CRESST REPORT 729

Jia Wang
David Niemi
Haiwen Wang

PREDICTIVE VALIDITY OF AN ENGLISH LANGUAGE ARTS PERFORMANCE ASSESSMENT

OCTOBER 2007

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Predictive Validity of an English Language Arts Performance Assessment

CRESST Report 729

Jia Wang, David Niemi, & Haiwen Wang
CRESST/University of California, Los Angeles

October 2007

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported in part by the Los Angeles Unified School District Validation Study, PR/Award Numbers LAPT 960496 and LAPA 0000112, and in part by the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research & Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the Los Angeles Unified School District or of the National Center for Education Research, the Office of Educational Research & Improvement, or the U.S. Department of Education.

PREDICTIVE VALIDITY OF AN ENGLISH LANGUAGE ARTS PERFORMANCE ASSESSMENT

Jia Wang, David Niemi, & Haiwen Wang
CRESST/University of California, Los Angeles

Abstract

The main goal of this report is to present evidence on the predictive validity of an English language arts (ELA) performance assessment (PA) administered in Grades 2–9 in a large urban school district. To account for the hierarchical structure of the data (students are nested within schools), we employed hierarchical linear modeling (HLM) to distinguish individual and aggregated explanatory variables. Based on a sub-sample of 5,427 students, we found that students' 2001 ELA PA scores were predictive of their probability of passing the California High School Exit Exam (CAHSEE). We also found a significant correlation between student performances on the ELA performance assessment and other standardized tests. We believe that the ELA PA may be a dependable and useful indicator to identify at-risk students.

Predictive Validity of an English Language Arts Performance Assessment

The main goal of this study is to examine the predictive validity of an English language arts (ELA) performance assessment (PA) that was implemented in a large urban school district, starting the 2000–2001 school year. The ELA PA was administered to students in Grades 2–9. Our report looked at a sub-sample of students who took the ELA PA test as 9th-graders in 2001 and then took the California High School Exit Exam (CAHSEE) as 10th-graders in 2002. The specific research questions were:

1. To what extent is students' performance on the ELA PA test related to their disadvantaged statuses (being a minority and having a low socioeconomic status [SES] etc.)? What are the partial effects of each explanatory variable when the outcome variables are ELA PA scores, Stanford Achievement Test, Ninth Edition (SAT-9) Reading scores, or SAT-9 Mathematics scores? Are these effects consistent for each outcome? Are the proportions between student variance and school variance in ELA PA scores similar to the proportions between SAT-9 Reading and SAT-9 Mathematics scores?
2. Do the ELA PA scores predict students' performance on the CAHSEE? What is the relationship between students' 2001 ELA PA scores and their 2002 CAHSEE scores after controlling for student background variables and other previous achievement measures?

We will briefly describe the development of the PA, refer to relevant literature on assessment and accountability, describe the data and methodology for analysis, summarize statistical results related to student background and validity, and provide our conclusions.

Background on Performance Assessment Development

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) began its collaborative work with a large urban school district on a comprehensive assessment system in 1996. As described by Niemi, Baker, and Sylvester (in press) the purpose of collaboration between CRESST and the large urban school district was to develop assessments that were (a) consistent with state plans to incorporate performance measures in its assessment system, (b) aligned with California ELA standards, (c) capable of providing better focus for standards-based instruction on writing than multiple-choice items, and (d) capable of measuring writing standards more effectively and directly than through existing multiple-choice tests.

To support the design of this new assessment system, CRESST drew on its extensive PA research and development work. Through their 15 years of model-based assessment research, CRESST researchers had shown that assessments designed to provide good models for instructional activities and formative assessment purposes could also be used for summative purposes, and that a model-based approach to designing assessments made it easier and less costly to design assessments for multiple purposes. The development and testing of CRESST's model-based approach are described in greater detail in Niemi, Baker, and Sylvester (in press).

Literature Review

Performance assessments typically ask students to show the processes of their thinking and reasoning so educators can make direct inferences on the nature and depth of students' understanding (Lane, Liu, Ankenmann, & Stone, 1996; Messick, 1994). Linn, Baker, and Dunbar (1991) further stated that both logical and empirical evidence should be presented in order to draw valid inferences from performance assessments. They specified consequential validity and fairness as necessary criteria for evaluating performance assessments.

As pointed out by Messick (1995), performance assessment construct validation can be determined by the relationship between PA scores and other target construct measurements. Past research has found substantial associations for performance assessments with well-established measurements. Examining the Maryland School Performance Assessment Program (MSPAP), Yen and Ferrara (1997) found that the reading, writing, language, and math assessments of MSPAP show a substantial correlation (a = .54 to .78) with the reading,

language, and math assessments of the Comprehensive Tests of Basic Skills, Fourth Edition (CTBS-4). Hooper (1988) also detected significant correlations (a = .36 to .80) between the reading components of the Boder Test of Reading-Spelling Patterns, a performance assessment, and reading subtests from the SAT-9 scores.

Investigating the validity and generalizability of mathematics performance assessment—QUASAR Cognitive Assessment Instrument (QCAI), Lane, Liu, Ankenmann, and Stone (1996); and Messick (1994) detected modest to moderately high correlations (a = .48 to .72) between QCAI scores and the Mathematics Problem Solving and Concepts subtests of the Iowa Test of Basic Skills, Grade 4 (ITBS-4). Yoon and Young (2000) found that the New Standards Science Reference Examination (for middle schools), a standard-based assessment with mostly performance-based items, is moderately correlated with the SAT-9 and Otis-Lennon School Aptitude Test, seventh edition (OLSAT-7) scores, with correlation coefficients of .63 and .60 respectively. The authors concluded that the three assessments ranked student performance in similar ways.

Performance assessments can also have fairly strong predictive validity on future achievements. Davis, Caros, Grossen, and Carnine (2002) found that the score components of a writing benchmark assessment significantly predicted achievement in SAT-9 and High School Exit Exam (HSEE) scores. The function, based on the score components, correctly identified 77% of students in the upper or lower 50th percentiles on the SAT-9 Writing score distribution and 67% of students in the upper or lower 50th percentiles on the HSEE Writing score distribution.

