Plans Tied to State Standards	16
C. Instruction and Forecasting for Subgroups	17
D. Instruction and AYP Forecasting	17
V. Conclusions	18
VI. References	19

ACKNOWLEDGEMENTS

The authors wish to thank Dave Thissen for his careful review and helpful comments regarding this paper. We also wish to thank Kathy Bergan and Jody Jepson for their thorough review of the manuscript. Finally, we wish to extend our appreciation to the students, teachers, and administrators in the Marana Unified School District for their efforts, which resulted in the assessment data used in this forecasting analysis.

Introduction: Forecasting Statewide Test Performance and Adequate Yearly Progress from District Assessments

By John Richard Bergan and John Robert Bergan, Assessment Technology Incorporated

As demands for student accountability have increased (e.g. Crocker, 2003), school districts across the nation have moved toward the adoption of standards-based approaches to education. Standards-based initiatives are characterized by efforts to align district curriculums and assessments to state and local standards and to use local assessment information to inform instruction (Jamentz, 2002). Local assessments typically include teacher-made classroom tests and District-wide benchmark tests used to provide teachers and administrators with information on student learning occurring at multiple points during the school year. With the advent of the No Child Left Behind (NCLB) Act of 2001, 2002, districts have focused increasingly on the relationship between standards covered in local curriculums and assessments and the standards covered on statewide tests. States make the coverage of standards available through test blueprints detailing the content covered in statewide assessments. These blueprints assist districts in their efforts to align curriculum activities and local assessments to goals articulated in state standards and measured by statewide assessments.

Establishing the link between local curriculum and assessment and statewide test performance is a critical element in educational initiatives designed to meet NCLB requirements. Under NCLB, statewide tests play a defining role in evaluating the effectiveness of schools and the adequacy of individual student learning. The determination of adequate yearly progress (AYP) as required by NCLB (2001, 2002) is directly linked to statewide assessment initiatives. The determination of AYP calls for the establishment of annual measurable objectives specifying the percentage of students required to meet state standards. Over time, the required percentages increase until all students are required to meet standards. The effort to relate local assessment and instructional initiatives to AYP requirements is enhanced by the ability to forecast statewide test performance from local assessments. Forecasting assists teachers to use local assessment information to guide instruction toward improved skill acquisition and as a result improved performance on statewide tests. In addition, forecasting provides administrators with advance information to allocate resources aimed at promoting the mastery of state standards used in evaluating the performance of schools and individual students.

This paper discusses techniques for forecasting statewide test performance and AYP classifications from local benchmark tests. Data are presented illustrating the application of forecasting techniques in a standards-based instructional initiative in Arizona. The data analyzed include performance on local benchmark assessments in mathematics and reading and literature administered at three points during the school year. Statewide test performance was assessed using the Arizona Instrument to Measure Standards (AIMS). The key points made in the discussion are as follows:

1. Forecasting statewide test performance using standard techniques can provide useful information for guiding instruction and for setting cut points indicating that standards have been met on district benchmark tests.
2. Forecasting statewide test performance represents a useful first step in forecasting AYP. However, standard forecasting techniques do not provide all of the information necessary to forecast AYP. Determining AYP requires breaking a continuous statewide test score distribution into discrete categories. The fundamental forecasting

question is to determine the probability that a given student will meet state standards given his or her performance on benchmark tests.

3. Mastery classification procedures (e.g., Bergan, Schwarz, & Reddy, 1999) can be used to determine the probability of meeting state standards based on benchmark test performance. Mastery classification techniques can detect the probability of false-positive errors (i.e., false predictions that state standards will be met) and false-negative errors (i.e., false predictions that state standards will not be met). When the probabilities associated with these two types of errors are known, forecasts of AYP classifications can be made.
4. It is possible that the relationships between statewide test scores and benchmark test performance may vary across groups of students reflecting diverse backgrounds. If subgroup differences exist and are not detected, predictions of statewide test scores and forecasts of AYP classifications may produce misleading results. This paper demonstrates procedures for detecting subgroup differences and accommodating those differences in the forecasting process.

I. Forecasting Statewide Test Performance

Effective forecasting of statewide test performance and AYP classifications necessitates focus on three issues. The first is the minimization of errors associated with the prediction of statewide test performance from district assessment performance. The second has to do with the possibility that forecasting errors may vary across subgroups, and the third is related to the role of forecasting information in guiding instruction.

A. Minimizing Forecasting Errors

Regression is the time-honored statistical technique designed to minimize prediction errors. However, other techniques such as factor analysis and structural equation modeling may also be used in prediction initiatives (Dorans, 2004). In the simplest case, prediction involves computing the covariance or correlation between two tests. For example, we could predict statewide scores by correlating those scores with scores on a district benchmark test. Errors associated with predicted statewide test scores can be quite large even when the correlation between a benchmark test and a statewide test is of substantial magnitude. For example, let's assume a correlation of .70 between a benchmark test and a statewide test. The standard error of estimate for that correlation would be .71 standard deviations. The 95% confidence interval for that standard error would be roughly plus or minus 1.4 standard deviations from the estimated statewide test score. Dorans (2004) has pointed out that even with correlations as large as reliability coefficients (e.g., .87) the amount of uncertainty associated with a predicted score may be substantial. For example, a correlation of .87 would produce a confidence interval of plus or minus one standard deviation.

Correlations between benchmark tests and statewide tests are not likely to be as large as reliability coefficients. Benchmark tests and statewide tests are intended to serve different purposes. Benchmark tests are typically designed to assess a limited set of objectives that have been targeted for instruction. Statewide tests used in accountability initiatives are designed to evaluate overall student accomplishments covering an extended time span. These differences in purpose call for differences in test specifications.

