DOCUMENT RESUME

ED 469 290 EC 309 217

AUTHOR Davis, Betsy; Caros, Jennifer; Grossen, Bonnie; Carnine, Douglas

TITLE Initial Stages in the Development of Benchmark Measures of Success: Direct Implications for Accountability. Research Report.

INSTITUTION Kansas Univ., Lawrence. Inst. for Academic Access.
SPONS AGENCY Special Education Programs (ED/OSERS), Washington, DC.
REPORT NO RR-11
PUB DATE 2002-00-00
NOTE 40p.
CONTRACT 84.324S
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS Academic Ability; Accountability; *Benchmarking; *Disabilities; *Educational Assessment; Grade 9; *HIgh Stakes Tests; High Schools; Interrater Reliability; Mathematics Achievement; Predictor Variables; Reading Ability; *Student Evaluation; *Test Reliability; Writing Ability

ABSTRACT

The purpose of this study was to take an initial step toward developing sound and versatile predictive instruments (i.e., benchmark measures) in reading, writing, and math that can be used to assess high-school students' academic performance in a manner that will not only predict scores on high-stakes tests but will also be amenable to repeat administrations. Data were collected on 200 ninth-graders (6% with disabilities) enrolled in 8 math courses with a difficulty level ranging from remedial math to algebra. The sample for the reading benchmark assessment included 235 ninth-graders (33% were English language learners and 2% had disabilities), enrolled in 14 general education language arts courses. The sample for the writing benchmark assessment included 162 students, a subset of the reading sample. Results for the reading comprehension benchmark test indicated good levels of internal consistency. Additionally, variability in performance related to the sixth-grade reading passage was found to have the strongest predictive strength to overall performance variability on both SAT9 and HSEE reading sub-scales. For the mathematics benchmark test results, internal consistency estimates fell within an acceptable range. For the writing benchmark test, adequate levels of interrater reliability were established on the individual scoring criteria. (Contains 23 references and 7 tables.) (Author/CR)

Institute for Academic Access

Research Report #11

Initial stages in the development of benchmark measures of success: Direct implications for accountability

Betsy Davis, Jennifer Caros,
Bonnie Grossen, and Douglas Carnine

2002

IDEAs that Work
U.S. Office of Special Education Programs

Sponsored by the U.S. Department of Education,

Abstract

The purpose of this study was to take an initial step toward developing sound and versatile predictive instruments (i.e., benchmark measures) in reading, writing, and math that can be used to assess high-school students' academic performance in a manner that will not only predict how students are going to score on high-stakes tests but will also be amenable to repeat administrations so that student gains can be reported across the school year. The goal associated with developing these instruments is to help alleviate the problems facing secondary special and general educators relative to accountability.

Data were collected on ninth-grade students in two high schools located in one city in California. The sample for the math benchmark assessment included 200 ninth graders (6% of whom were students with disabilities) who were enrolled in eight math courses with a difficulty level ranging from remedial math to mainstream ninth-grade algebra. The sample for the reading benchmark assessment included 235 ninth graders (33% of whom were ELL and 2% were students with disabilities) who were enrolled in 14 general education language arts courses. The sample for the writing benchmark assessment included 162 students who were a subset of the reading sample. Outcome criterion measures included the math and language arts subscales of the Stanford Achievement Test and the California High School Exit Exam. Results for the reading comprehension benchmark test indicated good levels of internal consistency for all grade-level passages within two potential parallel forms. Additionally, variability in performance related to the sixth-grade reading passage was found to have the strongest predictive strength to overall performance variability on both the SAT9 and HSEE reading sub-scales. For the mathematics benchmark test results, internal consistency estimates fell within an acceptable range for the different subcomponents of the test (i.e., open-response computation items; multistep, multichoice word problems; open-response, prealgebra items). The strongest predictive relations among the math subcomponents were found for more complex open-ended questions covering content upwards to prealgebra skills. For the writing benchmark test, adequate levels of interrater reliability were established on the individual scoring criteria. In addition, a strong, simple and meaningful factor structure was developed from the scoring criteria, resulting in four scoring factors representing the following: (a) writing production skills; (b) writing clarity skills; (c) technical writing skills; and (d) sentence structure skills. Strong continuous predictions and adequate discriminant capacity were obtained, with clarity of writing in combination with a high quantity of output being the strongest contributors to success on both the SAT9 and the HSEE writing outcome measures. These results form the foundation for future IAA efforts regarding test development.

"Only strong academic standards can provide the sturdy foundation we need to dramatically improve student achievement and win back the confidence of the public"
(American Federation of Teachers, 1998, p. 3)

Accountability has historically been, and continues to be, the rallying cry for politicians relative to the improvement of our pubic education system. Over the past five decades, political swings in ideology have resulted in different foci of accountability efforts, but all have centered on the importance of the use of some type of student assessment-based decision framework (Linn, 2000). This move towards a "different" system of accountability in education has resulted in negative outcries relative to the demands placed on students, teachers, and schools to achieve the goals of accountability put forth by governing parties (Madaus, 1985). At no point in time, however, has the outcry against educational accountability been as loud as that heard in the current decade.

The current decade's trend of accountability began with President Clinton's call for the setting of "world-wide" standards (Goals 2000: Educate America Act), whereby social promotion within our schools was to be ended (Linn, 2000). The Clinton policy was followed by the current administration's goal of "testing every child, every year" (Goals 2000: Educate America Act), in order to ensure that no child gets left behind. These political accountability efforts have led to the development of the term "high-stakes" assessments in order to describe aspects of current reform efforts within our schools. Given the political push to meet exceedingly more demanding standards and placing important life decisions on those who are responsible for meeting the set-forth standards, it is not surprising that accountability efforts and high-stakes exams have led to widespread stress for all parties affiliated with schools.