Data

The data for this study came from a sub-sample of 5,427 students who took the ELA PA test and SAT-9 Reading and Mathematics test in Spring 2001 as 9th-graders, and then took the CAHSEE ELA test in Spring 2002 as 10th-graders. The passing rate was 47% for 10th-graders in Spring 2002, with 7,128 students passing and 7,953 students failing to pass. This data from 12,081 students was reduced down to approximately 9,000 when matched up with the Spring 2001 demographic files and the SAT-9 test file, then reduced down to approximately 6,400 when we further excluded students without 2001 ELA PA scores, and then finally reduced down to 5,427 students when we excluded the students for whom we did not have school characteristics variable information.

Table 1 presents the demographic characteristics of the students used in the analysis. Hispanic students (81.1%) made up the majority of students and 79.4% of the students were English language learner (ELL) students or former ELL students. Please note that ELL

students had to be at English language development (ELD) Level 5 in order to take the ELA PA test. Immigrants made up 26.2% of the students and 80.5% of the students spoke Spanish or a language other than English at home. Students who received free or reduced-fee lunch at school made up 73.1% and 74.6% of the students were Title 1 recipients. Less than 4.2% of the students were in special education programs or were classified as gifted.

The scale for student scores for the ELA PA test was: 1 (*not proficient*), 2 (*partially proficient*), 3 (*proficient*), and 4 (*advanced*). The district set the passing score for the ELA PA test at 2 (*partially proficient*) for the 2000–2001 academic year. Table 1 shows the overall ELA PA passing rate at approximately 70%, with 51.8% of the students scoring 2 (*partially proficient*), 16.5% of the students scoring 3 (*proficient*), and only 2.1% of the students scoring 4 (*advanced*). For the CAHSEE ELA tests administered in Spring 2002, the overall passing rate for our sample was 54.7%. Table 1 also presents the mean of 2001 SAT-9 Reading and SAT-9 Mathematics scores in normal curve equivalents (NCE), as well as the mean of ELA PA scores, and the proportion of students passing the 2002 CAHSEE. This table shows some preliminary comparisons among sub-groups.

Table 1 also shows the mean SAT-9 Reading scores varying from 16.07 NCE for special education students to 43.32 NCE for gifted students. Similarly, there was a significant amount of variation among sub-groups in the mean SAT-9 Mathematics scores, varying from 29.56 NCE for special education students to 57.48 NCE for gifted students. The mean ELA PA scores range from 1.36 for special education students to 2.42 for gifted students. For the percentage of students passing the 2002 CAHSEE, special education students again had the lowest passing rate at 9%, whereas students who scored 4 (advanced) on the 2001 ELA PA test had the highest passing rate at 92%.

Table 1

Student Distribution on Demographic and Other Variables (N = 5,427)

Definition	Value label	(P) of students	(N) of students	(M) SAT-9 Reading score	(M) SAT-9 Math score	(M) PA score	Students passing CAHSEE
Gender	Male	51.1	2,774	29.03	41.30	1.86	53%
	Female	48.9	2,653	28.05	39.22	1.97	57%
Ethnicity	Asian	2.4	128	27.80	48.72	1.95	63%
	Black, not Hispanic	9.6	519	27.47	36.24	1.82	45%
	Hispanic	81.1	4,399	28.22	40.09	1.92	54%
	White, not Hispanic	4.8	263	34.68	45.40	2.00	71%
	All other	2.2	118	32.48	44.83	1.92	69%
English proficiency	ELL	32.2	1,750	21.96	36.58	1.71	29%
	Former ELL	47.2	2,564	31.01	42.58	2.02	68%
	EP	20.5	1,110	31.20	40.05	1.93	59%
Immigrant	Non-immigrant	73.8	4,003	29.00	40.27	1.93	56%
	Immigrant	26.2	1,424	27.27	40.31	1.86	51%
Home language survey	Other	5.9	319	30.11	46.48	1.96	65%
	English	19.5	1,057	30.51	39.03	1.93	56%
	Spanish	74.6	4,051	27.91	40.12	1.91	53%
Title 1	Non-Title 1	25.4	1,379	30.51	41.13	1.90	58%
	Title 1	74.6	4,048	27.88	39.99	1.92	54%

(*table continues*)

Table 1 (continued)

Definition	Value label	(*P*) of students	(*N*) of students	(*M*) SAT-9 Reading score	(*M*) SAT-9 Math score	(*M*) PA score	Students passing CAHSEE
Meal program	Normal	26.9	1,462	30.23	40.05	1.92	58%
	Free or reduced-fee	73.1	3,965	27.93	40.37	1.91	53%
Special education	Non-special education	98.7	5,358	28.71	40.42	1.92	55%
	Special education	1.3	69	16.07	29.56	1.36	9%
Gifted	Non-gifted	97.1	5,269	28.10	39.77	1.90	54%
	Gifted	2.9	158	43.32	57.48	2.42	87%
2001 PA scores	Not proficient	29.6	1,604	24.69	36.79		38%
	Partially proficient	51.8	2,809	28.69	40.58		56%
	Proficient	16.5	898	33.53	44.34		75%
	Advanced	2.1	116	39.93	49.91		92%
2002 CAHSEE (ELA)	Not passing	45.3	2,458	21.83	35.91	1.69	
	Passing	54.7	2,969	34.11	43.90	2.10	

Note. SAT-9 = Stanford Achievement Test, Ninth edition. PA = Performance assessment. CAHSEE = California High School Exit Exam. ELL = English language learner. EP = English proficient. ELA = English language arts.