It is reasonable to expect that the relationship between benchmark assessment and statewide assessment can be improved by including multiple predictors in the regression equation. Benchmark tests are typically administered at multiple points during the school year. Moreover, the tests generally vary in content. Insofar as there is content variability, it is possible that each benchmark test could contribute to statewide test variability and thereby reduce forecasting errors. Multiple regression was used in the present study to address the question of whether or not multiple benchmark predictors would be effective in reducing forecasting errors.

B. Forecasting Errors in Student Subgroups

It is possible that forecasting errors may vary across subgroups in the student population. For example, forecasting errors may vary between students from minority groups and students in the majority group. Subgroup forecasting errors can lead to erroneous conclusions regarding forecasted performance for majority and/or minority students. Accordingly, it is important to consider the possibility of subgroup differences in research on the forecasting of statewide test performance from benchmark test performance.

Research aimed at detecting differences in the relationships between tests associated with subgroup variations has been a concern for many years. Early work focused on the issue of test bias (e.g. Cleary, 1968). The problem addressed in this study is different from the test-bias issues that stimulated the thinking of early researchers in the field. The problem of interest here is how best to use test information to promote learning in different subgroups. The question of central concern is whether or not it will be helpful to disaggregate forecasts of statewide tests by subgroup. Disaggregation will be useful if the regression coefficients in multiple regressions differ by subgroup. Statistical techniques associated with confirmatory factor analysis and structural equation modeling are particularly well suited to test hypotheses regarding subgroup variations (e.g. Jöreskog & Sörbom, 1996). These techniques were used in the present study to test the hypothesis of equal regression coefficients in multiple regression models for different subgroups.

C. Using Forecasting Information to Guide Instruction

Forecasting statewide test performance from benchmark assessments is useful because it provides a best estimate of a student's likely score on the statewide test. This information can be helpful in guiding instruction. Forecasting can be used to indicate the cut points on a benchmark test that are likely to be associated with meeting standards on the statewide test. For example, if the state test includes cut points for approaching, meeting, and exceeding the standard, forecasting information can be used to identify corresponding cut points on one or more benchmark tests. The benchmark cut points can serve as a basis for establishing instructional goals. For instance, if a student's projected score falls below the benchmark cut point for meeting the standard, instruction can be targeted toward the acquisition of skills that will lead toward meeting the standard.

Using forecasts of statewide test performance as a basis for establishing district instructional goals makes it possible to link state standards and goals set at the district level. When forecasting is implemented, the benchmark cut points informing local goals are based on two sources of information, the cut points for the state standards and the regression of the statewide test on benchmark tests. These sources of information create a link between district instructional goals and state standards. The validity of the link can be assessed empirically by testing hypotheses that the regression coefficients relating statewide and benchmark test performance contribute significantly to variation in statewide test performance.

II. Forecasting AYP Classifications Using Classification Analysis

Forecasts of statewide test performance provide an initial step toward determining estimates of AYP classifications. Statewide test forecasts provide an estimate of the statewide test score a student is likely to attain given the score actually attained on one or more benchmark tests. This information is necessary, but not sufficient for estimating the AYP classifications. In order to estimate a student's AYP classification, it is necessary to determine whether or not the estimated statewide test score is above or below the relevant AYP cut point. Then, it is necessary to determine the probability that the estimated classification will match the actual classification attained based on statewide test performance. It also will be of interest to forecast the proportion of students likely to meet the state standard and the proportion of students likely not to meet the state standard. This information is useful in determining whether or not annual measurable objectives used in establishing AYP are being met.

Classification Analysis (CA) is among the most widely used techniques for determining the probability that estimated classifications will match actual classifications (Bergan, Schwarz & Reddy, 1999). CA is a process for examining the cross-classification of predictor and criterion variables. The cross classification of concern here is that between benchmark classifications and classifications based on statewide test performance. There are four possible outcomes in this cross classification: The first is the case in which benchmark assessment and statewide assessment conclude that the student has met the established standard. The second is the case in which both assessments conclude that the standard has not been met. The remaining outcomes cover the two possible cases of disagreement between benchmark and statewide assessment classifications.

A. Classification Errors

The major purpose of CA is to identify classification errors or conversely, the accuracy of classification predictions. CA can be used to identify both false positive errors and false negative errors. **False positive errors** occur when students who do not meet state standards based on statewide test performance are classified as meeting standards based on district assessment results. False positive errors produce overly optimistic conclusions about the percentage of students meeting state standards. **False negative errors** occur when students meeting state standards are classified as not meeting standards based on district assessment performance. False negative errors underestimate the percentage of students meeting standards. Both types of errors can play a role in estimating the proportion of students likely to meet or not meet a standard. For example, if the probability of a false negative error is .15, 15% of those students designated as failing to meet the standard based on benchmark test performance can be expected to meet the standard based on statewide test performance. This information can be used in estimating the proportion of students likely to meet the state standard.

CA includes procedures for minimizing a particular type of classification error. For example, an administrator using CA may choose to minimize false positive errors. This is done through the use of a weight indicating the cost of misclassification. If the cost of false positive classifications is deemed to be high, the weight can be applied to raise the classification bar thereby reducing false positive classification errors.

B. Errors in Classifications Based on Statewide Test Performance

One limitation of CA is that it assumes that there is no measurement error in the criterion variable under examination (Bergan, Schwarz, & Reddy, 1999). For example, the analyses conducted in the present study assumed no measurement error in the classifications based on statewide test performance. Thus, if a student were classified as having met the standard based on benchmark test results and categorized as having not met the standard based on statewide test performance, the outcome was categorized as a false positive result. This would not necessarily be the case if measurement error were assumed to exist in the statewide test.