For parents and students, stress is invoked when children are not allowed to enter into successive grades or, at the end of their educational careers, graduate from high school unless test results indicate certain standards have been met. A reflection of this stress is seen in widespread accounts of parents demonstrating to boycott state standardized testing and being supported by teachers and administrators (Robbins, 2001). Within the school structure, teachers and administrators feel stressed when states tie teacher merit raises and school-level financial allocations to the results of imposed student assessments. Sources of this within-school stress may stem from aspects pointed to by the American Educational Research Association's recent position paper regarding policy implications of high-stakes testing (AERA, 2000). The AERA

authors note "setting of high standards of achievement will inspire greater efforts on the part of students, teachers and educational administrators" (pg. 2). They also note, however, potential problems that may arise to preclude this beneficial impact from achieving fruition. Particularly problematic is the lack of evidence for the alignment between the imposed test and the curriculum being taught in the schools and a lack of opportunities for meaningful remediation for all students who fail. These two problems play upon one another within any school system. If school personnel cannot be sure that what they have deemed important to teach will lead their students towards success on high-stakes assessment devices because of a misalignment, then certainly no meaningful form of remediation to promote success can be developed. In essence, student success becomes relegated to a chance occurrence with stress accompanying this uncertainty.

The above-noted problems and associated stress are particularly salient for students with disabilities who by definition lack proficiency with important academic skills. While students with disabilities currently may be exempt from the barrier function of high stakes assessments, if these tests are in fact indicators of important outcome skills, then virtually all students should be able to meet the standards and pass them. This is the ultimate responsibility of public education. Parents and educators of students with disabilities have fought hard to end the segregation of their children and students within the school system and ensure that these children are engaged in a rigorous program of educational attainment. This effort should be rewarded with systems that promote the success of all students relative to the same standard of performance.

Currently, for all student populations, it may be difficult for schools to understand how they are to achieve the goals set forth by both the federal and state accountability mandates. Educational goals stemming from these political arenas come forth as broad, sweeping statements that are difficult, if not impossible, to translate into meaningful and consistent curricula goals. Furthermore, some states have noted that there is no financial allocation within their state departments of education to monitor and ensure that there is a match between what is tested and what is actually taught in the classroom.

Adding to this dilemma, the scores from state-mandated tests are broad in nature and do not translate easily into planning for the remediation of difficulties. Even if specific reporting of scores allowed for the explicit identification of instructional needs, state mandated tests are summative in nature as they are administered towards the end of the school year well after any

meaningful intervention efforts can be made. Interestingly, Meyer (1996) has argued that in high-stakes accountability situations such as those currently evident, teachers, administrators and students are likely to employ all avenues at their disposal in order to "improve" student performance. These avenues do not preclude the manipulation of student scores or inappropriate alteration of teaching processes (i.e., teaching to the specific test items) in order to gain the needed student performance levels valued by politicians. Quite clearly, accountability within an environment of uncertainty relative to the process of moving students successfully towards ambitious and important performance goals has produced a certain anathema towards accountability testing. This anathema is particularly evident in those populations upon whose shoulders' accountability rests and whose lives are directly affected by the decisions made in the name of accountability.

Although problems exist in the stating of educational goals that can be aligned with curricular and remedial efforts, and these problems may not be addressable in short-order, the fact still remains that teachers, students and schools are required to meet the standards put forth. This fact is particularly salient for students at the secondary level of education. One could feasibly argue that within secondary education, the stress is paramount in light of the fact that high school is the last step for students within the public education system, with little time remaining for remediational efforts. Thus, researchers must assist schools, particularly secondary schools, in establishing reliable measures of success with an emphasis on evaluating instructional materials and methods that show promise for moving the educational system in a meaningful manner towards student success.

Unfortunately, the field of research has its own measurement problems when linking student performance to the relative effectiveness of educational interventions. Namely, there is no consistent means of reporting student performance trends across studies to allow for meaningful comparisons. Researchers have called for the development of measures that could be used not only across research studies in a standard, consistent, format but also could be used to assist in identifying deficiencies in specific areas of academic functioning (e.g., Mayer, 1993). However, a recent meta-analytic review of math instruction has revealed that such consistency and specificity has not yet been achieved (Xin & Jitendra, 1999). Add to this fact the paucity of research aimed at, and thereby knowledge relative to, content-area assessment at the secondary level, and the problems associated with current accountability efforts exponentially increase.

The purpose of the present study, therefore, is to take the initial step in developing sound and versatile predictive instruments (i.e., benchmark measures) that can be used to assess high-school students' academic performance in a manner that will not only predict how students are going to score on high-stakes tests but also will be amenable to repeat administrations so that student gains can be reported across the school year. The goal of developing these instruments is to assist in alleviating the reality of problems facing secondary educators relative to accountability efforts as well as researchers relative to their search for effective interventions that move secondary students toward success.

As regards accountability within individual school systems, benchmark measures should possess the following characteristics: (a) be efficient to score and administer so as not to overtax an already stressed educational environment; (b) be amenable to the development of multiple forms to assist in the formative evaluation of student progress as students approach high-stakes assessment conditions; (c) provide meaningful information to teachers and administrators through clearly summarized and interpretable results that inform a range of decisions (including intervention selection decisions); (d) be strongly predictive of success relative to important academic outcomes for students; (e) be sensitive to effective instruction for both low and high performing students so progress trends within a school year may be frequently reviewed and responded to; and (f) provide motivating feedback to teachers and students.

Quite clearly, the above criteria, if met, would not only assist in reducing stress within school systems relative to high-stakes assessment but also address the problematic issues facing educational researchers. Benchmark measures should provide schools and parents with reliable and useful information about student performance as well as provide researchers with consistent measurement devices to serve as research outcomes across studies. Within a research framework, such measures would allow for direct comparisons of the effectiveness of intervention efforts, as well as provide known measures that would be sensitive to change based on intervention implementation. Further, the use of multiple parallel forms of the benchmark measures would assist in the conduct of pre-post-follow-up research designs, eliminating the potential of practice effects or incongruity of dependent measures across time.

The impact of the development of the proposed benchmark measures cannot be overstated relative to their importance in assisting schools and researchers in developing and implementing effective intervention efforts that move students towards the levels of success

demanded by accountability efforts. In fact, it would be difficult to argue against testing educational interventions against a standard of eliciting change in measures displaying strong predictive success to future high-stakes outcomes for children. In fact, this type of meaningfully predictive research outcome should be the sine qua non of all educational research endeavors.