Table 2

Descriptive Information for Achievement Measures

Variable definition	*M*	*SD*
All students (*N* = 5,427)		
Spring 2001 SAT-9 NCE Reading total scores	28.55	11.42
Spring 2001 SAT-9 NCE Mathematics total scores	40.28	12.15
2001 GPA	2.30	0.72
Spring 2001 PA scores	1.91	0.73
Students not passing the 2002 ELA CAHSEE (*N* = 2,458)		
Spring 2001 SAT-9 NCE Reading total scores	21.83	8.68
Spring 2001 SAT-9 NCE Mathematics total scores	35.91	10.25
2001 GPA	2.12	0.72
Spring 2001 PA scores	1.69	0.65
Students passing the 2002 ELA CAHSEE (*N* = 2,969)		
Spring 2001 SAT-9 NCE Reading total scores	34.11	10.37
Spring 2001 SAT-9 NCE Mathematics total scores	43.90	12.40
2001 GPA	2.45	0.69
Spring 2001 PA scores	2.10	0.75

Note. SAT-9 = Stanford Achievement Test, Ninth edition. NCE = Normal curve equivalents. GPA = Grade point average. PA = Performance assessment. ELA = English language arts. CAHSEE = California High School Exit Exam.

Table 2 provides descriptive information on students' test scores and grade point averages (GPAs). Students had a mean score of 28.55 NCE on the SAT-9 Reading test, about 40.3 NCE on the SAT-9 Mathematics test, 1.91 for the 2001 spring PA test, and 2.3 for GPA. Table 2 also includes the means and standard deviations of these variables by students' CAHSEE ELA results. As indicated in the table, students who passed the CAHSEE ELA had higher scores on all four measures used in the analysis.

Figure 1 shows cross-tabulation information in percentages between 2001 ELA PA scores and 2002 CAHSEE ELA results. We found that students who scored higher on the ELA PA test in 2001 also had higher passing rates on the 2002 CAHSEE ELA. For example, 92.2% of the 9th-graders who scored 4 (*advanced*) on their 2001 ELA PA test also passed the CAHSEE in 2002, whereas 37.5% of the students who scored 1 (*not proficient*) on their 2001 ELA PA test later passed the CAHSEE ELA. (The latter somewhat high result suggests that there may be scaling or difficulty differences between the ELA PA test and the CAHSEE ELA, or some students may have improved their skills in the year between the tests.)

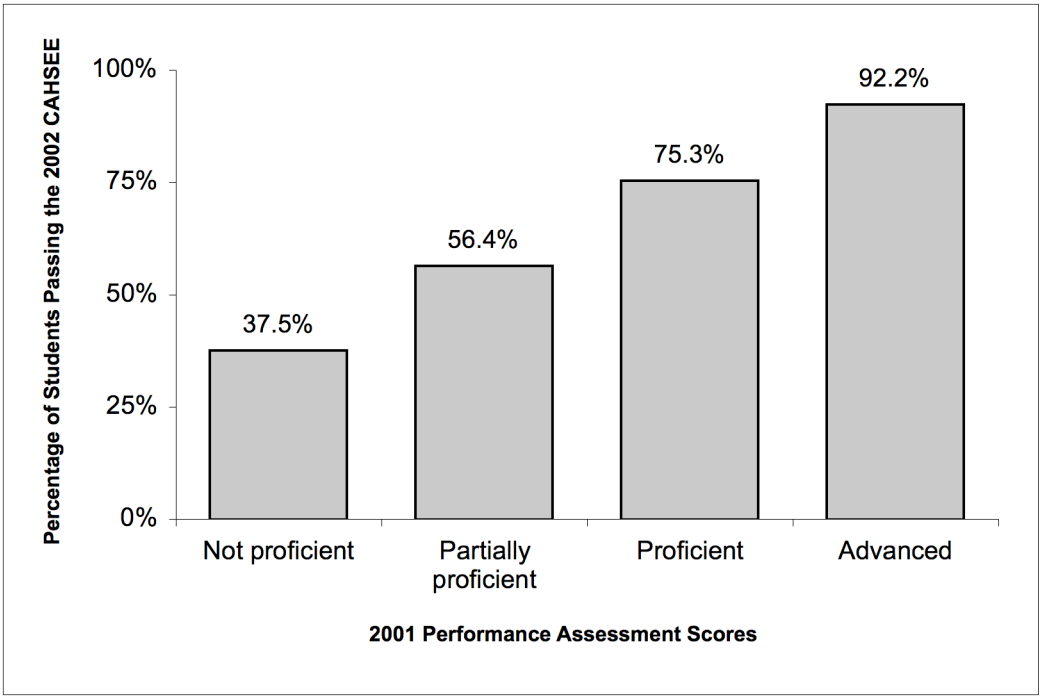

Figure 1. Percentage of 10th-grade students passing CAHSEE ELA as predicted by their 9th-grade ELA PA scores (N = 5,427).
Note. CAHSEE = California High School Exit Exam. ELA = English language arts.

Table 3 contains the means and standard deviation values for the four school variables we used in the analysis. The results were based on 50 schools. The average class size was about 27 students per class. The average school size was approximately 3,200 students, ranging from 1,020 to 5,140 students. The 50 schools we included in the analysis varied substantially in the mean percentage of students receiving free or reduced-fee lunch. Some schools had 85% of their students in the free or reduced-fee lunch program, whereas some schools had about 11%. The schools also differed a great deal in their 2001 Similar Schools rankings, ranging from 1 through 9 on a scale of 1 to 10. Similar Schools ranking system places each school relative to other schools in California.

Table 3

Means and Standard Deviations of School Level Variables (N = 50)

Variables	*M*	*SD*
Average class size	27.19	1.63
% of students in lunch program	0.56	0.21
Similar schools rank in 2001	2.88	2.09
School enrollment size (in 1,000s)	3.20	0.88

Table 4

Pearson Correlation Coefficients Among Measures

	2001 PA	2001 Grade point average	2001 SAT-9 Reading	2001 SAT-9 Mathematics
2001 PA	1.00	0.22**	0.29**	0.24**
2001 Grade point average	0.22**	1.00	0.20**	0.28**
2001 SAT-9 Reading	0.29**	0.20**	1.00	0.46**
2001 SAT-9 Mathematics	0.24**	0.28**	0.46**	1.00

** Correlation is significant at the 0.01 level (2-tailed).
Note. PA = Performance assessment. SAT-9 = Stanford Achievement Test, Ninth edition.