The assumption that measurement error is zero for classifications based on statewide test performance is useful in modeling conditions that generally exist in state accountability systems. Of course, as is the case for any test, there is measurement error associated with statewide test performance. States address this fact in a number of ways. For example, students who fail a statewide test may be given an opportunity to take the test again. Nonetheless, in many instances classification decisions associated with state accountability systems are consistent with the assumption that there is no measurement error in the criterion variable. For instance, state calculations of AYP classifications for the data analyzed below are consistent with the assumption of no measurement error in the criterion variable. Accordingly, this is the assumption made in the present study. However, a companion to this paper examines the case in which measurement error in the criterion variable is recognized (Bergan, Bergan, & Guerrero, 2005).

III. Illustration of Forecasting Techniques

The following sections illustrate our approach to prediction analysis with samples of 3rd, 5th, 8th, and 10th grade students from schools in Arizona using the *Galileo Educational Management System*. These students took the AIMS exam in 2004 and three benchmark exams included in the Galileo System over the course of the year. We begin by illustrating a procedure for detecting subgroup differences. We use structural equation modeling procedures for this purpose. One of the major benefits of structural equation modeling is that it provides a direct statistical test of the hypothesis that structural coefficients, including regression coefficients, are the same across subgroups.

We divided the sample into two ethnic groups: Those labeled Caucasian and those in any other category, which we labeled Diverse Ethnicities. These included students classified as Hispanic, Asian, Indian, and Other. Most of these students were classified as Hispanic. Of course, it would be possible to examine differences among specific ethnic groups. The split we used was chosen to illustrate the technique for detecting subgroup differences. Table 1 shows the means and standard deviations for the Caucasian subgroup and the Ethnically Diverse subgroup on the AIMS scales and the benchmark tests.

**FORECASTING STATEWIDE TEST PERFORMANCE AND
ADEQUATE YEARLY PROGRESS FROM DISTRICT ASSESSMENTS**

TABLE 1
Means and Standard Deviations for AIMS and Benchmark Tests

Subject	Group	AIMS		Benchmark 1		Benchmark 2		Benchmark 3	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
3 rd Math	Caucasian	539.50	63.10	758.60	71.49	896.30	93.90	954.30	99.30
	Diverse Ethnicities	528.90	61.90	746.10	74.90	874.90	95.60	943.80	103.30
3 rd Reading	Caucasian	524.00	42.00	772.30	91.70	848.00	95.90	838.00	102.10
	Diverse Ethnicities	518.30	33.80	761.30	90.40	838.20	99.00	830.60	96.40
5 th Math	Caucasian	515.70	53.90	1003.10	91.40	1094.70	110.10	1147.90	122.60
	Diverse Ethnicities	498.84	55.70	980.70	87.70	1059.70	119.40	1101.80	129.60
5 th Reading	Caucasian	511.10	38.20	967.80	81.80	1072.60	97.50	961.20	85.70
	Diverse Ethnicities	505.00	40.80	945.80	86.10	1044.40	102.30	938.70	88.70
8 th Math	Caucasian	468.00	44.20	1311.30	93.90	1419.80	85.40	1417.00	96.50
	Diverse Ethnicities	455.89	44.20	1293.40	101.30	1400.50	92.50	1402.30	91.00
8 th Reading	Caucasian	500.50	43.70	1313.70	85.80	1390.00	83.60	1312.40	96.40
	Diverse Ethnicities	487.90	40.10	1304.30	85.90	1377.80	93.10	1293.40	96.20
10 th Reading	Caucasian	528.90	47.20	1556.00	90.00	1494.80	91.30	1540.60	119.40
	Diverse Ethnicities	509.10	36.90	1523.20	105.10	1457.30	95.90	1502.10	111.70

Overall, the means for the Caucasian group are somewhat higher than those for the Ethnically Diverse group. The mean differences underscore the need to increase learning opportunities for students from diverse ethnic backgrounds. The standard deviations are generally similar in the two groups. However, there are a few cases in which the standard deviations are somewhat larger in one group than in another. It is important to note that the subgroup differences revealed in Table 1 do not in themselves suggest the need to forecast statewide test results separately for the Ethnically Diverse and Caucasian subgroups. For example, it is quite possible to have equivalent regression coefficients in circumstances in which there are mean differences between subgroups. The question of whether or not to forecast separately by subgroup is addressed by testing the hypothesis that subgroup regression coefficients are equivalent.

Table 2 shows the results of analyses testing the hypothesis that the regression coefficients for the Caucasian and Ethnically Diverse Subgroups do not differ significantly.

TABLE 2
Tests of Subgroup Equivalence in Regression Coefficients Predicting AIMS Scores

Grade	Subject	Chi-square	df	p
3 rd	Math	3.37	3	.34
	Reading	92.18	3	<.01
5 th	Math	22.96	3	<.01
	Reading	13.14	3	<.01
8 th	Math	34.36	3	<.01
	Reading	14.15	3	<.01
10 th	Reading	90.13	3	<.01

In these analyses, maximum likelihood estimates of the regression coefficients for the Caucasian and Ethnically Diverse Subgroups were constrained to be equal. These constraints added three degrees of freedom to the regression models under examination, one for each of the three benchmark tests included in the analyses. The fit of the constrained regression models to the data was assessed by comparing the values in the actual covariance matrices to the values in

**FORECASTING STATEWIDE TEST PERFORMANCE AND
ADEQUATE YEARLY PROGRESS FROM DISTRICT ASSESSMENTS**

each of the expected covariance matrices for the constrained models using the likelihood-ratio chi-squared statistic (Jöreskog & Sörbom, 1996). Table 2 shows the values for each of the chi-squared tests, the degrees of freedom for each test, and the p value for each test.