In the present study, the initial stages of development for benchmarks measures in high-school reading, writing and math are reported. Preliminary evidence is presented for both the reliability and predictive validity of these measures in relation to high-stakes state outcomes. Since test development of any kind is an iterative process, not accomplishable within the scope of one study, the results of the current study will hopefully provide heuristic value to the continuing pursuit of developing strong and versatile benchmark measures.

Method

Participating Schools

Study participants were ninth-grade students in two high schools located in one city in California. The first school (School A) had a total enrollment of 1,448 ninth- through twelfth-grade students with an ethnicity distribution of 28% African American, 10% Asian, 28% Caucasian, 27% Hispanic, and 3% other. The second school (School B) had a total enrollment of 1,886 students with an ethnicity distribution of 26 % African American, 33% Asian, 14% Caucasian, 23% Hispanic, and 2% other. The population of School A included 3% English language learners (ELL), 6.3% students with disabilities, 24% students receiving free and reduced lunch, and 7% families receiving aid for dependent children (AFDC). The population of School B included 33% ELL students, 6% students with disabilities, 87% receiving free and reduced lunch, and a 65% families receiving AFDC.

Participating Students

Math Sample 1. The total sample for the math benchmark measure included 200 ninth-graders from School A who were enrolled in eight math courses, with a difficulty level ranging from remedial math to mainstream ninth-grade algebra. Of the 200 potential participants, 154 students completed the computation (open-response) and word problem (multiple-choice) components of the math benchmark measure. The demographic description of this computation/word problem sample was as follows: 57% females and 43% males, 2% ELL, and 6% students with disabilities.

Math Sample 2. A subset of the above-referenced sample, which included 89 ninth-graders, completed the advanced open-response component of the math benchmark measure. Given the close proximity of the scheduled benchmark administration to the scheduled administrations of state mandated tests, teachers and administrators were concerned about over-taxing this population of students with repeated test administrations. Therefore, the advanced component of the Math benchmark measure was not administered to students with disabilities and at-risk students. The demographic description of the advanced sample was as follows: 73% females and 27% males, 0% ELL, and 1.5% students with disabilities.

Reading Comprehension Sample. The total sample for the reading benchmark measure included 235 ninth-graders from School B enrolled in 14 general education language arts courses. Of the 235 potential participants, 116 students completed Form A of the reading benchmark and 119 completed the parallel Form B. The demographic distributions of the Form A and Form B reading samples were not significantly different and, overall, included 55% females and 45% males, with an ethnicity distribution of 26% African Americans, 33% Asians, 14% Caucasians, 20% Hispanics, 4% Pacific Islanders, and 2% Other. Further, the total sample included 33% ELL and 2% students with disabilities.

Writing Sample. The total writing benchmark sample included 162 ninth-graders who were a subset of the reading sample described above. Demographically, the writing sample did not differ from the description given above for the reading sample.

Procedures

Consent. In both schools, passive consent procedures were utilized for several reasons. To appropriately evaluate any research endeavor, collected information must be accurate and generalizable. Active consent procedures can serve to reduce the representativeness of the sample upon which evaluation is conducted and reduce participation by children most at risk for behavioral, emotional and academic difficulties (Severson & Ary, 1983; Anderman, Cheadle, Curry, Diehr, Shutlz, & Wagner, 1995; Dent, Galaif, Sussman, Stacy, Burtun, & Flay, 1993; Ellickson & Hawes, 1989; Noll, Zeller, Vannatta, Bukowski, & Davies, 1997; Iverson & Cook, 1994). These are the very children most in need of having the educational system work for them.

Indeed, according to the researchers cited above, active consent results in participants who are less likely to demonstrate academic, social or behavioral problems and more likely to be Caucasian, live in two-parent families, have parents with more education and be involved in

extracurricular activities. Thus, active consent procedures can result in non-representative samples. Within the current context, when the goal of the research is to assess the viability of benchmark measures for all students, such goals are compromised with the use of the active consent procedure. Given that a large number of students within the current samples reside in poverty settings, the effects of active consent would severely undermine the evaluation of student performance for all students and potentially preclude from evaluation those most likely to be at risk. As regards the impact of passive consent procedures on parental rights, the noted research also indicates that non-response to active consent typically does not reflect active refusal but rather "latent consent" and that non-response to passive consent typically does reflect consent.

Within the current research project with its passive consent framework, participating schools and districts gave the research team written approval to collect the benchmark academic data necessary to achieve the project goals. Acknowledgement was made that the data collected was owned by the school/district and provided to our research team for research purposes only, with no associated identifying information. Further, the schools/district agreed to provide summary information relative to demographic and personal information on our respective samples. Lastly, data for the high-stakes outcomes were collected by school personnel within the participating school district and provided to our research team by identification number only.

Administration of the benchmark instruments. Classroom teachers administered the benchmark instruments during their usual instructional periods (i.e., the math benchmark instruments were administered by teachers during math classes, and the reading and writing benchmark instruments were administered by teachers during language arts classes). Each teacher was provided with and asked to follow standard administration directions exactly for each benchmark instrument. The administration directions are explained within each measure description below.

Benchmark Instruments

Development. A critical issue in the development of benchmark instruments is the question of whether the same benchmark instruments can be used in different states. To address this issue, the math, reading and writing standards and sample test items from the seven largest states (California, Florida, Illinois, New York, Ohio, Pennsylvania, and Texas) plus Washington and Kansas were reviewed. Despite the great variety in the verbal statements of the standards across states, there was remarkable commonality and overlap in test content across states. The

benchmark instruments developed for the current study were aligned with core test content that was included in at least five states' standards and/or test items. The content subsumed in each benchmark instrument is described below, with the scoring for each measure explained within the Results section. As well, reliability estimates for the benchmarks are presented in the Results section as part of the study outcome.

