A Comparison of Alternative Models for Estimating School Performance in Mathematics and Reading/Language Arts in Four State Accountability Systems: Arizona Results

NCAASE Technical Report
January, 2018

Ann C. Schulte[b]
Joseph J. Stevens[a]
Joseph F. T. Nese[a]
Nedim Yel[c]
Gerald Tindal[a]
Daniel Anderson[a]
Stephen N. Elliott[b]

[a] University of Oregon
[b] Arizona State University
[c] Indiana University

Address all correspondence to Ann C. Schulte, Arizona State University, Sanford School of Social and Family Policy, Arizona State University, PO Box 873701, Tempe, AZ 85287-3701, ann.schulte@asu.edu

This research was funded in part by a Cooperative Service Agreement from the Institute of Education Sciences (IES) establishing the National Center on Assessment and Accountability for Special Education – NCAASE (PR/Award Number R324C110004); the findings and conclusions expressed do not necessarily represent the views or opinions of the U.S. Department of Education. For more information about NCAASE see www.ncaase.

Table of Contents

List of Tables

A Comparison of Alternative Models for Estimating School Performance in Mathematics and Reading/Language Arts in Four State Accountability Systems: Results

Background and Introduction

This technical report is one of a series of four reports that describe the results of a study comparing eight alternative models for estimating school academic performance using data from Arizona, North Carolina, Oregon, and Pennsylvania accountability systems. Our purpose was not to evaluate the accountability systems in use by these states, but to evaluate a broader range of models commonly used for estimating school performance in many states and frequently reported in the school effectiveness research literature. This introduction describes the study background and details the methods and procedures we used to estimate the eight school performance models and compare model results in all four states. The individual state technical reports including details on each state's accountability data, assessment instruments, and results are provided at: http://www.ncaase.com/publications/tech-reports.

Despite the central importance of analytic models used in evaluating teacher and school effects in modern accountability systems, there are relatively few studies of the reliability and validity of these high-stakes systems (see, for example, Goldschmidt, Choi, & Beaudoin, 2012). The results reported here examine eight models using operational state accountability data in mathematics and reading/language arts from the participating states. We addressed four questions surrounding the use of analytic models for the evaluation of school performance:

1. Are estimates of school performance stable across successive cohorts of students?
2. How well do estimates of school performance correlate among models?
3. How do estimates of school performance correlate with variables describing the student composition of the school?
4. Do estimates of school performance vary from one model to another based on the school composition of students with disabilities (SWD)?

General Method Description

Sample

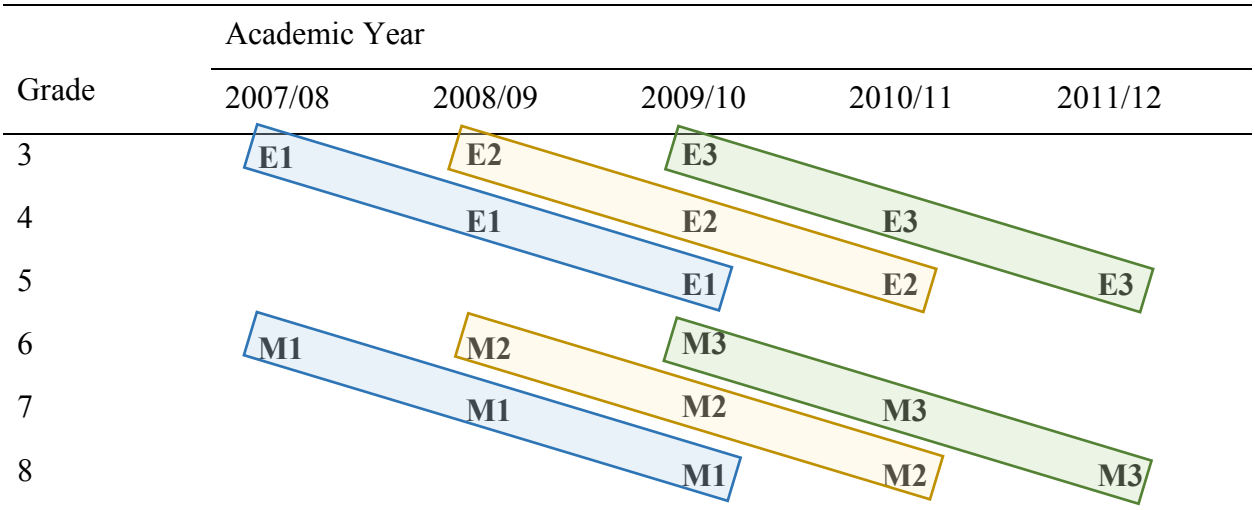
The sample from each state is described in each individual state technical report. In three of the four states, we began by creating longitudinal cohorts drawn from all students who took the state's mathematics or reading/language arts general assessment in any one school year from 2007/08 through 2011/12, and whose records in each year were included in the state's calculation of Adequate Yearly Progress (AYP). Samples were separated into two grade level bands: a longitudinal elementary school sample (Grades 3 through 5) and a longitudinal middle school sample (Grades 6 through 8), each consisting of three cohorts (a) 2007/08 through 2009/10; (b) 2008/09 through 2010/11; and (c) 2009/10 through 2011/12 (see research design schematic below). In Arizona, only one elementary and middle school cohort was used (2006/07 through 2008/09) due to changes in the Arizona testing program in 2010.

Instruments

The outcome measures for all analyses were the standardized mathematics and reading/language arts tests used for accountability in each state. In three of the states, the instruments used vertically linked developmental scales created using item response theory (IRT) methods. In Pennsylvania, the test was not vertically linked over grades preventing the

estimation of certain school performance models described in the next section. Additional details on each state's accountability test are provided in its individual state technical report.

Research design indicating academic years and longitudinal cohorts studied:



Note. E denotes an elementary school cohort, M denotes a middle school cohort.

School Performance Models

For all models, we estimated school performance in the last focal year (Grade 5 or 8) of the two grade level bands, or using prior years of achievement data as dictated by the particular model. We applied eight alternative analytic models of school performance to the mathematics and reading/language arts achievement data in elementary and middle school for each state. The eight school performance models were: Percent Proficient (PP), gain score (Gain), transition matrix (TM), student growth percentile (SGP), value-added model (VAM), and three Multilevel Linear Model (MLM) estimates: focal year intercept or status (MLM0), focal year growth rate (Grate), and average MLM growth rate across the three years (AvGrate).

Percent Proficient (PP). PP was the NCLB required metric used by the state that calculated the percentage of students in each school that met or exceeded state benchmarks for proficiency in either mathematics or reading/language arts in each grade.