The results shown in Table 2 indicate that the equivalence hypothesis can be safely rejected in all cases with the exception of the math results from the 3rd grade students. These findings indicate the need to consider subgroup differences in making predictions of scores on statewide tests or in forecasting AYP classifications. As will become apparent in the examination of data presented below, in most instances the relationship between benchmark assessments and the statewide test are of similar magnitude across subgroups despite the fact that they are not equivalent. However, there are cases in which the relationships are not similar. In some cases, failure to take into account subgroup differences may result in misleading conclusions about the relationships between benchmark test performance and performance on statewide tests. The suggestion for districts is that examination of subgroup differences should be an essential component of any forecasting initiative.

A. Forecasting Statewide Test Performance with Single and Multiple Predictors

As soon as a benchmark test has been completed, the results can be used to guide instruction. The value of each benchmark test as an instructional guide is enhanced to the extent that benchmark test performance is related to statewide test performance. Thus, it is important to examine the relationship between each benchmark test and statewide test performance. Tables 3, 4, 5, and 6 present correlation matrices for AIMS and Benchmark Math and Reading and Literature Tests. All of the correlations between the benchmarks and AIMS are significantly different from zero.

TABLE 3
Correlation Matrices for Third-Grade AIMS and Benchmark Tests

<i>Group</i>	<i>Test</i>	1	2	3	4
All Students	1 Math Benchmark 1	1.00			
	2 Math Benchmark 2	.68	1.00		
	3 Math Benchmark 3	.62	.73	1.00	
	4 Math AIMS	.61	.66	.62	1.00
Caucasian Students	1 Reading Benchmark 1	1.00			
	2 Reading Benchmark 2	.76	1.00		
	3 Reading Benchmark 3	.71	.81	1.00	
	4 Reading AIMS	.47	.55	.56	1.00
Ethnically Diverse Students	1 Reading Benchmark 1	1.00			
	2 Reading Benchmark 2	.75	1.00		
	3 Reading Benchmark 3	.74	.78	1.00	
	4 Reading AIMS	.75	.76	.77	1.00

**FORECASTING STATEWIDE TEST PERFORMANCE AND
ADEQUATE YEARLY PROGRESS FROM DISTRICT ASSESSMENTS**

TABLE 4
Correlation Matrices for Fifth-Grade AIMS and Benchmark Tests

<i>Group</i>	<i>Test</i>	1	2	3	4
Caucasian Students	1 Math Benchmark 1	1.00			
	2 Math Benchmark 2	.74	1.00		
	3 Math Benchmark 3	.69	.80	1.00	
	4 Math AIMS	.72	.78	.79	1.00
Ethnically Diverse Students	1 Math Benchmark 1	1.00			
	2 Math Benchmark 2	.77	1.00		
	3 Math Benchmark 3	.70	.82	1.00	
	4 Math AIMS	.71	.81	.79	1.00
Caucasian Students	1 Reading Benchmark 1	1.00			
	2 Reading Benchmark 2	.74	1.00		
	3 Reading Benchmark 3	.70	.76	1.00	
	4 Reading AIMS	.52	.51	.54	1.00
Ethnically Diverse Students	1 Reading Benchmark 1	1.00			
	2 Reading Benchmark 2	.78	1.00		
	3 Reading Benchmark 3	.76	.81	1.00	
	4 Reading AIMS	.53	.56	.58	1.00

TABLE 5
Correlation Matrices for Eighth-Grade AIMS and Benchmark Tests

<i>Group</i>	<i>Test</i>	1	2	3	4
Caucasian Students	1 Math Benchmark 1	1.00			
	2 Math Benchmark 2	.69	1.00		
	3 Math Benchmark 3	.67	.65	1.00	
	4 Math AIMS	.75	.68	.72	1.00
Ethnically Diverse Students	1 Math Benchmark 1	1.00			
	2 Math Benchmark 2	.64	1.00		
	3 Math Benchmark 3	.71	.64	1.00	
	4 Math AIMS	.79	.66	.76	1.00
Caucasian Students	1 Reading Benchmark 1	1.00			
	2 Reading Benchmark 2	.53	1.00		
	3 Reading Benchmark 3	.55	.62	1.00	
	4 Reading AIMS	.56	.58	.66	1.00
Ethnically Diverse Students	1 Reading Benchmark 1	1.00			
	2 Reading Benchmark 2	.52	1.00		
	3 Reading Benchmark 3	.58	.61	1.00	
	4 Reading AIMS	.54	.59	.67	1.00

TABLE 6
Correlation Matrices for Tenth-Grade AIMS and Benchmark Tests

<i>Group</i>	<i>Test</i>	1	2	3	4
Caucasian Students	1 Reading Benchmark 1	1.00			
	2 Reading Benchmark 2	.63	1.00		
	3 Reading Benchmark 3	.69	.67	1.00	
	4 Reading AIMS	.61	.60	.61	1.00
Ethnically Diverse Students	1 Reading Benchmark 1	1.00			
	2 Reading Benchmark 2	.55	1.00		
	3 Reading Benchmark 3	.58	.56	1.00	
	4 Reading AIMS	.62	.66	.65	1.00

In some cases these correlations are almost as high as might be expected for a reliability coefficient. For instance in Table 4 for Ethnically Diverse Students the correlation between the second math benchmark exam and the math scaled score on AIMS was .81. It is also notable that the relationship between the benchmark exams and AIMS in some cases is quite different for the two ethnic subgroups. For example, the correlation between the 3rd grade reading benchmarks and AIMS was appreciably higher for the Ethnically Diverse group than it was for the Caucasian group. Overall, the results show that each of the benchmark tests is significantly related to performance on the statewide test. Accordingly, each of these tests can play a useful role in guiding instruction toward the achievement of educational standards measured by performance on AIMS.