The correlation coefficients reported in Table 4 indicates that the achievement measures were only marginally, but significantly correlated, in the approximate range of 0.30, with one exception: The coefficient between SAT-9 Reading and SAT-9 Mathematics was 0.46, much higher than the others reported in the table. Despite the fact that teachers rated their own students' PA scores, its correlation coefficients with GPA, SAT-9 Reading, and SAT-9 Mathematics scores were similar (approximately 0.30). In addition, the correlations between GPA and the SAT-9 scores were also modest. These modest coefficients were similar to other research (e.g. Hooper, 1988).

It is important to note that due to the limited range of the scale for the ELA PA scores and GPA, the correlations involving either ELA PA scores or GPA were attenuated. Furthermore, the correlations did not reflect potential non-linear relationships among these measures. This particularly was the case for ELA PA scores, as they were not likely to be continuous: that is, the difference in performance between scores 1 and 2 was not necessarily the same as the difference in performance between scores 2 and 3. Although simple correlations and cross-tabulations provided some evidence of correlation between measures, they could not provide a complete picture of the relationship among the measures. Therefore we adopted additional methodology.

Methodology

Given the complex nature of the research questions and the data itself, it is important to use appropriate methodological techniques. HLM is one approach for analyzing the relationship between different achievement measures. This is due to the natural structure of the data, which is hierarchical: that is, students attend (are nested within) schools. Although students are the unit of analysis, school context is also an important aspect to investigate. Taking this naturally nested data structure into account is important because mixing

individual and aggregated explanatory variables can lead to both statistical and substantive errors in the interpretation of group effects (Aitkin & Longford, 1986; Bryk & Raudenbush, 2002; Burstein, 1980). For example, a student's native language may limit opportunity to learn (OTL) if the teacher is not teaching in a language the student understands; but when a student's native language categories are aggregated to the classroom or school level, they become an indicator of school language diversity and the normative environment (Burstein, 1980).

Group effects may be important because students with the same characteristics may have different learning outcomes if they attend schools with different quality, policies, organization, and practices (Akin & Garfinkel, 1977). Hence, we consider school context in the model, while also accounting for differences in mean school achievement due to differences in enrollment among schools. Ordinary Least Square (OLS) regression cannot accomplish this task. Furthermore, if there is a large variance in test scores (whether it is PA, SAT-9, or CAHSEE) attributable to differences between schools, OLS regression analysis will severely understate standard errors and overestimate the significance of parameter estimates, thereby leading to falsely rejecting null hypotheses.

Goldschmidt and Martinez-Fernandez (2002) examined whether ELA PA scores need to be analyzed as separate categories or can be treated as a continuous variable. They found that ELA PA scores behave linearly with respect to SAT-9 Reading scores. Therefore we would treat ELA PA scores as a continuous variable even though the scores only range from 1 (*not proficient*) to 4 (*advanced*). The other outcome variable (besides 2001 ELA PA scores and SAT-9 Reading scores) analyzed in this report is whether or not a student passes the CAHSEE ELA. We adopted the logistic model to accommodate the fact that this outcome is binary (passing or not passing). In light of these analyses, we use HLM when the outcome variables are ELA PA scores and SAT-9 scores, and we use logistic HLM when the outcome variable is CAHSEE ELA.

HLM Results on Student Background

Research question 1. We address the question of how students' ELA PA scores related to their disadvantaged status by looking at both variance partition and effects of student and school variables across all three achievement measures. Table 5 presents the variance component values for both student and school level models, for SAT-9 test scores and ELA PA scores. Table 5 also shows the variation between students and schools, and the percentage of variance reduction due to the explanatory variables in the model specification. The variance components themselves cannot be directly compared due to different measures

having different scales. However, the variance partitioning into percentages was directly comparable.

The variation found in ELA PA scores was mainly associated with student characteristics (92.3%) and marginally with school context (7.7%). Compared to SAT-9 Reading and SAT-9 Mathematics scores, ELA PA scores were relatively slightly more homogeneous between students than the SAT-9 scores, judging by the proportion of variance attributable to students. Because teachers scored their own students' ELA PA tests, we found a relatively larger variation between schools in ELA PA scores than SAT-9 scores. These two small differences could also be caused by the general unreliability of ELA PA scores. With that said, the amount of variation found in the student and school level for ELA PA scores differ only by about 2% with SAT-9 Mathematics scores, and by about 1% with the SAT-9 Reading scores. These differences were not substantially large enough for concern or attention. We conclude that the variance partitions of these three achievement measures are similar and that the ELA PA test was as valid an assessment has similar differentiating power to SAT-9 Reading and SAT-9 Mathematics tests, at least in terms of score variation between students and schools. This gives evidence regarding the validity of the ELA PA test.

The last three columns in Table 5 reports the percentages of variance reduced with predictors by including the student and school variables in the estimation. The combination of student variables explains 6.1% of the variation found in ELA PA scores; one-third and one-half of what were found for SAT-9 Reading scores (18%) and SAT-9 Mathematics scores (12.5%), respectively. The differences are much larger at the school level. The four school variables we used explained 65.3% between school variation for SAT-9 Reading scores, 47.3% for the SAT-9 Mathematics scores, and only 9.8% for ELA PA scores. The results may have the following three explanations:

1. Although we found similar proportions of variation in ELA PA scores between students and schools (as with SAT-9 scores), we needed a different set of explanatory variables to explain the found variance in ELA PA scores.
2. This could also imply that teachers who scored their own students' ELA PA tests were potentially equalizing scores among their students. That is not to say that teachers were artificially raising student scores, but rather that the teachers took student circumstances into account when rating the tests.
3. The last explanation may be that the ELA PA test was a more egalitarian test than the SAT-9 test, as these traditionally used variables did not have much explanatory power over the ELA PA test.