Average Gain Score. Gain scores were calculated as the prior academic year (Grade 4 or Grade 7) scale score in mathematics or reading/language arts subtracted from the focal year scale score (Grade 5 or Grade 8):

$$\text{Gain}_i = \Delta_i = Y_{it} - Y_{i(t-1)} \qquad (1)$$

where Y_{it} was the assessment outcome for student *i* at time *t*. Student gain scores were averaged for each school (labeled "Gain" below).

Transition Matrix (TM). School performance estimates were computed from a table of the state's proficiency categories in the prior year crossed with the proficiency categories in the focal year (Grade 5 or Grade 8) which, in the case of five proficiency categories, created a

transition matrix table of 25 cells. The percentage of students occurring in each of the cells was entered and then a weighting scheme was applied to each cell and the products were summed to create a TM school performance index. The weighting scheme awarded one of three scores: (a) -1 was recorded if the student moved down one or more categories from the previous year, (b) 0 was recorded if the student stayed in the same category, and (c) +1 was recorded if the student moved up one or more categories from the previous year (see Tindal, Nese, & Stevens, 2017). The weighted values were averaged across all cells to create an overall school TM index.

Student Growth Percentiles (SGP). Student growth percentiles were computed at the student level using the approach described by Betebenner (2009). A student's SGP was calculated by taking the current year test score and regressing it on the two prior years of test scores. Betebenner's (2009) approach uses ordinal methods (quantile regression) as well as B-spline, cubic polynomial smoothing of the resulting normative distribution of conditional regression estimates. The analysis results in a relative rank for each student in a conditional distribution of those who had similar scores in previous years. We used the R package *SGP* (Betebenner, & Iwaarden, 2011) to compute student estimates based on the regression of the two prior years of test scores on the current year's test score and then we aggregated student SGP for each school to create a median SGP as each school's SGP performance estimate.

Value-added Models (VAM). This mixed effects approach examined performance gains over years and included indicators for student membership in a particular school. This model is known generally as the "layered model" because layers of equations are added with each year of schooling (Ballou, Sanders, and Wright, 2004). For example, the model for our case with students with three years of data would be specified as follows:

$$Y_{0ij} = b_0 + u_0 + e_0 \quad (2a)$$
$$Y_{1ij} = b_1 + u_0 + u_1 + e_1 \quad (2b)$$
$$Y_{2ij} = b_2 + u_0 + u_1 + u_2 + e_2 \ , \quad (2c)$$

where Y_{tij} represents an assessment for student *i* at time *t* (grade) attending school *j*. The fixed mean for all students in the combination of grades and schools was μ_{tij}, while e_{tij} was the random deviation for student *n* from the mean, μ_{tij}. The layered model we used was limited to a maximum of three years and was applied separately to mathematics and reading/language arts.

Multilevel Linear Growth Model Initial Status, Focal Year Growth, and Average Growth (MLM0, MLM Growth Rate and MLM Average Growth Rate). We modeled student growth over the three elementary or three middle school grades with multilevel longitudinal analyses (Raudenbush & Bryk, 2002) using HLM 7.1 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011) and full maximum likelihood estimation. The conditional models included a level-1 model that specified student mathematics or reading/language arts scores predicted by a quadratic function of time of measurement, a level-2 model composed of the prediction of level-1 model parameters as a function of student mean values, and a level-3 model composed of the prediction of level-2 parameters as a function of school mean parameter values. Time was centered on the focal year (Grades 5 or 8) for computation of MLM0 and MLM growth rate but was centered on the middle year (Grades 4 or 7) for computation of MLM average growth rate. We used a quadratic model based on previous findings (Bloom, Hill, Black, & Lipsey,

2008) as well as inspection of the data and statistical testing of alternative growth functions. Because only three time points were present, the model intercept and linear slope were random parameters but the variance of the quadratic parameter was fixed (note the omission of a residual term in equation 4c below) to obtain a model solution. We used two different centering definitions to take into account the curvilinear nature of growth. Although centering in the last, focal year is most consistent with the definition of other models, it likely underestimates the amount of growth that occurs over the three year period because of deceleration. We therefore also centered on the middle grade in the three year span to produce an average growth rate over the three years. The resulting MLM model equations were:

Level 1 (Time):

$$(Y_{tij}) = \pi_{0ij} + \pi_{1ij}(\text{time}_{tij}) + \pi_{2ij}(\text{time squared}_{tij}) + e_{tij} \quad (3)$$

Level 2 (Students):

$$\pi_{0ij} = \beta_{00j} + r_{0ij} \quad (4a)$$
$$\pi_{1ij} = \beta_{10j} + r_{1ij} \quad (4b)$$
$$\pi_{2ij} = \beta_{20j} \quad (4c)$$

Level 3 (Schools):

$$\beta_{00j} = \gamma_{000} + u_{00j} \quad (5a)$$
$$\beta_{10j} = \gamma_{100} + u_{10j} \quad (5b)$$
$$\beta_{20j} = \gamma_{200} + u_{20j} \quad (5c)$$

where Y_{tij} was the mathematics or reading/language arts scale score for student i at time t in school j, π_{0ij} was the initial status or intercept for student i at time 0 in school j, π_{1ij} was the linear rate of change, π_{2ij} was the quadratic curvature representing the acceleration or deceleration in each student's growth trajectory and e_{tij} was the residual for each student. At level-2, the level-1 parameters were modeled using mean parameter values across students (β_{k0j}) and at level-3, the level-2 parameters were modeled using mean parameter values across schools (γ_{k0j}).

Comparison of Model Estimates

We used several comparison criteria to evaluate the comparability and stability of school estimates across school performance models and across cohorts. In each state technical report we describe the results of our evaluation of school performance estimates. We examined: (a) correlations of model estimates for each school across the three cohorts, (b) correlations among school estimates from one model to another, (c) correlations among the school estimates and school composition variables (e.g., percent economically disadvantaged in the school, percent minority students in the school), and (d) correlations of each model with the percentage of students with disabilities in the school.

Comparison of School Ranks Based on Model Estimates

Many states and districts create school ranks based on their accountability system results. To compare the alternative school performance models using this metric, we created school percentile ranks (from 1 to 99, with 99 being the highest performance) based on each of the school performance models described above. In one of the only studies evaluating school performance models, Goldschmidt, Choi, and Beaudoin (2012) compared models using quintiles.