As results for more than one benchmark test become available during the course of the school year, multiple regression analyses become a possibility. As indicated earlier, multiple regression may reduce forecasting errors in that the relationship between multiple predictor variables and a criterion variable is likely to be higher than would be the case if only a single predictor variable were used. In the present study, multiple regression was used to examine the prediction achieved when multiple benchmark tests were used to forecast AIMS. The contribution of each regression coefficient to the regression model was assessed using *t* tests (Jöreskog & Sörbom, 1996). All of the regression coefficients made a significant contribution to the fit of the regression models to the data. The multiple correlations and squared multiple correlations attained in the regression analyses are shown in Table 7.

**FORECASTING STATEWIDE TEST PERFORMANCE AND
ADEQUATE YEARLY PROGRESS FROM DISTRICT ASSESSMENTS**

TABLE 7
Multiple Correlations and Squared Multiple Correlations for Caucasian and Ethnically Diverse Groups

Grade	Subject	Group	N	Multiple-R	R-Squared
3 rd	Reading	Caucasian	508	.72	.52
		Diverse Ethnicities	226	.82	.68
5 th	Math	Entire Sample	734	.72	.51
		Caucasian	497	.84	.71
	Reading	Diverse Ethnicities	241	.84	.71
		Caucasian	497	.58	.34
8 th	Math	Diverse Ethnicities	241	.60	.36
		Caucasian	473	.81	.66
	Reading	Diverse ethnicities	206	.84	.71
		Caucasian	473	.72	.52
10 th	Reading	Diverse Ethnicities	206	.72	.52
		Caucasian	337	.68	.47
		Diverse Ethnicities	112	.77	.59

The multiple correlations in these analyses are all of substantial magnitude. Moreover, they are invariably higher than the correlations for single predictor variables. This is to be expected given that all of the regression coefficients contributed significantly to the fit of the models to the data.

As indicated earlier, multiple-regression results such as those presented here provide information that can assist in determining AYP classifications. However, regression results do not directly address the question of determining the percentage of students who are likely to make adequate yearly progress. Since the percentage of students who make adequate yearly progress is the central component used in determining AYP, forecasting that percentage is a useful activity. The data of central interest in forecasting AYP is categorical data including the number of students who meet state standards and the number of students who do not meet state standards. Forecasting AYP requires the analysis of categorical data. Classification Analysis is a useful technique for forecasting involving categorical information.

B. Classification Analysis

In the present study, we initiated Classification Analysis by using the results of the regression analyses presented above to determine whether or not each student's predicted score on AIMS would fall at or above or below the state cut point indicating that standards were met. Predicted scores falling below the state cut point were classified as not meeting the standard. Scores falling at or above the state cut point were classified as meeting the standard. We cross-classified the predicted classifications with the actual classifications based on AIMS performance. The results are presented in tables 8, 9, 10, and 11.

**FORECASTING STATEWIDE TEST PERFORMANCE AND
ADEQUATE YEARLY PROGRESS FROM DISTRICT ASSESSMENTS**

TABLE 8
Classification Analyses for the 3rd Grade Sample

Group	Cross Classification				Forecasts	Values	
<i>All Students</i>				Predicted Math		Sensitivity	.94
	AIMS Math	<i>Met</i>	<i>Not Met</i>	<i>Total</i>		<i>Specificity</i>	.66
	<i>Standard Met</i>	513	32	545		<i>AYP Accuracy</i>	.96
	<i>Standard Not Met</i>	63	125	188			
	<i>Total</i>	576	157	733			
<i>Caucasian Students</i>				Predicted Reading		Sensitivity	.93
	AIMS Reading	<i>Met</i>	<i>Not Met</i>	<i>Total</i>		<i>Specificity</i>	.80
	<i>Standard Met</i>	375	29	404		<i>AYP Accuracy</i>	.98
	<i>Standard Not Met</i>	21	83	104			
	<i>Total</i>	396	112	508			
<i>Ethnically Diverse Students</i>				Predicted Reading		Sensitivity	.92
	AIMS Reading	<i>Met</i>	<i>Not Met</i>	<i>Total</i>		<i>Specificity</i>	.77
	<i>Standard Met</i>	157	13	170		<i>AYP Accuracy</i>	1.00
	<i>Standard Not Met</i>	13	43	56			
	<i>Total</i>	170	56	226			

The data in all four tables are arranged in the same fashion. We depict this arrangement by describing the data in Table 8. The column on the far left indicates the group of students whose data are analyzed. The columns in the center of the table present the cross classifications for each group included in the analyses. The last two columns in the table indicate the types of forecasts included in the analysis and the values obtained for each forecast.

As indicated earlier, the hypothesis of equivalent regression coefficients across subgroups could not be safely rejected for the regression of third-grade AIMS math scores on third-grade math benchmark tests. The classification analysis for third-grade math was conducted on all students rather than on the Caucasian and ethnically diverse subgroups. The cross classification for math indicates that 513 students were correctly predicted to meet the standard based on their performance on the math benchmark tests. By contrast, 125 students were correctly predicted to have achieved a score that did not meet the standard on AIMS. Sixty-three students were incorrectly predicted to have met the standard on AIMS and thirty-two students were incorrectly predicted not to have met the AIMS standard.

Indices of sensitivity, specificity, and AYP Accuracy are given to the right of the cross-classification of actual and predicted AIMS outcomes. As indicated earlier, the sensitivity index indicates the proportion of students accurately forecasted to meet state standards. For math, 545 students actually met the standard. Of these, 513 students were accurately predicted to meet the standard. The resulting value for the sensitivity index is .94. The specificity index gives the proportion of students correctly identified as having not met state standards. For the third-grade group, 125 students were correctly identified as having not met the state standard. However, 188 students actually failed to meet state standards. The resulting value for the specificity index is .66.