Math benchmark. The math test consisted of three components: (1) computation (open-response), (2) multi-step word problems (multiple-choice), and (3) pre-algebra (open-response). The computation component consisted of 24 items requiring students to compute answers using addition, subtraction, multiplication, division, fractions, ratios, decimals, square roots, exponents, and percentages. The multi-step, multiple-choice, word-problem component consisted of 18 items that required students to apply the range of calculation skills covered on the computation component to solve multi-step word problems. Half the items required students to identify the set of steps that showed the correct way to solve the problem, and half the items required students to work through and identify the correct answer to the problem. These two components are sections of the Multilevel Academic Survey Inventory (MASI) (Howell, Zucker & Morehead, 1994), and were selected based on their established reliability and predictive strength for general math performance. However, since the MASI only covers math content up through the eighth-grade level, the Math benchmark test included a third component which also covered ratios, fractions and exponents, as well as more advanced content (e.g., volume, geometry, integers, equations) within an open-ended response format. The total possible range of scores for all components of the Math benchmark was 0-142 points (i.e., 42 possible points for the MASI-derived components; 100 possible points for the open-ended, higher-level content).

Students were given one hour to complete the mathematics benchmark test. For the open-response computation items, students were asked to "Look at the math problems on your answer sheet. Work these problems on your answer sheet. Make sure your answers are clear. You have 12 minutes to answer as many of these problems as you can." For the multiple-choice, multi-step work problems, students were given 13 minutes to complete two sections (5 minutes for Part 1, 8 minutes to Part 2). In Part 1, students were instructed as follows: "Below each problem there are four ways to find the answer shown. Only one is the correct way. Mark your choice by darkening the circle on your answer sheet that matches the answer you choose." For Part 2, students were instructed: "This time four different answers are given. Only one is correct. Decide how to work

the problem then work it. When you have your answer, choose the answer that matches yours. Mark your choice by darkening the circle on your answer sheet that matches the answer you choose." For the open-response pre-algebra component, students were given 35 minutes to complete it and they were instructed to "Show your work for each problem, as points are given for working the appropriate steps as well as for the correct answer." Given that points were award for each item based on the number of appropriate steps taken in solving the problem, 20% of the open-response tests were scored by a second person in order to estimate inter-scorer reliability. Percent agreement between the two scorers on the current sample was 98%.

Reading comprehension benchmark. The reading comprehension test consisted of three "maze" passages spanning a range of comprehension levels (fourth, sixth, and eighth grade). With this format, students read passages that have words missing (a blank line indicates each missing word), and under each blank line three answer choices are given (1 semantically incorrect word, 1 syntactically incorrect word, and the correct word). Parallel forms (each containing three passages spanning the three grade levels) of this measure were administered (i.e., Forms A and B were given to every other student in each class). Students recorded their answers on a scantron answer sheet, which was scored electronically, and a total of 15 minutes response time was given to complete all three passages (total of 105 and 107 items for Forms A and B, respectively).

Three advantages of the maze test format are: (1) empirical evidence of predictive ability for general reading comprehension, (2) relative ease of constructing multiple parallel forms (standard procedures exist for determining which words to delete from each passage and to select distracters for each answer), and (3) ease of administration and scoring. All six reading passages were selected from the MASI (Howell et al., 1994). An advantage of using MASI passages was that two passages for each grade level (fourth, sixth, and eighth grade) were already available, thus providing groundwork for constructing parallel forms.

Students were instructed as follows: "You are going to read three passages. Some of the words are missing. Below each blank line there are three words. One of these words is the missing word for that blank line. As you read the story, choose the word that best fits the story. Mark your choice by darkening the circle on your answer sheet that matches the answer you choose. You have fifteen minutes to do all of the items. Each passage gets a little more difficult. Concentrate and do your best."

Writing benchmark test. To sample writing performance, students were given ten minutes to write on each of two prompts (one expository writing topic and one persuasive writing topic). These writing samples were scored using a scoring rubric focused on ten aspects of writing: (1) overall message clarity; (2) lack of fragments; (3) lack of run-ons; (4) correct usage (e.g., subject-verb agreement, homonyms, slang); (5) correct capitalization; (6) correct punctuation; (7) correct spelling; (8) number of correctly structured sentences (i.e., not a fragment or run-on); (9) number of words written; and (10) number of sentences written. Feature 1 was scored using a three-point scale (1, 3, or 5), features 2-7 were scored using a four-point scale (0, 1, 3 or 5), and features 8-10 were simply tallied.

Advantages of the current writing and scoring format included: (a) relatively short administration time, (b) relatively short scoring time, (c) assessments of both expository and narrative writing performance, and (d) assessments of both the quality of writing and the quantity of production. Inter-rater reliability was assessed on 20% (n=32) of the writing samples, with equal numbers from the expository and narrative samples being double-scored by independent raters. For the categorical scoring criteria, exact agreement between raters was required and resulted in a range of 78%-97% agreement across scoring categories. The lowest agreement levels were found for overall clarity (78%) and correct usage (82%), with the highest agreement levels being found for capitalization (95%) and punctuation (97%). For the total-number variables, the correlation between raters was .83 ($p < .001$), .97 ($p < .001$), and .93 ($p < .001$) for correctly structured sentences, total number of words written, and total number of sentences written, respectively.

Students were given both an expository and narrative prompt and 15 minutes to write on each subject. At the beginning of each writing prompt, students were instructed as follows: "You will have 2 minutes to plan your writing, 10 minutes to write your response, and 3 minutes to proofread your work. I will tell you when to begin writing you response and when to begin proofing your work."

Outcome Measures

The math and language arts sub-scales of the Stanford Achievement Test (SAT9) (Harcourt Educational Measurement, 2000) and the High School Exit Exam (HSEE) (California Department of Education, Standards and Assessment Division, 2001) were the high-stakes academic outcome measures used to evaluate the predictive strength of the benchmark math,

reading, and writing tests. These outcome measures were used because they are mandated by the state of California and are used for high-stakes decisions within the state.

The SAT9 is a nationally norm-referenced achievement test widely used on an annual basis for rank ordering schools (and rank ordering students within schools) and for attempting to evaluate school improvement initiatives. The SAT9 reading scale consists of reading, vocabulary, and reading comprehension sub-scales. The SAT9 writing scale consists of writing mechanics and writing expression sub-scales.