They examined the percentage of times schools remained in the same quintile band based on one school performance model versus another. Similarly, Castellano and Ho (2013) compared SGP and conditional regression models by examining the percentage of times schools remained within 1, 5 or 10 percentile ranks for each model. To maintain some comparability with each of these studies, we used three levels of similarity in school ranks, computing the percentage of schools within 5, 10, or 20 ranks of each other. We also computed the Spearman's correlation of school ranks from one cohort to another or from one school performance model to another. As a final comparison metric, we computed the root mean squared difference (RMSD) between school ranks based on each pair of cohorts or each pair of school performance models (see Castellano & Ho, 2013):

$$RMSD_{c,c} = \sqrt{\frac{\sum_{j=1}^{j}(Rank_{jc} - Rank_{jc})^2}{n}} \quad (6)$$

In equation 6, for a particular school performance model, the RMSD computes the difference ($Rank_{it}$) between each school's rank in one cohort (*jt*) versus the school's rank in a second cohort (*ju*), squaring the difference, summing across all schools, dividing by the number of schools, *n*, and taking the square root of the result.

$$RMSD_{mn} = \sqrt{\frac{\sum(Rank_{jm} - Rank_{jn})^2}{n}} \quad (7)$$

Similarly, in equation 7, the school ranks arising from alternative school performance models are compared in which $Rank_{jm}$ and $Rank_{jn}$ represent the rank of school *j* using school performance model m compared to that school's rank using school performance model *n*. As in equation 6, differences in ranks are then summed, squared, divided by the number of schools and taken to the ½ power. The RMSD was a measure of similarity in school performance models where a lower value indicates a pair of models that rank schools most similarly.

Summary

We evaluated eight models for estimating school academic performance in mathematics and reading/language arts using operational state accountability data. In NC, OR, and PA, we examined stability in model estimates across three successive student cohorts in mathematics and reading/language arts in both elementary and middle school grades. In all four states, we also compared the estimates of school performance from one model to another to determine whether the models provided similar or different depictions of school performance, although several models could not be estimated in Pennsylvania because their test did not have a vertically linked score scale. We then compared the degree to which model estimates correlated with variables describing the student composition of the school, a likely indication of construct irrelevant variance. Ideally, estimates of school performance should not be related to the student composition of the school. Last, we evaluated the school performance models in terms of the way they ranked schools, the stability of school ranks across cohorts, and the degree of agreement in school rankings from one school performance model to another. Detailed results of these analyses and comparisons follow for the state of Arizona.

Arizona Study

Method

Sample

The sample was separated into an elementary school sample (Grades 3 through 5) and a middle school sample (Grades 6 through 8), each consisting of the cohort of students enrolled in school years 2006/07 through 2008/09. The initial sample included all students whose Grade 5 (elementary school sample) or Grade 8 (middle school sample) reading or mathematics scores on the general or alternate assessment were included in the state calculation of Adequate Yearly Progress (AYP). There was a small number of cases where a unique student identifier appeared to have been associated with more than one student in a year. When conflicting reading or mathematics scores were associated with a student identifier, all records were removed for that student identifier in that year. The initial elementary school sample for the mathematics test was 81,067 students. The initial middle school sample for the mathematics test was 79,424 students. The initial elementary school sample for the reading test was 81,081 students. The initial middle school sample for the reading test was 79,446 students.

To create an analytic sample that was appropriate for our research questions, we only included students with valid test scores in all three years in schools that served all three grades (Grades 3 through 5 or 6 through 8). Students who did not follow the typical grade level sequence due to grade retention, acceleration, or dubious progressions were excluded from the sample. We included only schools that served all three grades for a cohort, and schools with $N \geq 10$ students in the final reference year of the three-year grade level band (i.e., Grade 5 for elementary grades 3 to 5 and Grade 8 for middle grades 6 to 8). Students and schools that did not meet these criteria were excluded from analyses. Due to a lack of 2005/06 data, were unable to exclude students present in 2007/08 who had been retained or accelerated from the previous year, as we had done in other states included in the larger study. As is the case in most operational and research applications of these models, we made no attempt to account for student mobility in years prior to the focal year or to make any attributions of "school effects" based on how many years the student had been in the focal year school. Our concern in creating the analytic sample was to maximize the interpretation of comparisons of the models rather than to ensure complete representativeness of the samples. These inclusion rules were applied to ensure that there were no differences in the analytic samples for different school models so that comparisons of school models were a function only of differences in the models and not the composition of the sample analyzed. The final elementary school analytic sample for the mathematics test was 61,660 students (76.06% of the initial sample). The final middle school analytic sample for the mathematics test was 41,806 students (56.64%). The final elementary school analytic sample for the reading test was 61,713 students (76.11%). The final middle school analytic sample for the reading test was 41,837 students (52.66%). The greater losses in number of students from the initial to the analytic sample for the middle school mathematics and reading cohorts were largely due to the requirement that schools in the sample have served students in all three grades (Grades 6 to 8 for middle school cohorts) for each of the three years that comprised the longitudinal cohort.

Table 1 provides summary statistics describing the school-level analytical samples of elementary and middle school students in the mathematics and reading cohorts. School composition variables reported in the table include the percent of English Language Learners

(ELL), females, economically disadvantaged students (EDS), ethnic minorities, and students with disabilities (SWD). It should be noted that when we refer to "school" composition, it references variables representing a particular cohort in each school in our analytic samples. Because we excluded students and schools to create our analytic samples, "total school" characteristics may differ slightly from the variables reported here. The reading and mathematics samples for each grade band were quite similar. The elementary and middle school cohorts differed somewhat. For example, there were lower percentages of students classified as ELL in the middle school cohorts.

Table 1

Proportion and Standard Deviation (in parentheses) of Student Subgroups for the Arizona Analytical Samples by Content Area and Grade Level Band

		Proportion (SD)
Mathematics Elementary	ELL	0.102 (0.128)
	Female	0.492 (0.077)
	EDS	0.534 (0.304)
	Ethnic Minority	0.550 (0.299)
	SWD	0.130 (0.065)
Reading Elementary	ELL	0.102 (0.128)
	Female	0.493 (0.077)
	EDS	0.534 (0.304)
	Ethnic Minority	0.550 (0.299)
	SWD	0.130 (0.065)
Mathematics Middle	ELL	0.081 (0.107)
	Female	0.498 (0.093)
	EDS	0.529 (0.306)
	Ethnic Minority	0.576 (0.303)
	SWD	0.123 (0.084)
Reading Middle	ELL	0.081 (0.107)
	Female	0.498 (0.093)
	EDS	0.529 (0.306)
	Ethnic Minority	0.576 (0.303)
	SWD	0.124 (0.084)

Instrument

The outcome measures for all analyses were the mathematics and reading versions of the Arizona Instrument to Measure Standards (AIMS). The AIMS Reading (AIMS-R) and Mathematics (AIMS-M) are multiple-choice, standardized tests that were designed as dual-purpose assessments, providing both norm-referenced and criterion-referenced scores (CTB/McGraw-Hill, 2008). The criterion-referenced scores only made use of test items aligned

with the reading and mathematics content standards in the state curriculum that had been written by Arizona teachers or drawn from the TerraNova (CTB/McGraw-Hill, 2001; CTB/McGraw-Hill, 2008). The AIMS-R and AIMS-M criterion-referenced scores served as the primary outcome data in the state's school accountability model during the school years included in the study. The criterion-referenced portions of the AIMS were placed on a vertical scale across grades using items from the TerraNova embedded in the test and the item parameters from the three-parameter logistic model from the TerraNova's national standardization sample CTB/McGraw-Hill, 2008).