The AYP Accuracy index is given directly below the specificity index. Accurately forecasting the percentage of students who will meet statewide standards is particularly important in initiatives aimed at promoting AYP. The AYP Accuracy Index was designed to measure the

accuracy of predictions of the percentage of students meeting state standards. For the 3rd grade math group, the AYP Accuracy Index is .96. This value indicates that the percentage of students meeting state standards has been forecasted with 96% accuracy. This percentage is based on three numbers, the number of students who actually met the state standard (545), the number of students predicted to meet the standard (513), and the total number of students in the sample (733). In the 3rd grade group the absolute difference between the actual and predicted number of students meeting the standard is 31 students. These 31 students reflect the amount of error in the forecast. The percentage of error in the forecast is computed by dividing the forecasting error (31) by the sample size (733). The AYP Accuracy Index is obtained by subtracting the error percentage from 1.

Overall, the results in Table 8 indicate that for the third-grade sample, the sensitivity of benchmark predictions was higher than the specificity associated with those predictions. The benchmark tests were more effective in identifying students who met the standard on the statewide test than they were at identifying students who failed to meet the standard. The relative effectiveness of sensitivity and specificity estimates is affected by how high the bar is set for determining a predicted passing score. Lowering the bar would increase specificity and reduce sensitivity. In the present study the bar was set by requiring that the predicted AIMS score be at or above the cut point for meeting the standard established by the state. For the third-grade sample, this cut point produced more pleasant surprises in the form of unexpected instances of meeting state standards than disappointments in which students predicted to meet state standards failed to do so.

The AYP Accuracy Index was in almost all cases higher than either the sensitivity or specificity indices in the third-grade sample. This forecast is less demanding than either the sensitivity or specificity forecasts in that it does not require predicting which students will meet or fail to meet state standards. Rather, the forecast is based on a prediction of how many students will meet state standards.

**FORECASTING STATEWIDE TEST PERFORMANCE AND
ADEQUATE YEARLY PROGRESS FROM DISTRICT ASSESSMENTS**

Results for the 5th grade sample are presented in Table 9.

TABLE 9
Classification Analyses for the 5th Grade Sample

Group	Cross Classification				Forecasts	Values	
<i>Caucasian Students</i>				Predicted Math		Sensitivity	.91
	AIMS Math	<i>Met</i>	<i>Not Met</i>	<i>Total</i>		Specificity	.81
	<i>Standard Met</i>	284	27	311		AYP Accuracy	.98
	<i>Standard Not Met</i>	36	150	186			
	<i>Total</i>	320	177	497			
<i>Ethnically Diverse Students</i>				Predicted Math		Sensitivity	.81
	AIMS Math	<i>Met</i>	<i>Not Met</i>	<i>Total</i>		Specificity	.85
	<i>Standard Met</i>	96	22	118		AYP Accuracy	.98
	<i>Standard Not Met</i>	18	105	123			
	<i>Total</i>	114	127	241			
<i>Caucasian Students</i>				Predicted Reading		Sensitivity	.92
	AIMS Reading	<i>Met</i>	<i>Not Met</i>	<i>Total</i>		Specificity	.66
	<i>Standard Met</i>	292	27	319		AYP Accuracy	.93
	<i>Standard Not Met</i>	61	118	179			
	<i>Total</i>	353	145	498			
<i>Ethnically Diverse Students</i>				Predicted Reading		Sensitivity	.93
	AIMS Reading	<i>Met</i>	<i>Not Met</i>	<i>Total</i>		Specificity	.69
	<i>Standard Met</i>	106	8	114		AYP Accuracy	.87
	<i>Standard Not Met</i>	39	88	127			
	<i>Total</i>	145	96	241			

The values for sensitivity tended to be higher than the values for specificity. However, this was not the case for math in the ethnically diverse group. Here the sensitivity index was slightly lower than the specificity index. The AYP Accuracy Index exceeded both the sensitivity and specificity indices in all cases except one. High values for AYP Accuracy were found for all groups and all grades.

**FORECASTING STATEWIDE TEST PERFORMANCE AND
ADEQUATE YEARLY PROGRESS FROM DISTRICT ASSESSMENTS**

Results for the 8th grade sample are presented in Table 10.

TABLE 10
Classification Analyses for the 8th Grade Sample

Group	Cross Classification			Forecasts	Values	
<i>Caucasian Students</i>	Predicted Math			Sensitivity	.67	
	AIMS Math	<i>Met</i>	<i>Not Met</i>	<i>Total</i>	Specificity	.93
	<i>Standard Met</i>	71	35	106	AYP Accuracy	.98
	<i>Standard Not Met</i>	26	341	367		
	<i>Total</i>	97	376	473		
<i>Ethnically Diverse Students</i>	Predicted Math			Sensitivity	.60	
	AIMS Math	<i>Met</i>	<i>Not Met</i>	<i>Total</i>	Specificity	.95
	<i>Standard Met</i>	18	12	30	AYP Accuracy	.99
	<i>Standard Not Met</i>	9	167	176		
	<i>Total</i>	27	179	206		
<i>Caucasian Students</i>	Predicted Reading			Sensitivity	.78	
	AIMS Reading	<i>Met</i>	<i>Not Met</i>	<i>Total</i>	Specificity	.77
	<i>Standard Met</i>	195	55	250	AYP Accuracy	.99
	<i>Standard Not Met</i>	51	173	224		
	<i>Total</i>	246	228	474		
<i>Ethnically Diverse Students</i>	Predicted Reading			Sensitivity	.74	
	AIMS Reading	<i>Met</i>	<i>Not Met</i>	<i>Total</i>	Specificity	.87
	<i>Standard Met</i>	65	23	88	AYP Accuracy	.96
	<i>Standard Not Met</i>	15	104	119		
	<i>Total</i>	80	127	207		

In the eighth-grade sample, there is a reversal of the tendency for sensitivity to exceed specificity. Specificity exceeds sensitivity for math in both the Caucasian and Ethnically Diverse student groups. We note that in the eighth-grade sample the proportion of students meeting state standards in math is considerably lower than was the case in the earlier grades. For example, in the Ethnically Diverse Group only 30 students out of 206 students met the standard on the statewide assessment. We also note that the specificity indices for math were both above .90. The benchmark assessments accurately predicted that large numbers of students would fail to meet the State math standard. Predictions of this kind can be used to target instructional resources to promote learning.