The HSEE is a mastery test designed to help ensure that students are proficient in critical math, reading, and writing skills; it was administered for the first time in the spring of 2001 to all ninth-grade California students. Beginning in spring 2002, all tenth-grade students in California will be required to take the HSEE, and they will have until their senior year to pass all subscales. The HSEE math subscale consists of five components: (1) statistics, data analysis and probability, (2) number sense (e.g., computing rational numbers expressed in a variety of forms), (3) algebra and functions, (4) measurement and geometry, and (5) algebra. The student response format for the five math components is multiple-choice. The reading subscale consists of three components: (1) word analysis and vocabulary development (e.g., identifying and using literal and figurative meanings of words and understanding word derivations), (2) reading comprehension (e.g., comparing and contrasting features and elements of consumer materials to gain meaning from documents), and (3) literary response and analysis (e.g., analyzing interactions between main and subordinate characters in a literary text and explaining the way these interactions affect the plot). The student response format for all three reading components is multiple-choice. The HSEE writing subscale consists of three components: (1) writing strategies (multiple-choice), (2) writing conventions (multiple-choice), and (3) essay writing (open-response). Adequate internal consistency estimates for the multiple-choice portions of the HSEE have been reported as well as adequate discrimination between high- and low-performing students based on total scores (Wise, Sipes, George, Ford, & Harris, 2001). Additionally, psychometric information on the HSEE scoring for the essay writing portion indicate adequate inter-rater agreement, with approximately 70% absolute inter-rater agreement across two essay prompts and 29% inter-rater agreement within one point.

Results

Performance on the SAT9 and HSEE

Performance levels for each sample of students described above, relative to the state assessments that served as the predictive outcomes for the current analyses, are reported in Table 1. For the computation/word problem sample (Math Sample 1) on SAT9 math sub-scale, the indicated range of 0-41 correct items out of a total of 53 points possible corresponds to a percentile range of 1%-97%. The mean raw score performance of 18 corresponds to an average percentile of 38%. Relative to performance levels on the HSEE math sub-scale, the descriptive statistics presented in Table 1 should be viewed in relation to a possible scale score range of 250-450.

For the HSEE, a scaled-score passing criterion of 350 has been established by the state for all sub-scales included in the exam. For Math Sample 1, 23% achieved a passing score on the HSEE math sub-scale. Relative to SAT9 performance, no cut-off score has been established by the state, thus all current samples were viewed relative to a 50th percentile cut-off score as established by the national norms provided for the test. For Math Sample 1, 32% of the students achieved a SAT9 math score at or above the 50th percentile.

Performance levels for the Math Sample 2 indicated a SAT9 math raw score range of 0-41, corresponding to a percentile range of 1%-97%. The reported mean raw score performance of 20, corresponds to an average percentile ranking of 45%. For this sample, 36% achieved a passing score on the HSEE math sub-scale (i.e., at or above 350), with 43% of the students performing at or above the 50% percentile on the SAT9 math sub-scale.

The performance levels for the Form A reading sample, relative to the SAT9 reading results, ranged between 0 and 76 correct items out of a total of 90 points possible. The corresponding percentile range was 1%-91%. The mean raw score was 43, corresponding to the 25th percentile. Performance levels on the HSEE language arts sub-scale (reflecting a total of reading comprehension and writing sub-scales) ranged between 282 and 409 points out of a possible scale score range of 250-450. For the Form A sample, 62% achieved a passing score on the HSEE (i.e., at or above 350), with 15% performing above the 50th percentile on the SAT9 Reading sub-scale.

The Form B sample did not significantly differ from the Form A sample relative to performance levels, with a SAT9 Reading raw score range of 0-74 corresponding to a percentile

range of 1%-86%. The mean raw score was 44, corresponding to an average percentile score of 26%. Relative to performance levels on the HSEE language arts sub-scale, the scaled score range was 279-413 out of a possible score range of 250-450. For the Form B sample, 60% achieved a passing score on the HSEE, with 9% performing above the 50th percentile on the SAT9.

Performance levels for the writing sample on the SAT9 language sub-scale indicated a raw score range of 0-43. This range corresponds to a percentile range of 1%-95%. Relative to performance levels on the HSEE language arts sub-scale, the scaled score range was 279-413 out of a possible score range of 250-450. For the writing sample, 31% performed above the 50th percentile on SAT9 language arts sub-scale while 56% received a passing score on the HSEE language arts sub-scale.

Reliability Estimates for the Reading Comprehension Benchmark Test

Reliability estimates, as assessed through internal consistency estimates utilizing Cronbach's alpha (Coefficient Alpha) (Cronbach, 1951), for the subscales and total scores for the reading comprehension benchmark test are presented in Table 2. Coefficient alpha is a general reliability coefficient, encompassing both the Spearman-Brown prophecy formula as well as the Kuder-Richardson 20 formula (Carmines & Zeller, 1979). Alpha provides a lower bound to the reliability of a scale and, thus, the reliability of a scale can never be lower than alpha even if items depart substantially from being parallel measurements. In essence, alpha provides a conservative estimate of a measure's reliability (Novick & Lewis, 1967).

Each form (i.e., Form A and Form B) of the reading comprehension benchmark test consisted of three comprehension passages, with difficulty ranging from 4th grade to 8th grade reading level. Results in Table 2 are presented for each grade-level passage within each of the two forms as well as for the total form scale comprised of all items across reading difficulty levels. The reliabilities for the total reading scales and their respective subscales were acceptably high across the two different forms. It has been suggested that reliabilities in the 90's or high 80's are sufficient for most purposes that involve using test scores as information about individuals. Further, reliabilities in the 70's or low 80's are adequate for most purposes that involve using summaries of test scores as information about groups (Thorndike & Hagen, 1961).