Table 5

Variance Component Results for SAT-9 Reading, SAT-9 Math and PA Scores

Variance Component	Variance components						Variance reduced with predictors		
	Without predictors			With predictors					
	SAT-9 Reading	SAT-9 Math	PA	SAT-9 Reading	SAT-9 Math	PA	SAT-9 Reading	SAT-9 Math	PA
Level 1 - student									
Between student variation	123.65	138.97	0.49	101.42	121.55	0.46			
Proportion of variance attributable to students	93.4%	94.4%	92.3%						
% variance reduced due to student variables							18.0%	12.5%	6.1%
Level 2 - school									
Between school variation	8.73	8.19	0.04	3.03	4.32	0.04			
Proportion of variance attributable to schools	6.6%	5.6%	7.7%						
% variance reduced due to school variables							65.3%	47.3%	9.8%

Note. SAT-9 = Stanford Achievement Test, Ninth edition. PA = Performance assessment.

Table 6

HLM Results on Performance Assessment's Fairness

Independent Variables	Coefficients			Effect size		
	SAT-9 Reading (*SE*)	SAT-9 Math (*SE*)	PA (*SE*)	SAT-9 Reading (*SE*)	SAT-9 Math (*SE*)	PA
School-level variables						
School average	22.38 (5.39)	36.58 (5.33)	1.80 (0.53)			
Average classroom size	0.32 (0.19)	0.15 (0.18)	0.00 (0.02)	0.03	0.01	0.00
% of students in lunch program	4.31 (2.47)	4.39 (2.94)	0.31 (0.21)	0.38	0.36	0.43
Similar schools rank in 2001	0.60* (0.18)	0.52* (0.21)	0.01 (0.02)	0.05	0.04	0.02
School enrollment size	0.19 (0.40)	-0.03 (0.45)	0.01 (0.04)	0.02	0.00	0.01
Student-level variables						
Female	-0.80* (0.26)	-2.10* (0.31)	0.10* (0.02)	-0.07	-0.17	0.13
Ethnicity-Black	-5.32* (1.09)	-5.57* (1.04)	-0.06 (0.06)	-0.47	-0.46	-0.08
Ethnicity-Hispanic	-1.81 (0.99)	-3.35* (1.40)	0.07 (0.06)	-0.16	-0.28	0.09
Ethnicity-Asian	-3.82* (1.22)	3.16 (1.90)	-0.05 (0.09)	-0.33	0.26	-0.07
Ethnicity-other	-1.65 (1.31)	0.33 (1.90)	0.06 (0.07)	-0.14	0.03	0.08
English language learner	-7.19* (0.72)	-4.03* (0.69)	-0.19* (0.04)	-0.63	-0.33	-0.25
Re-designated fluent EP	1.25 (0.65)	0.90 (0.70)	0.10* (0.04)	0.11	0.07	0.14
Home language-Spanish	-1.24 (0.85)	1.69 (0.89)	-0.12* (0.05)	-0.11	0.14	-0.17

(table continues)

Table 6 (continued)

	Coefficients			Effect size		
	SAT-9 Reading (*SE*)	SAT-9 Math (*SE*)	PA	SAT-9 Reading (*SE*)	SAT-9 Math (*SE*)	PA
Home language-other	-0.48 (1.77)	2.91 (1.53)	0.02 (0.07)	-0.04	0.24	0.02
Immigrant status	-0.11 (0.27)	0.15 (0.41)	-0.04* (0.02)	-0.01	0.01	-0.05
Free or reduced-fee lunch	-1.29* (0.39)	0.41 (0.45)	-0.04 (0.02)	-0.11	0.03	-0.06
Title 1	-1.87* (0.69)	-1.83* (0.77)	-0.05 (0.03)	-0.16	-0.15	-0.07
Special education	-10.04* (1.13)	-8.73* (1.72)	-0.42* (0.10)	-0.88	-0.72	-0.58
Gifted	11.67* (1.62)	14.84* (1.72)	0.40* (0.06)	1.02	1.22	0.54

* Statistically significant at .05 level.
Note. HLM = Hierarchical linear modeling. SAT-9 = Stanford Achievement Test, Ninth edition. PA = Performance assessment. EP = English proficient.

Table 6 summarizes the HLM results using the same set of student and school variables to explain ELA PA scores and SAT-9 scores. The first three columns present the coefficients and standard errors (in parentheses), and whether or not the coefficient was statistically significant. The last three columns provide the corresponding effect sizes of the coefficients for compatibility. Using the results reported in Table 6, we compare whether the variables had consistent and similar effects across these three outcome measures using the effect size values and statistical significant signs.

At the school level, none of the school variables were significant for ELA PA scores. Similar Schools rank variable was found to be a significant predictor of SAT-9 Reading and SAT-9 Mathematics scores. This may be due to the fact that SAT-9 scores are used in schools where Similar Schools ranks are calculated.

At the student level, ELL designation and special education designation had a negative, significant effect on students' performances on all three measures, whereas a gifted designation had a positive, significant effect. Being female was associated with lower SAT-9

scores, but associated with higher ELA PA scores. This finding differed from the traditional gender gap in favor of males in school achievement beyond elementary schools. The female-in-favor gender effect had a size of 0.13, meaning that females performed 0.13 standard deviations higher than males, holding all other variables constant.

Expectedly, as an ELA assessment, PA scores were sensitive to students' language skills. We found that a former ELL designation, speaking Spanish at home, and an immigrant designation all had a significant effect on students' ELA PA scores, but no effect on SAT-9 scores. Unexpectedly, student ethnicity and family (SES) indicators (lunch program and Title 1 status) were found to be insignificant predictors of students' ELA PA scores, but both of these indicators were significant predictors of SAT-9 Reading scores. Title 1 status was a significant predictor of SAT-9 Math scores. These findings provide some evidence that the ELA PA test was an accurate measure of student ability and fairly indifferent to these typical, significant, demographic variables.

One may argue that the insensitivity of ELA PA scores to student ethnicity and family SES indicators suggests that the ELA PA test may not be very differentiating. ELA PA scores only have four categories, and the background predictors explained a much lower percentage of the variation in ELA PA scores than in SAT-9 scores. On the other hand, we did find ELA PA scores to be quite sensitive to student language skills, although not to student ethnicity and family SES. Nevertheless, this alone cannot guarantee that the ELA PA test is as valid as the SAT-9 test, although it suggests the possibility that the ELA PA test may be more fair than the SAT-9 test for disadvantaged students. Studies comparing more comprehensive results for the ELA PA test with the SAT-9 or other well-established tests are warranted.