As in other grades, the overall number of students meeting the standard on AIMS was predicted quite accurately. For example, the AYP Accuracy Index for math was .99 in the Ethnically Diverse Student Group. In this group, 27 students were predicted to meet the state standard in math while 30 students actually did meet the standard.

**FORECASTING STATEWIDE TEST PERFORMANCE AND
ADEQUATE YEARLY PROGRESS FROM DISTRICT ASSESSMENTS**

Table 11 contains the Classification Analyses for the 10th grade sample. For this sample, Benchmark Tests were administered in the area of Reading and Literature, but not in the area of Math. Math assessments in the District were linked to specific subjects such as algebra and geometry rather than to the statewide test, which covered a broad range of topics related to mathematics.

TABLE 11
Classification Analyses for the 10th Grade Sample

Group	Cross Classification			Forecasts	Values	
		Predicted Reading				
<i>Caucasian Students</i>	AIMS Reading	<i>Met</i>	<i>Not Met</i>	<i>Total</i>	Sensitivity	.91
	<i>Standard Met</i>	248	25	273	Specificity	.85
	<i>Standard Not Met</i>	8	44	52	AYP Accuracy	.95
	<i>Total</i>	256	69	325		
<i>Ethnically Diverse Students</i>	AIMS Reading	<i>Met</i>	<i>Not Met</i>	<i>Total</i>	Sensitivity	.83
	<i>Standard Met</i>	52	11	63	Specificity	.61
	<i>Standard Not Met</i>	16	25	41	AYP Accuracy	.95
	<i>Total</i>	68	36	104		

In the tenth-grade sample, the sensitivity and specificity indices were considerably higher for the Caucasian Student Group than for the Ethnically Diverse Student group. We note that the number of ethnically diverse students assessed at this grade was small. Moreover, the number was much smaller than was the case for the Ethnically Diverse Student Group at earlier grades. The reliability of results is affected by sample size. Results in large samples are likely to be more reliable than results obtained with small samples. As with other grades, the AYP Accuracy Index was quite high. Even for the small number of Ethnically Diverse Tenth Graders, the AYP Accuracy Index provided a close estimate of the number of students meeting the state standard.

IV. Forecasting and Instruction

The standard-based educational environment that characterizes contemporary education calls for an alignment between instruction, assessment, and educational standards and performance objectives. In a standards-based initiative, curriculum is aligned with standards and objectives, and instruction is targeted toward the attainment of objectives. Assessment serves two major purposes. The first is to provide information that can be used to guide instruction. The second is to measure instructional outcomes. In order to fulfill these purposes, assessment instruments must also be aligned with standards and objectives. Standards and objectives provide the organizing framework that guides instruction and promotes the use of assessment information to inform instruction.

The standards-based approach is designed to link standards-based initiatives in local schools to social policy goals articulated at state and national levels. Standards and performance objectives established at state levels provide the foundation for this link. However, by themselves they do not provide the information needed to insure that learning is progressing adequately toward the mastery of standards. For example, a teacher may attempt to align instruction with standards. Yet, his or her students may or may not acquire the skills needed to meet standards

based on their performance on statewide tests administered toward the end of the school year. Teacher-made classroom tests may be helpful in providing information regarding the extent to which specific objectives have been mastered. However, classroom tests do not provide information regarding the likelihood that students will meet standards based on their performance on statewide tests. The movement toward benchmark testing has its origins in the need for information regarding student achievement levels in advance of the time that achievement level information based on statewide test performance becomes available.

A. Instruction and Forecasting Statewide Test Performance

When appropriate forecasting procedures are in place, advance information obtained through benchmark testing can be used to inform instruction while there is still time for students to progress toward standards mastery. The research presented here illustrates procedures for forecasting statewide test performance from benchmark tests. The first step in the forecasting process is to establish the relationship between each benchmark test and statewide test performance. Establishing the relationship between a benchmark test and a statewide test provides evidence of the validity of the benchmark assessment. Tests that correlate significantly with a statewide test are to some degree measuring the same competencies as those assessed on the statewide test. The correlations presented in tables 3 through 6 indicate that all of the benchmark tests used in the present study are significantly related to performance on the statewide AIMS tests.

The data in tables 3 through 6 confirm the existence of error in the prediction process. The correlations between benchmark tests and AIMS are not perfect, nor should they be expected to be perfect. As indicated earlier, benchmark tests and statewide tests serve different purposes. Benchmark tests are designed to guide instruction and to measure progress during the school year. Statewide tests are summative assessments used primarily for accountability purposes. Despite these differences, the regression of a statewide test on one or more benchmark tests can be highly useful in informing instruction. The standard regression techniques illustrated here are designed to minimize squared errors of prediction. Accordingly, they provide a useful estimate of what level of achievement will be required to meet state standards.

B. Instructional Goals and Plans Tied to State Standards

The relationship between a benchmark test and the statewide test provides the foundation for setting instructional goals that are tied to state standards. When the covariance or correlation between a benchmark test and a statewide test is known, it is possible to set a cut point on the benchmark test representing an instructional goal that, when attained, predicts that the state standard as evidenced by statewide test performance will be met.