From an item-level perspective, the Form A 4th-grade reading passage had three items with zero variance with all students responding with the correct answer. Of the remaining items, the mean of the items was .93 with a range of .44-.99. Given that all items were scored as "0"

incorrect and "1" correct, each item mean would indicate the percentage of students answering each item with a correct response. Thus, an average of .93 across all items would indicate that a majority of students provided correct answers to the 4th-grade reading items, with an item variability of .06. No items on the Form A 6th-grade passage had zero variance. The mean of the items was .74, with a range of .55-.96. Overall, students answered fewer items correctly on the 6th-grade passage when compared to the 4th-grade passage, with a higher average item variance of .18. For the Form A 8th-grade passage, again, no items had zero variance; however the mean of the items was much lower than those of the previous grade-level passages at .29, with a range of .16-.47 and an average item variance of .20. In essence, relative to the 4th- and 6th-grade passage, fewer numbers of students answered items correctly on the 8th -grade passage.

For the Form B 4th-grade reading passage, unlike Form A, no items had zero variance. The mean of the 4th-grade Form B items was .93, with a range of .76-.99. As with Form A, overall students tended to homogenously respond with correct answers to the 4th-grade passage items, with an average item variance of .06. For the Form B 6th-grade passage, the mean of the items was .67, with a range of .44-.91. Again, students answered fewer items correctly on the 6th-grade passage when compared to the 4th-grade passage, with a high item variance of .20. For the Form B 8th-grade passage, the mean of the items was .22, with a range of .08-.39. As with Form A, the fewest number of students answered items correctly on the 8th-grade passage, with the average item variance being .17.

The item-level results indicated a homogenously correct response pattern on the 4th-grade passage for both forms of the reading measures, a slightly more variable response pattern for both forms on the 6th-grade passage, and the greatest response variability pattern on the 8th-grade passage for both forms.

The Parallel Nature of the Reading Comprehension Forms

The different forms of the reading comprehension benchmark test were given to two different samples of students and, as noted above, these two samples did not differ significantly on demographic or performance outcome variables (i.e., SAT9 and HSEE). It would be ideal to have parallel forms given to the same sample of children in order to assess the correlation between the two forms and lessen the potential influence of cohort effects. Unfortunately, the reality of collecting data in applied academic settings did not allow this luxury. Thus, the only comparison to be made between the two reading forms is to examine the potential of mean level

differences between forms using between-groups t-tests. Significant between-group differences would indicate potential differential difficulty levels between the two forms. For student scores on the individual grade-level passages for Forms A and B, there were no significant between-group differences. For the 4th-grade passage, the raw score means for Form A and B were 38 and 38 respectively (t (233)=.48, p=.63). For the 6th-grade passage, the raw score means for Forms A and B were 29 and 27, respectively (t (233)=1.69; p=.09). For the 8th-grade passage, the means for Forms A and B were 8 and 6, respectively (t (233)=1.57; p=.12). The total raw score across difficulty levels for Forms A and B were 75 and 72, respectively (t (233)=1.65; p=.10). Given the lack of significant between-group differences on the two forms of the reading comprehension benchmark test, paired with similarity in the internal consistency estimates across forms as well as lack of significant differences in demographic and performance levels of the two groups taking each form of the test, the two samples and their respective reading scores were treated as one unitary sample.

Reliability Estimates for the Mathematics Benchmark Test

Reliability estimates for the mathematics benchmark test are also presented in Table 2. The math test consisted of three sub-components (i.e., open-response computation problems, multiple-choice, multi-step word problems, and open-response pre-algebra problems), with reliability estimates for the total math scale and the open-ended subscale meeting sufficiency for the individual decision-making criteria, and the computation and word problem subscales meeting sufficiency for the group decision-making criteria.

From an item-level perspective, for the computation subscale, the average item mean was .60 with an item range of .03 to .99. Average item variance was .12. For the multiple-choice word problems, the average item mean was .68, with an item range of .12 to .97 and an average item variance of .16. Once again, item scores on the computation and multiple-choice subscales were dichotomous in nature, with the item means representing the percentage of students getting each item correct. Thus, across items on both subscales, the average number of students responding with a correct answer was relatively large (i.e., over 50%), however, with significant variability across items as represented by the range of item means and average item variance.

The items on the open-ended subscale of the benchmark test were scored based on the number of calculative steps necessary to solve each item. Thus, possible scores for all items ranged from 1 to 6 points. The internal consistency estimate for the open-ended subscale was

acceptably high. From an item-level perspective, the average item mean for the 33 1-point items was .42, with a range of .00 to .94 and an average item variance of .17. Overall, the average number of students responding with a correct response for the 1-point items was less than half the sample. The average item mean for the 11 2-point items was .69, with an item average range of .21 to 1.17 and an average item variance of .56. For the 2-point items, the item means cannot be interpreted as the percentage of students responding with a correct answer as could be done with the 1-point items. However, the overall item mean for these items reflect that, on average, students received less than 1 out of 2 possible points for these items.

For items scored on a 3-point scale (5 items), the average item mean was .53, with a range in item means of .26 to 1.01 and an average item variance of .80. Once again, on average, students received less than 1 point out of a possible 3 points for these items. There were two items for which the possible point totals were 4. These items involved first finding the area of different shaded shapes (e.g., triangles, parallelograms) on a gridded space, and then computing the total area shaded. For these two items, students got an average of .11 and .12 points correct, with over 90% receiving no points. Likewise, there were two items for which the possible point totals were 5. These items involved first determining the areas associated with two figures (e.g., cubes, circles, rectangles) and then subtracting the areas to determine the area remaining. The average points correct for these items were .49 and .58, with over 70% of students receiving no points. Lastly, there were two items whose possible point totals were 6. These items involved manipulating ratios to solve multi-level word problems. For these items, the average points correct were .08 and .13, with over 90% of students receiving no points correct for either item.

Comparability of the Writing Benchmark Prompts

The scoring criteria for the writing benchmark consisted of 1-5 ratings for multiple aspects describing the quality of the writing sample (e.g., capitalization, usage errors) based on the first 50 words of the written sample, as well as scores based on the entire written passage indicative of the overall clarity (1-5 rating), total number of words written, total number of sentences written, and total number of correctly structured sentences. Students were given two different writing prompts (expository and narrative) for which they supplied writing samples. To examine the extent to which students produced different writing samples based on the prompt given, between-prompt analysis was performed utilizing paired t-tests. These results are presented in Table 3.