HLM Results on Predictive Validity

Research question 2. Are ELA PA scores predictive of passing the CAHSEE ELA? We address this question from the following two aspects: What is the relationship between students' 2000–2001 ELA PA scores and their 2001–2002 CAHSEE ELA results? Do ELA PA scores provide additional effect on students' performances in CAHSEE ELA results?

Table 7

HLM Results on Performance Assessment's Predictive Validity

Independent Variables	β	*SE*	Log odds
School-level variables			
Intercept	-5.664*	1.169	0.00
Average class size	-0.009	0.040	0.99
% of students in lunch program	0.103	0.389	1.11
Similar Schools rank in 2001	0.023	0.037	1.02
School enrollment size	0.215*	0.081	1.24
Student-level variables			
PA 2001	0.385*	0.042	1.47
SAT-9 Reading 2001	0.123*	0.006	1.13
SAT-9 Math 2001	0.021*	0.004	1.02
GPA 2001	0.365*	0.067	1.44
Female	0.382*	0.080	1.46
Ethnicity-Black	-0.745*	0.304	0.47
Ethnicity-Hispanic	-0.394	0.313	0.67
Ethnicity-Asian	-0.002	0.209	1.00
Ethnicity-other	-0.117	0.255	0.89
ELL	-0.397*	0.121	0.67
Re-designated ELL	0.369*	0.108	1.45
Home language-Spanish	-0.188	0.162	0.83
Home language-other	-0.278	0.296	0.76
Immigrant status	-0.118	0.076	0.89
Free or reduced-fee lunch	-0.174	0.091	0.84
Title 1	-0.129	0.120	0.88
Special education	-1.493*	0.381	0.22
Gifted	0.152	0.193	1.16

*Statistically significant at .05 level.

Note. HLM = Hierarchical linear modeling. PA = Performance assessment. SAT-9 = Stanford Achievement Test, Ninth edition. GPA = Grade point average. ELL = English Language Learner.

As summarized in Table 7, the 2001 ELA PA score variables had a statistically significant effect (*) on a student's possibility of passing the CAHSEE ELA, even after controlling for other student- and school-level variables. At the school level, school enrollment size had a positive effect in a student's possibility of passing the CAHSEE ELA (0.215). At the student level, we found that females and re-designated ELL students had a

higher probability of passing the ELA CAHSEE (.382, .369 respectively), whereas Black students, ELL students, and students enrolled in special education programs were less likely to pass the ELA CAHSEE (-.745, -.397, -1.493, respectively). As expected, higher prior 2001 GPA and higher prior 2001 scores in SAT-9 tests increased a student's potential of passing the ELA CAHSEE (.365, .123, .021, respectively).

To further investigate any additional effect in the relationship between the ELA PA test to a student's probability of passing the ELA CAHSEE, we proceeded to calculate the passing probabilities for students at all four possible score points of the ELA PA test. In the calculation, we assumed the student to be the following: (a) White, (b) male, (c) not an immigrant, (d) proficient in English, (e) English spoken at home, (f) pays for lunch, (g) not enrolled in any special education programs, (h) not enrolled in any Title 1 program, and (i) not classified as gifted. It is also assumed that the student (j) scores at the mean level in GPA and (k) the SAT-9 tests, in addition to being enrolled in a school with all school variables at their mean values. We found that the probability of our example student to pass the ELA CAHSEE is 62% with a prior ELA PA score of 1 (*not proficient*); 71% with a prior ELA PA score of 2 (*partially proficient*); 78% with a prior ELA PA score of 3 (*proficient*); and 84% with a prior ELA PA score of 4 (*advanced*).

Figures 2 and 3 relax the requirement that the student scores at the mean SAT-9 Reading and Mathematics, respectively. Figure 2 shows the student's expected probability of passing the 2002 ELA CAHSEE in response to their 2001 SAT-9 Reading scores and different score points on the ELA PA test. These four prediction lines converge when the student scores 68 or higher in the SAT-9 Reading test. The student would pass the ELA CAHSEE even if the student was rated as "not proficient" in the ELA PA test (Please note here that the mean SAT-9 Reading score in our data was 28.55 with a standard deviation of 11.42. About 96% of the students scored in the range of 5.71 and 51.39. It is quite difficult to score 68 or higher on the SAT-9 Reading test.).