Results and Discussion

This technical report is organized in two sections: Section A describes school performance model estimates and Section B describes school ranks.

Section A: School Performance Estimates

Cohort stability. We were unable to examine the stability of model estimates across successive cohorts due to a change in test publisher and a change in mathematics test editions in 2010 in Arizona.

Comparison of models. We computed the correlations of school performance estimates from one model to another. Table 2 shows model correlations for mathematics and reading in the elementary school and middle school samples.

Table 2

Correlations of School Performance Model Estimates Across Models by Content Area and Grade Level Band

Elementary School Mathematics

Model	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
PP	0.909	0.260	0.271	0.448	0.486	0.159	0.415
MLM0		0.288	0.227	0.447	0.539	0.191	0.451
Gain			0.917	0.841	0.865	0.961	0.593
TM				0.799	0.782	0.88	0.535
SGP					0.943	0.702	0.852
VAM						0.729	0.898
Grate							0.383

Elementary School Reading

Model	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
PP	0.927	-0.063	0.013	0.345	0.406	-0.214	-0.042

MLM0	-0.137	-0.104	0.300	0.374	-0.283	-0.113
Gain		0.847	0.759	0.781	0.950	0.662
TM			0.665	0.668	0.803	0.564
SGP				0.904	0.614	0.724
VAM					0.644	0.791
Grate						0.542

Middle School Mathematics

Model	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
PP	0.928	0.494	0.461	0.498	0.524	0.589	0.256
MLM0		0.530	0.426	0.505	0.565	0.651	0.252
Gain			0.882	0.915	0.940	0.943	0.692
TM				0.822	0.822	0.820	0.611
SGP					0.947	0.809	0.824
VAM						0.855	0.867
Grate							0.492

Middle School Reading

Model	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
PP	0.919	0.425	0.273	0.330	0.430	0.587	0.434
MLM0		0.434	0.188	0.298	0.447	0.629	0.433
Gain			0.798	0.853	0.891	0.920	0.766
TM				0.723	0.697	0.685	0.586
SGP					0.886	0.773	0.847
VAM						0.898	0.960
Grate							0.770

Average Over Content Area and Grade Level Band

Model	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
PP	0.921	0.279	0.255	0.405	0.462	0.280	0.266
MLM0		0.279	0.184	0.388	0.481	0.297	0.256
Gain			0.861	0.842	0.869	0.944	0.678
TM				0.752	0.742	0.797	0.574
SGP					0.920	0.725	0.812
VAM						0.782	0.879
Grate							0.547

As evident in Table 2, substantial variability was present in the degree to which school performance estimates for one model were related to other models, and the correlations among models varied by content area and grade level band. For example, the correlation between the MLM0 model and the MLM Grate model ranged from -.283 to +.651 and between PP and MLM Grate ranged from -.214 to +.589. The least variation in model correlations across content area and grade level band was for the PP and MLM0 model, with correlations ranging from +.909 to +.928.

As shown in the last panel of Table 2, on average across content area and grade level band, the strongest correlations were between the Grate and Gain models (+.944), the MLM intercept (MLM0) and PP models (+.921), and the SGP and VAM models (+.920). The weakest average correlation was between the MLM0 and TM model (+.184). The average correlation of the two status models (PP, MLM0) with the remaining six multiyear models was only +.319. Average correlations among the six multiple year models ranged from +.547 to +.944 with all models showing average intercorrelations larger than .72, except for Grate, where some of the correlations with other multiyear models were in the moderate range (.50 to .70).

We also examined the degree to which school performance model estimates were consistent from one content area to the other. Table 3 shows model estimate agreement across content areas in each cohort. Correlations were generally higher between content areas in elementary than middle school. Correlations across content areas for the two status models (PP and MLM0) were greater than +.80, and more robust than the correlations across content areas for the other models, which ranged from .261 to .649.

Table 3
Correlations of School Performance Model Estimates between Mathematics and Reading by Grade Level Band

	Elementary Schools	Middle Schools
Model		
PP	0.879	0.833
MLM0	0.919	0.905
Gain	0.504	0.402
TM	0.466	0.261
SGP	0.595	0.515
VAM	0.649	0.511
Grate	0.503	0.518
AvGrate	0.593	0.484

Relation with school composition variables. We computed the correlation of model estimates with school composition variables to determine whether estimates were related to the aggregated student characteristics in each school. Table 4 shows the correlations of model estimates with school composition variables for mathematics and reading in the elementary school and middle school samples.

The rightmost column of Table 4 shows the average correlation of each school performance model with the school composition variables across all school composition variables. As can be seen, correlations of the status models, PP and MLM0, were negative and noticeably stronger than the correlations of the other school performance models with school composition variables. On average across content and grade level band, the correlation of the school composition variables was -.320 for the PP model and -0.331 for the MLM0 model. In contrast, the average correlations of the school composition variables with the remaining models were quite low ranging from -.032 to +.058. Thus, there was relatively little relation of the multiyear models with school composition, but for status models school performance estimates were higher when fewer students from protected subgroups were present in the school. No clear pattern was present for the relation between school size and model estimates.

Table 4

Correlations of Model Estimates with School Composition Variables by Content Area and Grade Level Band

Elementary School Mathematics

Models	EDS	EL	SWD	Female	Ethnic Minority	School Size	Mean
PP	-0.671	-0.561	-0.236	-0.007	-0.617	0.145	-0.324
MLM0	-0.713	-0.552	-0.201	-0.004	-0.639	0.189	-0.320
Gain	-0.071	-0.016	-0.020	0.028	-0.020	-0.024	-0.020
TM	-0.033	0.015	-0.032	0.030	0.006	-0.032	-0.008
SGP	-0.129	-0.070	-0.099	0.030	-0.078	0.032	-0.052
VAM	-0.195	-0.108	-0.079	0.025	-0.132	0.043	-0.074
Grate	-0.050	0.000	-0.004	0.023	-0.005	-0.043	-0.013
AvGrate	-0.091	-0.029	-0.071	0.023	-0.043	0.052	-0.026

Elementary School Reading/Language Arts

Models	EDS	EL	SWD	Female	Ethnic Minority	School Size	Mean
PP	-0.725	-0.654	-0.236	-0.020	-0.685	0.086	-0.372
MLM0	-0.796	-0.670	-0.217	0.001	-0.748	0.122	-0.385
Gain	0.313	0.290	0.008	0.027	0.343	-0.023	0.160
TM	0.255	0.257	-0.020	0.038	0.290	-0.002	0.136
SGP	-0.029	0.012	-0.070	-0.003	0.003	0.068	-0.003
VAM	-0.095	-0.043	-0.086	0.000	-0.057	0.028	-0.042
Grate	0.412	0.375	0.038	0.034	0.428	-0.054	0.206
AvGrate	0.279	0.276	0.035	-0.030	0.283	-0.027	0.136