Of course, it would be desirable to know which skills to teach and how many skills would have to be mastered in order to reach the established cut point. Item Response Theory (IRT) can be helpful in determining what skills to teach and how many will have to be mastered to reach the desired cut point. IRT is the accepted standard for conducting psychometric analysis for statewide tests. IRT is also useful in the construction of benchmark assessments. When IRT is used in the development of benchmark tests, it is possible to identify the specific skills reflected in performance objectives that need to be acquired in order to obtain a benchmark test score at or above the instructional goal reflected in the established cut point. IRT places measures of student achievement and measures of test-item difficulty on the same scale. When test items are

tied to performance objectives, it is possible to identify the specific objectives that the student is ready to learn and that will lead to the attainment of an achievement score that is likely to eventuate in the mastery of state standards as evidenced by statewide test performance. The teacher who knows the specific objectives that need to be targeted to maximize the likelihood of meeting state standards is in a good position to plan effective learning opportunities for students. The teacher using benchmark forecasting to plan instruction is able to link assessment to the achievement of state standards in a way that is not possible to obtain when assessment is limited to teacher-made classroom tests.

C. Instruction and Forecasting for Subgroups

The data on regression-coefficient equivalence across subgroups indicate the need to address the possibility of forecasting differences for different subgroups. The equivalence hypothesis was rejected in every case except for third-grade math. Although equivalence was rejected, in most cases subgroup differences were small. Nonetheless, when subgroup differences are not taken into account, the effectiveness of instructional goals and plans based on forecasting may be affected. For example, for third-grade reading, there was a ten point subgroup difference in the multiple correlation between AIMS and benchmark tests. The multiple correlation was higher for the ethnically diverse student group than for the Caucasian student group.

When Districts attempt to address the problem of possible subgroup differences, they may encounter circumstances in which the number of students falling in certain subgroups is too small to adequately detect subgroup differences using standard statistical procedures. In the present study, this problem was addressed by combining students from different ethnic backgrounds into a single group. Other remedies for this problem also exist. For example, Districts with similar student demographics may choose to combine their data for analysis purposes.

D. Instruction and AYP Forecasting

The procedures used for forecasting AYP classifications included three forecasting indices: sensitivity, specificity, and AYP Accuracy. The Results presented for all three of these indices support the assumption that benchmark assessments can be an effective tool in forecasting AYP. Each of these indices can be used in planning instruction in ways that take advantage of forecasting information. The sensitivity index can be used to identify how likely it is that students predicted to meet state standards will actually meet them. When the sensitivity index is high, schools can plan resource allocations with confidence that students predicted to meet state standards will very likely do so. When the sensitivity index is relatively low, schools must plan for the possibility that some students predicted to meet standards will not do so in the absence of additional instructional intervention. Similar adjustments can be made based on information provided by the specificity index. When the index is high, schools can be confident that students predicted to fail to meet standards will likely do so unless additional intervention is undertaken. When the index is low, schools can expect a number of pleasant surprises in which students predicted to fail to master standards will actually achieve mastery.

The results presented for the AYP Accuracy Index indicate that this index can provide an estimate of the number of students likely to meet state standards based on statewide test performance that will closely approximate the number of students who actually meet standards. In those instances in which the predicted number is lower than the number required for AYP, the index provides an advance warning signal making it possible to adjust instructional resources as needed to achieve AYP requirements.

V. Conclusions

The present study illustrates an approach to forecasting statewide test performance and AYP classifications that can yield useful information for goal setting and instructional planning designed to promote standards mastery. The approach illustrated ways in which standard regression techniques can be used to link goal setting and planning in local districts to social policy goals articulated at state and national levels. In addition, classification analysis techniques and related forecasting indices were presented that can be used in forecasting AYP classifications and in instructional resource allocations related to the attainment of AYP.

In order for the approach illustrated here or other forecasting approaches to be useful to schools, forecasting initiatives must be applied in a manner that takes account of the rapid change that is the hallmark of the current accountability landscape. For example, the statewide test under examination in the present study is currently undergoing significant change. A new state contractor is now responsible for statewide testing. The number of grades in which the test is administered is being expanded. There have been recent modifications in state standards and test content is being adjusted to relate to national norms. At the same time benchmark initiatives are undergoing change. Many schools are introducing new curriculums and altering benchmark tests to reflect curricular changes. Textbook publishers are also making curricular changes to reflect new state standards. These changes create additional demands for alterations in benchmark assessment.

The task of making forecasting useful to schools requires a continuing research program. Data must be collected and analyzed rapidly to accommodate continual changes occurring at state and local levels. Test equating procedures must be implemented to provide continuity in benchmark assessments conducted over multiple years. Finally, technology must be applied to make forecasting information available in meaningful ways and in a timely fashion to educational stakeholders. Clearly, the challenges associated with forecasting initiatives are great. At the same time the potential realization of improved instructional effectiveness through forecasting represents a unique and potentially significant advancement in the conduct of education.

VI. References

- Bergan, J.R., Bergan, J.R., & Guerrero, C. (2005). *Standards Mastery Determined by Benchmark and Statewide Test Performance*. Tucson, AZ: Assessment Technology Inc.
- Bergan, J.R., Schwartz, R.D., & Reddy (1999). Latent structure analysis of classification errors. *Applied Psychological Measurement*, 23 (1), 69-86.
- Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 10. 43-56.
- Crocker, L. (2003). Teaching for the test: validity, fairness, and moral action. *Educational Measurement: Issues and Practice*. Vol. 22, No. 3 (pp. 5-11).
- Dorans, N. J. (2004). Equating, Concordance, and Expectation. *Applied Psychological Measurement* Vol. 28, No. 4 (pp. 227-246). Thousand Oaks, SAGE Publications.
- Jamentz, K. (2002). *Isolation is the Enemy of Improvement: Instructional Leadership to Support Standards-Based Practice*. San Francisco, CA. WestEd.
- Jöreskog, K.G. & Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Chicago, IL: Scientific Software International.