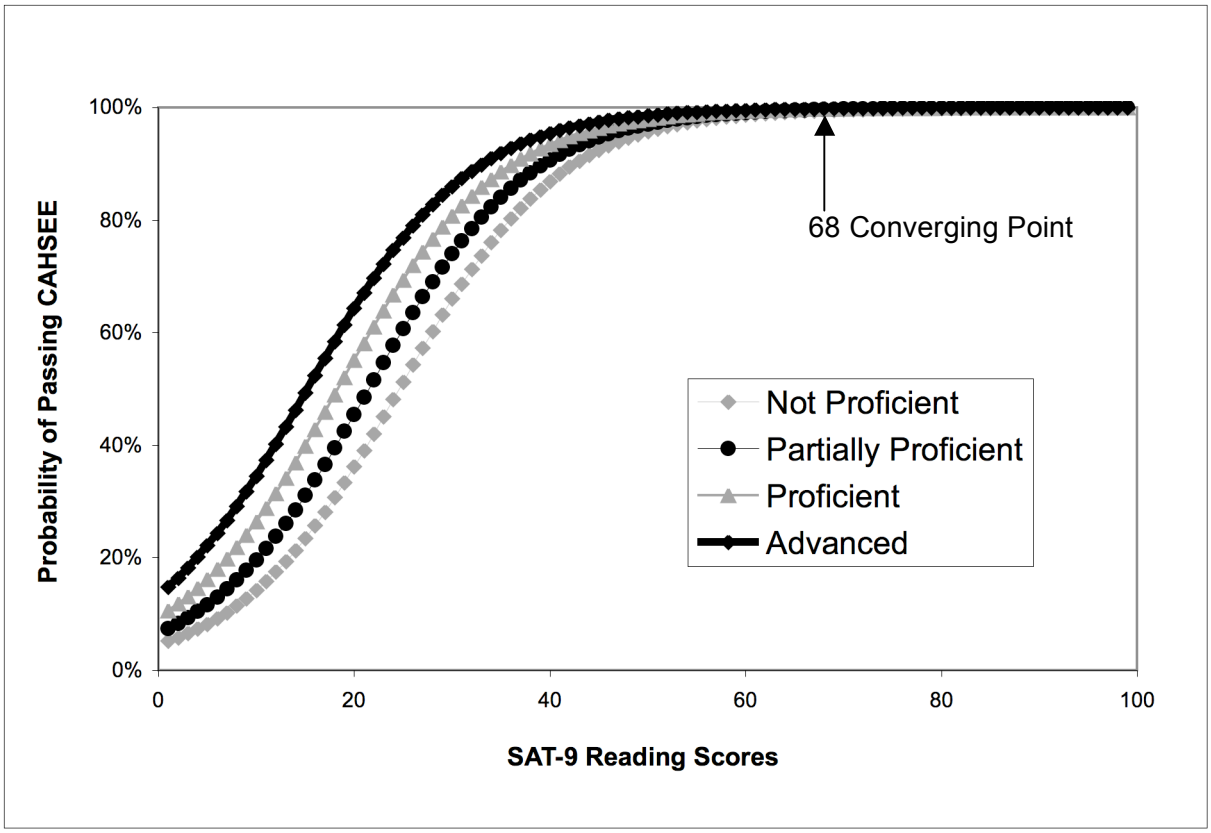

Figure 2. Probability of passing ELA CAHSEE by SAT-9 Reading scores.

The calculation was done assuming the student to be the following: (a) White, (b) male, (c) non-immigrant, (d) proficient in English, (e) English spoken at home, (f) pays for lunch, (g) not enrolled in special education, (h) non-Title 1, (i) not classified as gifted, (j) scores at the mean level in GPA, (k) scores at the mean level in SAT-9 tests, (l) enrolled in a school with mean school variables.
Note. CAHSEE = California High School Exit Exam. SAT-9 = Stanford Achievement Test, Ninth edition, ELA = English language arts, PA = Performance assessment. GPA = Grade point average.
Not Proficient = PA score 1, Partially Proficient = PA score 2, Proficient = PA score 3, Advanced = PA score 4.

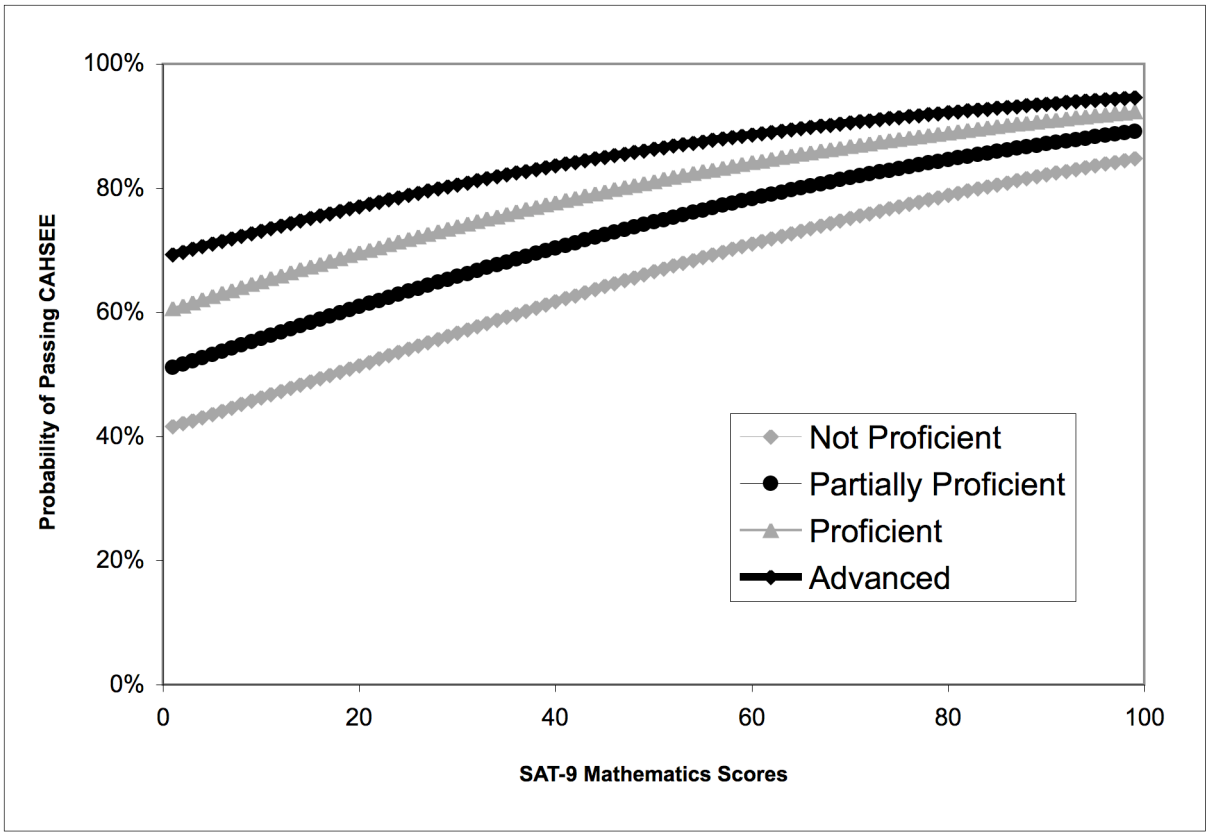

Figure 3. Probability of passing ELA CAHSEE by SAT-9 Mathematics scores.