As noted in Table 3, significant differences were found between prompts. It appears that students produced a higher quantity of writing to Prompt A when compared to Prompt B. As well, there was a trend towards better quality for Prompt A.

Reliability Estimates for the Writing Benchmark Test

As noted in the Methods Section, inter-rater reliability was within an acceptable range for each scoring criterion for both the expository and narrative prompts. The reliable mean-level prompt differences for the criteria noted above may be related to numerous factors influencing the writing sample produced by students, including student interest in, and knowledge of, the topic presented. In light of these differences, the average of each criterion was taken to represent the overall quality and quantity of student writing across prompts. Internal consistency estimates for the averaged scoring criteria, however, indicated that a general unitary scoring rubric was not appropriate (i.e., scores for all criteria did not converge into an internally consistent structure). Internal consistency estimates for the scoring criteria across the two writing prompts ranged from alpha = .14 to alpha = .40. All estimates were well below the criteria set for summing all items to create a unitary writing score for individual and group decision-making.

Given that a unitary factor was not forthcoming, exploratory factor analysis was performed to determine if a simple factor structure reflecting multiple scoring domains was plausible. Varimax and oblique rotation procedures were performed to determine the best fit for the potential factor structure. Varimax rotation creates independent factors and attempts to spread item variance across these derived factors. Oblique rotation tests a factor structure wherein the derived factors are correlated. Results indicated that the varimax procedure produced the best fit to the data. The varimax solution was produced in half the number of iterations as that of the oblique rotation procedure and inter-correlations between factors produced by the oblique rotation were minimal (r's ranging from -.01 to .03). The results of the rotated varimax factor structure are presented in Table 4.

Factor loadings above .30 are listed in the table. As can be seen, varimax rotation resulted in a simple structure wherein scoring criteria highly loaded on only one factor. The resulting 4-factor structure, with eigenvalues of 2.58, 1.69, 1.14, and 1.07 respectively, accounted for 65% of the variance in the scoring criteria. As well, all scoring criteria accounted for significant amounts of factor variance, with communality estimates ranging from .35 to .90. The resulting

factors represented: (a) a quantity-related production factor; (b) a quality-related clarity factor; (c) a quality-related technical factor; and (d) a quality-related sentence structure factor.

Descriptive Statistics for the Benchmark Tests

The items for each benchmark scale are discussed above as regards the internal consistency of the scales. Herein are presented descriptive statistics for the total scores that were computed for each benchmark measure and utilized in our predictive relation analyses. These results are presented in Table 5. For the reading comprehension benchmark test, total scores were computed by summing the number of items correct across all grade-level passages within a Form (i.e., Form A or B). For the mathematics benchmark test, wherein items were scored on different scoring metrics, the total score reflects the sum of the percent correct on each sub-scale. This allowed for an equitable score for each component contributing to the total score. For the writing benchmark test, factor scores were created utilizing unit weighting to represent the factor structure produced from the varimax rotation procedure. The factor scores were then summed to create a total writing score.

As noted in the Methods Section, only a subset of students completed the open-ended sub-scale of the mathematics benchmark test. The statistics for this sample are also presented in Table 5. In order to equalize the sample size for the different math subscales, a regression estimation procedure was utilized whereby the available open-ended scores were regressed on the computation and multiple-choice word problems. Through this procedure, open-ended scores were estimated for the entire sample. In computing the equation for the estimated open-ended scores, both the computation and multiple-choice word problem subscales were significant predictors (beta = 1.62, p = .003; beta = 2.03, p = .000, respectively), predicting over 50% of the variance in the available open-ended scores. Descriptive statistics for the estimated scores are also presented in Table 5. Within the prediction analyses, the available open-ended data were utilized and the predictive results were compared to the regression-based estimated sample results.

As can be seen in Table 5, there are no significant skew or kurtosis issues related to the benchmark total scores. As well, the true and estimated benchmark scores for math are relatively equivalent in mean levels and distribution properties, the only difference being that a lower range of scores was estimated.

Predictive Relation of Benchmark Performance to High-Stakes Outcomes

Caveats related to outcome measures. Before beginning the presentation of the predictive relations between performance on the benchmark tests and high-stakes outcomes tests, several caveats need to be expressed related to the different forms of the outcome measures to be utilized in the analyses. As noted in the Methods Section, both the SAT9 and the HSEE are salient outcomes for California students. In the current analyses, a continuous form of both outcomes was utilized to assess the amount of variance explained in each by the benchmark measures. Statistics for the SAT9 provides percentile rankings for scores within the domains of reading, mathematics, and writing. For the HSEE, a scale score is produced for the domain of mathematics. However, for the domain of language arts that incorporates both reading and writing sub-scales, only one scale-score is produced and reported. In order to separate out the HSEE language arts score into meaningful outcomes relative to the reading comprehension and writing benchmark tests, scores were created from the different HSEE sub-scale scores.

The HSEE reading outcome score was comprised of the word analysis, reading comprehension, and literary analysis sub-scales. For these scales, the proportion of correct items relative to the total possible items for each scale was determined. Then the average of these ratios was calculated to create a score representing each student's reading outcome. The HSEE writing outcome score was comprised of the writing strategies and writing conventions multiple-choice sub-scales, as well as scores on two written essays scored on a 1-4 scale. To create a total writing outcome score, the proportion of correct responses relative to the total correct possible for each of the multiple-choice sub-scales was determined and then these proportions were averaged. The proportion of points scored on each essay relative to the total possible score of four was also determined and an average essay score was calculated. Noting that the HSEE counts the essays as 30% of a student's total score, the averaged multiple-choice sub-scale score was weighted at 70% and the essay average score at 30% and these weighted averages were summed to create a total HSES writing outcome score.