Middle School Mathematics

Models	EDS	EL	SWD	Female	Ethnic Minority	School Size	Mean
PP	-0.534	-0.433	-0.366	0.090	-0.537	0.142	-0.273
MLM0	-0.587	-0.457	-0.327	0.097	-0.545	0.167	-0.275
Gain	-0.067	-0.066	-0.125	0.105	-0.056	0.054	-0.026
TM	-0.038	-0.013	-0.061	0.070	-0.021	0.035	-0.005
SGP	-0.013	-0.012	-0.133	0.106	0.005	0.050	0.000
VAM	-0.028	-0.038	-0.122	0.101	-0.026	0.036	-0.013

Grate	-0.223	-0.190	-0.172	0.083	-0.199	0.088	-0.102
AvGrate	0.230	0.164	-0.031	0.088	0.209	-0.037	0.104

Middle School Reading

Models	EDS	EL	SWD	Female	Ethnic Minority	School Size	Mean
PP	-0.645	-0.535	-0.344	0.154	-0.601	0.097	-0.312
MLM0	-0.726	-0.576	-0.314	0.102	-0.676	0.134	-0.343
Gain	-0.098	-0.027	-0.127	0.020	-0.068	0.115	-0.031
TM	0.053	0.076	-0.051	-0.019	0.061	0.079	0.033
SGP	0.067	0.093	-0.123	0.078	0.091	0.098	0.051
VAM	-0.037	0.034	-0.152	0.088	0.008	0.062	0.001
Grate	-0.276	-0.178	-0.196	0.074	-0.233	0.097	-0.119
AvGrate	-0.006	0.066	-0.172	0.122	0.048	0.046	0.017

Relation of model estimates to SWD school composition. Because of the NCAASE emphasis on the performance and academic growth of SWD, we also focused more specifically on the relations between the percentage of SWD students served by a school and the school performance model estimates. Table 5 shows the correlation of model estimates with the percentage of SWD in each school for mathematics and reading in the elementary school and middle school samples averaged over cohorts. As can be seen in the bottom row of Table 5, average school performance estimates based on the single-year, status models (PP and MLM0) had substantially higher correlations with school SWD composition than the other school performance models. With the PP and MLM0 models, school performance estimates were higher the smaller the percentage of SWD students in the school and lower to the extent that the school served larger proportions of SWD.

Table 5

Average School Performance Model Estimates as a Function of the Percentage of SWD in the School by Content Area and Grade Level Band

Content Area and Grade Level Band	PP	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
Math Elementary	-0.236	-0.201	-0.020	-0.032	-0.099	-0.079	-0.004	-0.071
Reading Elementary	-0.236	-0.217	0.008	-0.020	-0.070	-0.086	0.038	0.035
Math Middle	-0.366	-0.327	-0.125	-0.061	-0.133	-0.122	-0.172	-0.031
Reading Middle	-0.344	-0.314	-0.127	-0.051	-0.123	-0.152	-0.196	-0.172
Mean	-0.296	-0.265	-0.066	-0.041	-0.106	-0.110	-0.084	-0.060

Summary of Section A. We evaluated eight alternative models for estimating school academic performance in mathematics and reading using operational state accountability data. We compared the estimates of school performance from one model to another and found substantial disagreement across models. In general, correlations within model type (i.e., single year or multiyear) were stronger than correlations where a status model was paired with a model using multiple years of data.

We also compared school performance estimates in mathematics with those in reading. Again, agreement was greater across content areas for the status models than for the multiple year models. The correlations of the status models (PP and MLM0) with student composition were stronger than the correlations of the multiple year models with student composition. Larger proportions of protected student subgroups were associated with lower school performance. Finally, we correlated school performance estimates with the percentage of SWD in each school. Ideally, estimates of school performance should be unrelated to the student composition of the school, but as with the other school composition variables, we found that the status models were more strongly correlated with SWD school composition than multiyear model estimates.

Section B: School Ranks Based on School Performance Estimates

In this section, we focus on the examination of school ranks based on the school performance estimates reported in the previous section. It is a common practice for states and other jurisdictions to rank schools as a method for evaluating academic performance. Therefore, using the estimates of school performance generated by the eight models, we computed percentile ranks for each school. We then compared the school ranks for each model to the ranks obtained from each of the other models. Finally, we examined the relation between school ranks from each model with variables describing the student composition of each school. Three criteria were used to evaluate the comparisons of school ranks: (a) the Spearman's correlation between school ranks, (b) the proximity of absolute school ranks, and (c) the root mean square difference (RMSD) in school ranks.

Comparison of cohorts. Due to changes in test publisher and state standards in AZ, we only had one cohort for each subject area and grade band and were not able to complete the cohort comparison portion of our study for AZ.

Comparison of models. We were able to compare school ranks from one model to another for AZ. We first computed the Spearman's correlations among school ranks for the different models. These values were quite similar to the Spearman's correlations among school model estimates (see Table 2) and for this reason they are not included in this report. Our second criterion for comparing school ranks was to determine how much a school's rank changed from one model to another. For each pair of school performance models, Table 6 shows the percentage of schools that were within 5, 10, or 20 percentile ranks in one model versus the other. As can be seen in the table, two pairs of models produced results that were quite similar, Gain with Grate, and PP with MLM0. For these two model pairings, over 76% of schools were within 10 ranks of each other and over 94% were within 20 ranks of each other.

When a single year model (PP or MLM) was paired with a model that made use of multiyear results, the level of agreement in school ranks was much lower than when a single year

model was paired with another single year or status model (PP and MLM) or a multiyear model with a multiyear model.

Table 6

Proportion of Elementary or Middle Schools within 5, 10, or 20 Ranks of Each Other for Each Pair of School Performance Models in Mathematics and Reading

Model Comparison:	r = 5	r = 10	r = 20
PP vs. MLM0			
Math Elementary	0.496	0.773	0.951
Reading Elementary	0.520	0.783	0.962
Math Middle	0.612	0.853	0.969
Reading Middle	0.559	0.808	0.947
Mean	0.547	0.804	0.957
PP vs. Gain			
Math Elementary	0.145	0.269	0.445
Reading Elementary	0.103	0.200	0.333
Math Middle	0.174	0.301	0.521
Reading Middle	0.167	0.330	0.550
Mean	0.147	0.275	0.462
PP vs. TM			
Math Elementary	0.149	0.260	0.445
Reading Elementary	0.093	0.180	0.333
Math Middle	0.163	0.303	0.519
Reading Middle	0.145	0.278	0.483
Mean	0.138	0.255	0.445
PP vs. SGP			
Math Elementary	0.165	0.292	0.513
Reading Elementary	0.155	0.267	0.450
Math Middle	0.187	0.301	0.541
Reading Middle	0.171	0.303	0.490
Mean	0.170	0.291	0.498