The calculation was done assuming the student to be the following: (a) White, (b) male, (c) non-immigrant, (d) proficient in English, (e) English spoken at home, (f) pays for lunch, (g) not enrolled in special education, (h) non-Title 1, (i) not classified as gifted, (j) scores at the mean level in GPA, (k) scores at the mean level in SAT-9 tests, (l) enrolled in a school with mean school variables.
Note. CAHSEE = California High School Exit Exam. SAT-9 = Stanford Achievement Test, Ninth edition, ELA = English language arts, PA = Performance assessment. GPA = Grade point average.
Not Proficient = PA score 1, Partially Proficient = PA score 2, Proficient = PA score 3, Advanced = PA score 4.

Figure 3 has the corresponding information in response to the SAT-9 Mathematics scores. The base passing probability for students who scored 1 on the SAT-9 Mathematics test was 42% if the student was rated as *not proficient*; 51% if *partially proficient*; 61% if *proficient*; and 69% if the student was rated *advanced* on the ELA PA scores. Note that the student is never predicted to have a 100% probability of passing the ELA CAHSEE. The predicted probability is 95%, even if the student scored 99 on the SAT-9 Mathematics test. This could imply that the ELA PA scores were more highly correlated to SAT-9 Reading scores than SAT-9 Mathematics scores, once we controlled for other student and school variables.

Summary and Conclusions

This report investigates the predictive validity of CRESST language arts performance assessment implemented in a large, urban school district. Using HLM to take the hierarchical structure of the data into account, we distinguished student and aggregated school explanatory variables, and therefore improved the estimation accuracy. The analysis was based on students who took the ELA PA test as 9th-graders in Spring 2001 and the CAHSEE ELA as 10th-graders in Spring 2002.

Research question 1. The results suggest that the ELA PA test to is not sensitive to students' disadvantaged status, judging by the evaluation of the variance components and partition, and also how student and school variables relate to ELA PA scores. Specifically, we found a similar proportion of variance between students and schools among the three achievement measures we examined: ELA PA scores, SAT-9 Reading, and SAT-9 Mathematics scores. The same set of student background variables and school context variables was less related to the variance found in ELA PA scores, than to SAT-9 test scores. We found no ethnicity and family SES effects on ELA PA scores, and as expected, ELA PA scores were sensitive to student variables associated with English language proficiency, home language, and immigrant status. These were essential pieces of evidence to suggest that the ELA PA test measures student ability and is indifferent to these typical, significant, demographic variables. On the other hand, the insensitivity to these variables may also suggest the overall insensitivity of the ELA PA test to student background and academic skills alike. Therefore further studies investigating the sensitivity of the ELA PA test to student language skills and other aptitudes are warranted.

Research question 2. The predictive validity results also suggest that ELA PA scores predict students' performance in the following year's CAHSEE ELA, even after controlling for student and school characteristic variables. Our study indicates that our example student's (see page 15) probability of passing the CAHSEE ELA was 62%, 71%, 78%, and 84% respective of the student scoring a 1, 2, 3, or 4 in the previous year's ELA PA test. In other words, the probability of passing the CAHSEE ELA improved by at least 6% with each one-point improvement on the ELA PA test. This is a significant effect of the ELA PA test on the exit exam results.

In summary, the results of this study suggest that the ELA PA test was indifferent to typical, significant, demographic variables including ethnicity and family SES. ELA PA scores also significantly predicted students' later performance on the CAHSEE ELA, even after controlling for multiple student and school variables. This suggests that the district

could use ELA PA scores as an early indicator of students' CAHSEE ELA performance in addition to other more traditional indicators. Furthermore, ELA PA scores could also help districts identify students who are at risk of failing the CAHSEE ELA and provide specific interventions or resources to help at-risk students prepare for the CAHSEE ELA.

There are several limitations to this study. First, the study uses only 1-year's worth of data of students with ELA PA scores in one year and CAHSEE ELA scores in the following year. This may limit the generalizability of the ELA PA's predictive validity. At the time of the study, not all students were required to take the CAHSEE ELA in the 10th grade. Consequently, the students in this study are more likely to be higher performing students. Secondly, instead of centralized raters, individual teachers rated the student responses on the ELA PA tests. This lowers the rater reliability of the ELA PA scores used for analysis. However, the other technical reports published from this same project indicate that the level of agreement between centralized raters was of an acceptable level (Goldschmidt & Martinez-Fernandez, 2002). The third caveat is that we investigated the relationships between student background and ELA PA scores without controlling for student academic aptitudes, of which we were short of good indicators. Therefore, we encourage further studies with multiple years of data, more centralized rating, and good control of student academic aptitudes to comprehensively investigate the predictive validity of the ELA PA test.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*(1), 1-43.

Akin, J. S., & Garfinkel, I. (1977). School expenditures and the economic returns to schooling. *Journal of Human Resources, 12*(4), 460-481.

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

Burstein, L. (1980). *The analysis of multilevel data in educational research and evaluation. Review of research in education* (Vol. 8, pp. 158-233). Washington, DC: American Educational Research Association.

Davis, B., Caros, J., Grossen, B. & Carnine, D. (2002). Initial stages in the development of benchmark measures of success: Direct implications for accountability (Research Report No. RR-11). Washington, DC: Special Education Programs (ED/OSERS). (ERIC Document Reproduction Service No. ED469290).

Goldschmidt, P., & Martinez-Fernandez, J. (2002) *The relationship among measures as empirical evidence of validity: Performance assignments, SAT-9, and high school exit exam performance incorporating effects of school context* (CRESST Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hooper, S. R. (1988). Relationship between the clinical components of the Boder Test of Reading—Spelling patterns and the Stanford Achievement Test: Validity of the Boder. *Journal of School Psychology, 26*, 91-96.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*(1), 71-92.

Linn, L. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Niemi, D., Baker, E. L., & Sylvester, R. M. (in press). Scaling up, scaling down: Seven years of performance assessment development in the Nation's second largest school district. *Educational Assessment, 12*(3,4).

Yen, W. M., & Ferrara, S. (1997). The Maryland school performance assessment program: Performance assessment with psychometric quality suitable for high stakes use. *Educational and Psychological Measurement, 57*(1), 60-84.

Yoon, B., & Young, M. J. (2000). *Validating standards-referenced science assessments* (CRESST Tech. Rep. No. 529). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).