PP vs. VAM

Math Elementary	0.172	0.320	0.534
Reading Elementary	0.177	0.307	0.495
Math Middle	0.169	0.341	0.539
Reading Middle	0.149	0.265	0.512
Mean	0.167	0.308	0.520

PP vs. Grate

Math Elementary	0.117	0.239	0.411
Reading Elementary	0.102	0.167	0.303
Math Middle	0.207	0.388	0.615
Reading Middle	0.200	0.341	0.599
Mean	0.156	0.284	0.482

PP vs. AvGrate

Math Elementary	0.166	0.283	0.495
Reading Elementary	0.102	0.197	0.326
Math Middle	0.122	0.220	0.396
Reading Middle	0.147	0.243	0.503
Mean	0.134	0.236	0.430

MLM0 vs. Gain

Math Elementary	0.128	0.247	0.452
Reading Elementary	0.097	0.171	0.332
Math Middle	0.174	0.296	0.557
Reading Middle	0.178	0.325	0.532
Mean	0.144	0.260	0.468

MLM0 vs. TM

Math Elementary	0.141	0.248	0.426
Reading Elementary	0.085	0.153	0.318
Math Middle	0.163	0.301	0.510

Reading Middle	0.127	0.229	0.434
Mean	0.129	0.233	0.422

MLM0 vs. SGP

Math Elementary	0.148	0.292	0.520
Reading Elementary	0.127	0.252	0.425
Math Middle	0.147	0.301	0.537
Reading Middle	0.156	0.261	0.461
Mean	0.144	0.276	0.486

MLM0 vs. VAM

Math Elementary	0.190	0.331	0.563
Reading Elementary	0.166	0.290	0.493
Math Middle	0.169	0.332	0.561
Reading Middle	0.178	0.310	0.486
Mean	0.176	0.316	0.526

MLM0 vs. Grate

Math Elementary	0.137	0.231	0.427
Reading Elementary	0.071	0.167	0.297
Math Middle	0.203	0.374	0.657
Reading Middle	0.227	0.383	0.610
Mean	0.160	0.289	0.498

MLM0 vs. AvGrate

Math Elementary	0.161	0.278	0.506
Reading Elementary	0.103	0.187	0.325
Math Middle	0.118	0.232	0.410
Reading Middle	0.156	0.274	0.490
Mean	0.134	0.243	0.433

Gain vs. TM

Math Elementary	0.471	0.704	0.908
Reading Elementary	0.376	0.573	0.822
Math Middle	0.421	0.664	0.871
Reading Middle	0.363	0.579	0.793
Mean	0.408	0.630	0.848

Gain vs. SGP

Math Elementary	0.365	0.579	0.814
Reading Elementary	0.287	0.461	0.724
Math Middle	0.443	0.648	0.893
Reading Middle	0.392	0.575	0.831
Mean	0.372	0.566	0.816

Gain vs. VAM

Math Elementary	0.362	0.578	0.819
Reading Elementary	0.269	0.452	0.706
Math Middle	0.470	0.697	0.933
Reading Middle	0.396	0.666	0.906
Mean	0.374	0.598	0.841

Gain vs. Grate

Math Elementary	0.580	0.816	0.978
Reading Elementary	0.593	0.837	0.976
Math Middle	0.499	0.768	0.973
Reading Middle	0.595	0.833	0.962
Mean	0.567	0.814	0.972

Gain vs. AvGrate

Math Elementary	0.222	0.376	0.599
Reading Elementary	0.225	0.408	0.651
Math Middle	0.238	0.385	0.592
Reading Middle	0.292	0.479	0.753
Mean	0.244	0.412	0.649

TM vs. SGP

Math Elementary	0.315	0.533	0.762
Reading Elementary	0.230	0.416	0.655
Math Middle	0.383	0.566	0.802
Reading Middle	0.305	0.488	0.748
Mean	0.308	0.501	0.742

TM vs. VAM

Math Elementary	0.303	0.496	0.736
Reading Elementary	0.216	0.384	0.647
Math Middle	0.307	0.535	0.788
Reading Middle	0.263	0.452	0.715
Mean	0.272	0.467	0.722

TM vs. Grate

Math Elementary	0.370	0.583	0.868
Reading Elementary	0.319	0.514	0.764
Math Middle	0.359	0.532	0.786
Reading Middle	0.307	0.488	0.715
Mean	0.339	0.529	0.783

TM vs. AvGrate

Math Elementary	0.185	0.332	0.558
Reading Elementary	0.190	0.347	0.567
Math Middle	0.200	0.350	0.568
Reading Middle	0.189	0.365	0.628
Mean	0.191	0.348	0.580

SGP vs. VAM

Math Elementary	0.522	0.786	0.964
Reading Elementary	0.436	0.670	0.904

Math Middle	0.548	0.817	0.971
Reading Middle	0.392	0.646	0.906
Mean	0.475	0.730	0.936

SGP vs. Grate

Math Elementary	0.270	0.436	0.683
Reading Elementary	0.203	0.371	0.618
Math Middle	0.294	0.486	0.739
Reading Middle	0.292	0.477	0.744
Mean	0.265	0.442	0.696

SGP vs. AvGrate

Math Elementary	0.350	0.562	0.811
Reading Elementary	0.243	0.409	0.655
Math Middle	0.339	0.474	0.742
Reading Middle	0.332	0.584	0.833
Mean	0.316	0.507	0.760

VAM vs. Grate

Math Elementary	0.272	0.444	0.675
Reading Elementary	0.213	0.376	0.612
Math Middle	0.330	0.539	0.804
Reading Middle	0.428	0.639	0.860
Mean	0.311	0.500	0.738

Grate vs. AvGrate

Math Elementary	0.172	0.310	0.513
Reading Elementary	0.214	0.350	0.592
Math Middle	0.180	0.301	0.506
Reading Middle	0.283	0.497	0.715
Mean	0.212	0.364	0.582

Our last criterion for comparing school ranks across cohorts was the RMSD between pairs of school performance model rankings. Table 7 shows the RMSD by content area and grade level band. The RMSD values reflect the same patterns of results for models as described

previously. The Gain versus Grate, and PP versus MLM0 pairings produced school rankings that were quite similar.

When a single year model (PP or MLM) was paired with a model that made use of multiyear results, the level of agreement in school ranks was much lower (difference of about 33 ranks on average across all model pairings of this type) than when the two single year models were paired (MLM and PP pairs differed by 9 ranks on average), or a multiyear model was paired with another multiyear model (difference of about 18 ranks, on average).

Table 7

Average of RMSD in School Ranks between School Performance Models by Content Area and Grade Level Band

Elementary School Mathematics

Model	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
PP	9.753	34.704	34.616	30.376	28.814	36.959	31.244
MLM0		34.458	35.309	30.172	27.812	36.841	30.373
Gain			11.73	16.635	15.846	8.073	26.419
TM				18.669	19.148	14.131	27.856
SGP					9.154	22.663	15.657
VAM						22.471	13.091
Grate							32.366

Elementary School Reading

Model	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
PP	9.283	42.308	40.877	33.269	31.630	44.993	41.952
MLM0		43.589	42.727	34.350	32.620	46.185	43.122
Gain			16.274	20.724	19.708	7.971	24.345
TM				23.835	23.464	17.852	27.179
SGP					12.510	25.463	22.291
VAM						25.170	19.797
Grate							28.161

Middle School Mathematics

Model	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
PP	7.948	29.300	30.244	29.455	27.980	25.816	35.901
MLM0		28.582	30.815	29.110	27.370	24.703	35.968
Gain			14.158	12.307	10.866	9.213	25.045
TM				17.358	17.176	16.880	26.967

SGP				8.368	17.962	18.321
VAM					16.595	16.932
Grate						30.892

Middle School Reading

Model	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
PP	9.348	30.138	33.625	33.457	30.828	25.892	31.195
MLM0		30.285	35.750	34.437	31.436	25.664	32.124
Gain			18.820	15.248	11.812	8.607	19.153
TM				20.948	20.757	21.004	24.772
SGP					12.282	18.624	15.018
VAM						13.935	8.519
Grate							20.390

We also evaluated the extent to which school ranks agreed from one content area to the other. Table 8 shows the Spearman's correlation of school ranks in mathematics with school ranks in reading by cohort and grade level band. The table also shows the mean correlation across cohorts at the two grade level bands. As can be seen in Table 8, on average correlations of school ranks across mathematics and reading in elementary schools ranged from about .44 to .91 for the different school performance models. For middle schools, the average correlations ranged from about .31 to .89. Correlations were higher for the two status models, and lower for the multiyear models at both grade level bands. Average correlations at the middle school level were slightly lower than for the elementary level for all models.

Table 8

Spearman's Correlations of School Performance Model Estimates Across Mathematics and Reading

	Elementary Schools	Middle Schools
Model		
PP	0.879	0.825
MLM0	0.911	0.891
Gain	0.472	0.450
TM	0.440	0.315
SGP	0.575	0.503
VAM	0.622	0.526
Grate	0.457	0.520
AvGrate	0.556	0.470

Table 9 shows the proportion of schools that shared similar ranks in mathematics as in reading for each school performance model by school level and averaged over grade level band. Similar to results previously described, Table 9 shows greater agreement for the PP and MLM0 models than the other school performance models with about 80% or more of the schools having ranks within 20 places across grade level bands. In contrast, there was substantially less agreement across the two content areas for the remaining, multiyear models with only approximately 50-60% of schools agreeing within 20 ranks for most models in either grade level band.

Table 9

Proportion of Elementary or Middle Schools within 5, 10, or 20 Ranks of Each Other in Mathematics Versus Reading for Each School Performance Model Averaged

Model Comparison	r = 5	r = 10	r = 20
PP			
Elementary	0.366	0.619	0.874
Middle	0.390	0.579	0.797
Mean	0.378	0.599	0.836
MLM0			
Elementary	0.393	0.648	0.912
Middle	0.423	0.639	0.880
Mean	0.408	0.644	0.896
Gain			
Elementary	0.188	0.313	0.531
Middle	0.223	0.323	0.519
Mean	0.206	0.318	0.525
TM			
Elementary	0.181	0.335	0.523
Middle	0.165	0.314	0.512
Mean	0.173	0.324	0.518
SGP			
Elementary	0.203	0.349	0.573
Middle	0.220	0.356	0.568
Mean	0.212	0.352	0.570
VAM			
Elementary	0.205	0.371	0.617

Middle	0.196	0.334	0.575
Mean	0.200	0.353	0.596
Grate			
Elementary	0.184	0.305	0.535
Middle	0.214	0.363	0.561
Mean	0.199	0.334	0.548
AvGrate			
Elementary	0.199	0.340	0.570
Middle	0.196	0.303	0.530
Mean	0.198	0.322	0.550

Calculation of the RMSD in school ranks for mathematics versus reading by grade level band showed similar results (see Table 10). The difference in school ranks for the PP and MLM0 models ranged from about 12 to 17. Average differences in rank across the two content areas were substantially greater for the remaining models ranging from 25 to 33 depending on model and grade level band.

Table 10

RMSD in School Ranks for Mathematics and Reading by Grade Level Band

	Elementary Schools	Middle Schools
Model		
PP	14.016	16.824
MLM0	12.056	13.314
Gain	29.334	29.848
TM	30.199	33.319
SGP	26.322	28.352
VAM	24.811	27.701
Grate	29.749	27.927
AvGrate	26.874	29.367

Relation with school composition variables. We computed the correlation of school ranks based on each school performance model with school composition variables to determine whether estimates were related to the aggregated student characteristics in each school. Table 11 shows these correlations for mathematics and reading in the elementary school and middle school samples. The rightmost column of Table 11 shows the correlation of each school performance model averaged over all of the school composition variables. As can be seen, average correlations of the status models, PP and MLM0, ranged from -.289 to -.410 depending

on content and grade level band and were noticeably stronger than the correlations of the other school performance models with school composition variables, which ranged from -.120 to +.211 depending on content and grade level band.

Table 11

Spearman's Correlations of School Ranks With School Composition Variables by Content Area and Grade Level Band

Elementary School Mathematics

Model	EDS	EL	SWD	Female	Ethnic Minority	School Size	Mean
PP	-0.695	-0.646	-0.244	-0.011	-0.642	0.123	-0.352
MLM0	-0.711	-0.640	-0.196	-0.023	-0.651	0.182	-0.340
Gain	-0.066	-0.042	-0.046	0.026	-0.025	-0.022	-0.029
TM	-0.040	-0.024	-0.041	0.037	-0.006	-0.033	-0.018
SGP	-0.132	-0.107	-0.107	0.020	-0.092	0.027	-0.065
VAM	-0.190	-0.153	-0.095	0.020	-0.139	0.043	-0.086
Grate	-0.044	-0.029	-0.018	0.025	-0.007	-0.046	-0.020
AvGrate	-0.082	-0.053	-0.079	0.014	-0.043	0.051	-0.032

Elementary School Reading

Model	EDS	EL	SWD	Female	Ethnic Minority	School Size	Mean
PP	-0.746	-0.719	-0.251	-0.019	-0.706	0.068	-0.395
MLM0	-0.804	-0.755	-0.227	-0.022	-0.763	0.111	-0.410
Gain	0.332	0.294	0.008	0.041	0.342	-0.023	0.166
TM	0.258	0.236	-0.014	0.061	0.267	-0.011	0.133
SGP	-0.026	-0.015	-0.084	0.017	-0.005	0.068	-0.008
VAM	-0.087	-0.078	-0.100	0.012	-0.056	0.035	-0.046
Grate	0.421	0.377	0.042	0.053	0.422	-0.049	0.211
AvGrate	0.296	0.281	0.040	-0.020	0.288	-0.034	0.142

Middle School Mathematics

Model	EDS	EL	SWD	Female	Ethnic Minority	School Size	Mean
PP	-0.556	-0.472	-0.333	0.027	-0.553	0.151	-0.289
MLM0	-0.590	-0.505	-0.328	0.031	-0.572	0.171	-0.299

Gain	-0.092	-0.063	-0.144	0.025	-0.072	0.115	-0.038
TM	-0.055	-0.020	-0.095	0.057	-0.027	0.062	-0.013
SGP	-0.024	-0.015	-0.125	0.030	-0.001	0.099	-0.006
VAM	-0.047	-0.050	-0.133	0.020	-0.042	0.094	-0.026
Grate	-0.236	-0.197	-0.195	0.016	-0.224	0.118	-0.120
AvGrate	0.239	0.176	-0.002	0.008	0.226	0.011	0.110

Middle School Reading

Model	EDS	EL	SWD	Female	Ethnic Minority	School Size	Mean
PP	-0.669	-0.576	-0.342	0.090	-0.632	0.080	-0.342
MLM0	-0.734	-0.635	-0.331	0.052	-0.708	0.113	-0.374
Gain	-0.123	-0.057	-0.181	0.054	-0.111	0.160	-0.043
TM	0.041	0.060	-0.086	0.024	0.018	0.095	0.025
SGP	0.064	0.076	-0.151	0.100	0.088	0.129	0.051
VAM	-0.027	0.007	-0.198	0.098	0.007	0.105	-0.001
Grate	-0.267	-0.191	-0.231	0.073	-0.240	0.147	-0.118
AvGrate	0.010	0.027	-0.207	0.128	0.058	0.073	0.015

Relation of school ranks with SWD school composition. We specifically examined the relations between the percentage of SWD students served by a school and the school ranks based on the school performance model. Table 12 shows these correlations for mathematics and reading in the elementary school and middle school samples averaged over cohorts. As can be seen in the bottom row of Table 12, on average, there was a substantially stronger correlation of the status models (PP and MLM0) with school SWD composition than found with the other school performance models. With the PP and MLM0 models, school ranks were higher with lower percentages of SWD students in the school and school ranks were lower as schools served larger proportions of SWD. Little relation was present between school ranks based on the other models and SWD school composition.

Table 12

Average School Rank as a Function of the Percentage of SWD in the School by Model, Content Area, and Grade Level Band

Content Area and Grade Level Band	PP	MLM0	Gain	TM	SGP	VAM	Grate	AvGrate
Math Elementary	-0.244	-0.196	-0.046	-0.041	-0.107	-0.095	-0.018	-0.079
Reading Elementary	-0.251	-0.227	0.008	-0.014	-0.084	-0.100	0.042	0.040
Math Middle	-0.333	-0.328	-0.144	-0.095	-0.125	-0.133	-0.195	-0.002

Reading Middle	-0.342	-0.331	-0.181	-0.086	-0.151	-0.198	-0.231	-0.207
Mean	-0.292	-0.270	-0.091	-0.059	-0.117	-0.132	-0.100	-0.062

Summary of Section B. We evaluated the school ranks arising from eight alternative models for estimating school academic performance in mathematics and reading. As with the school performance estimates described in Section A, substantial variability in school ranks was present. When we compared school ranks arising from one model to school ranks from other models, we found two pairs of models produced similar results across the members of a pair. Those models were Gain with Grate, and MLM intercept (MLM0) with the PP model. In general, pairs of models that combined a status model with a model making use of multiple years of test data showed the most discrepant results.

Comparison of model estimates to school composition variables showed that the status models (PP and MLM0) were more strongly related to school composition than the remaining school performance models. Finally, we correlated school ranks arising from the eight performance models with the percentage of SWD in each school. As with the school performance model estimates, we found the status models were more strongly correlated with SWD school composition, but there was little relation of the other model estimates with the percentage of SWD students in the school.

Conclusion

This report described the results of a large study examining eight alternative methods of estimating school performance. The eight alternative methods were representative of types of models often used in state accountability models, although none were the actual model used in AZ at the time of data collection. We represented school performance in two ways, the actual model estimates and school ranks based on model estimates. In addition to this report, there are reports describing results for the three other states (NC, OR, PA) included in the study. Our primary interest in these comparisons was estimating the impact of cohort and student composition (including the percent of SWD) on school performance estimates, as well examining the extent to which different estimates of school performance correlated with each other.

A number of general conclusions can be drawn from the results of these analyses. First, there was agreement between the two status model estimates (PP and MLM0) that were based on a single year of data, but these two models did not agree with the remaining multiyear models. However, there was substantial agreement of the multiyear models with each other with some variations. In general, the AvGrate model showed the least agreement with the other multiyear models.

We also examined the relation of school performance model estimates with variables describing the student composition of the schools. These results showed a pattern of results that differed between the status and the multiyear models. The two status models had substantially stronger correlations with school composition variables than the multiyear models. This was also true in terms of the percentage of SWD students served by a school. The larger the percentage of SWD in the school, the lower the status model estimates of school performance.

Thus, the results showed different estimates of school performance depending on the model chosen, especially for status versus multiyear models, and stronger relations of status models with the student composition of the school than multiyear models. Taken together, these results suggest the need for substantial caution in the way that school performance models are

used and interpreted. The substantial disagreement among the eight school performance models suggests that the choice of model matters a great deal. This choice should be made very carefully. A single model estimate of school performance may not be trustworthy and may need to be augmented by the results from additional models or metrics of school performance.

References

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics,* 29(1), 37-65.

Betebenner, D. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42-51.

Betebenner, D. W., & Iwaarden, A. V. (2011). *SGP: An textupR package for the calculation and visualization of student growth percentiles* [Computer software manual]. (R package version 0.4-0.0 available at http://cran.r-project.org/web/packages/SGP/)

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289‑328.

CTB/McGraw-Hill (2001). *TerraNova* (2nd edition). Monterey, CA: Author.

CTB/McGraw-Hill (2008). *Arizona's Instrument to Measure Standards: 2008 technical report*. Monterey, CA: Author.

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and Quantile Regression Approaches to Student "Growth" Percentiles. *Journal of Educational and Behavioral Statistics, 38*, 190–215.

Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012, February). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Council of Chief State School Officers: Washington, DC.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *MLM 7: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Tindal, G., Nese, J. F. T., and Stevens, J. J. (2017). Estimating school effects with a state testing program using transition matrices. *Educational Assessment, 22*, 189-204.