After the continuous variable analyses, we utilized dichotomous outcomes to determine the most salient benchmark aspects that discriminated between those who passed selected outcome criteria and those who did not. For the HSEE, a scaled-score of 350 is the criteria set forth by the state of California as a passing score for the different academic domains. This cut-off score was used for the HSEE mathematics outcome. For the created HSEE reading and

writing outcomes, the cut-off criterion representing an average of 50% correct was selected for each scale. For the SAT9, there is no official cut-off criterion; therefore, dichotomous criteria of "above the 50%" and "below the 50%" was selected to determine those who "passed" and those who "did not pass" each SAT9 academic domain.

Correlations of benchmark measures to outcomes. Table 6 presents the univariate relations between the reading comprehension, mathematics and writing benchmark measures and the continuous high-stakes outcome measures described above.

As can be seen, all relations among the benchmark totals and the outcome measures were highly significant. For the mathematics true and estimated samples, the correlations were lower for the estimated sample when compared to the true sample; however, both were significantly related to the outcome measures.

Regression and discriminant analyses predicting outcomes. With significant relations established between the benchmark and outcome measures, each outcome measure was regressed on the respective components of the benchmark sub-scales comprising the total score. A step-wise procedure for the removal of variables from the equation was utilized. The intended purpose of this analysis was to better elucidate the strongest contributor to the prediction of the high-stakes outcome, controlling for the inter-correlation between the component parts of the benchmark measures. The step-wise regression results are presented in Table 7.

For the prediction of both the SAT9 and HSEE reading outcomes, the 6th-grade level reading passage provided the most significant, albeit small, predictive contribution. Thus, for the current sample, the minimum competency reading-level predictive of higher performance on state-level high-school outcomes is 6th-grade. Discriminant function analysis was performed, utilizing a Wilks-Lambda criterion and a .05 threshold for variables entering the equation, to determine if the significant regression results in the continuous outcome analysis also held when discriminating between those who received a "passing" score on the outcome from those who "did not pass." For the SAT9 Reading outcome, the 6th-grade reading-level passage was the strongest discriminator with a standardized canonical coefficient of 1.0, with the 8th-grade passage loading .61 on the canonical discriminant function and the 4th-grade passage having the smallest loading at .23. The discriminant function for the SAT9 outcome, however, did not discriminate well those who fell within the upper 50% of the SAT9 percentile distribution; rather

it predicted that all students (100%) would fall under the 50th percentile on SAT9 reading subtest.

As regards the HSEE reading outcome, once again, the 6th-grade reading-level passage was the strongest discriminator with a standardized canonical coefficient of 1.0, with the 8th-grade passage loading .54 on the canonical discriminant function and the 4th-grade passage having the smallest loading at .20. The discriminant function based on these loading did significantly discriminate between those who performed within the upper 50% on the reading sub-scales and those who did not, with 88% correctly identified. Likewise, the discriminant function correctly predicted 60% of those who fell below the 50th percentile on the HSEE reading sub-scales, with an overall correct prediction rate of 77%.

For the prediction of both the SAT9 and HSEE continuous mathematics outcomes, the open-response pre-algebra problems were the strongest predictor, with the multiple-choice multi-step word problems contributing a smaller but significant predictive relation. The open-response pre-algebra component, however, had the potential for larger variability given the larger number of items as well as a larger distribution of skill difficulty, thus the predictive results are not surprising. However, regardless of the sample utilized in the analysis, the results were relatively consistent.

Discriminant function analysis was performed as above to determine if the significant regression results in the continuous outcome analysis also held when discriminating between those who received a "passing" score on the outcome measure from those who "did not pass." Given that the benchmark mathematics sub-scales consisted of different formats (direct computation, multiple-choice word problems covering computation skills, open-ended word problems requiring computation and problem-solving) and covered overlapping and unique skill difficulty levels, an iterative item-level approach to the discriminant analysis was utilized. First, items from each benchmark sub-scale were independently assessed for their discriminant relation to the outcome measures. Those items surviving the independent analysis were then combined into one discriminant analysis to determine the strongest combination of skill performance that best discriminated those who "passed" and "did not pass."

Regarding the SAT9 mathematics outcome, five open-ended and multiple-choice items were the strongest discriminators of those who passed. The five items represented skills related to (a) multiplication of decimals and percentages; (b) correct application of multiplication and

division to a word problem; (c) multiplication of opposite-sign integers; (d) geometric computation of area; and (e) division applied to complex problem-solving (with standardized canonical coefficients of .58, .43, .54, .42, and .60, respectively). The discriminant function based on these variables discriminated quite well; correctly identifying 85% of those who fell within the upper 50th percentile of the SAT9 Math distribution and 92% of those fell below the 50th percentile. The overall correct classification was 89%.

For the HSEE mathematics outcome, where a scaled score of 350 determined the "pass"/"no pass" distinction, three open-ended items were the strongest discriminators of those who passed. These items reflected skills related to (a) determining exponents; (b) geometric understanding of degrees; and (c) computation of area within a complex geometric figure (with standardized canonical coefficients of .56, .85, and .61, respectively). The discriminant function based on these variables correctly identified 77% of those who received a HSEE of 350 or above, and 82% of those who did not. The overall correct prediction rate was 80%.

For the prediction of the SAT9 and HSEE writing outcomes, both the clarity and production factors contributed significantly to the explanation of variability, with the sentence structure factor adding prediction to SAT9 performance and the Technical factor adding prediction to the HSEE. Interestingly, in the discriminant analysis for the SAT9 Writing outcome, clarity and production remained strong discriminators of those who passed and those who did not, with standardized canonical coefficients of .59 and .70 respectively. However, instead of the sentence structure factor, which was found to be significant in the continuous analysis, the technical factor added significantly to the discriminant function, with a standardized coefficient of .44. The discriminant function based on these variables correctly identified 40% of those students who scored at or above the upper 50th percentile on the SAT9 writing distribution and 93% of those fell below the 50th percentile, with an overall correct classification of 77%.

In examining the HSEE writing outcome, the production factor in addition to the clarity factor were the strongest discriminators of those who performed within the upper 50% on the writing sub-scales, with standardized canonical coefficients of .86 and .68, respectively. The discriminant function based on these variables correctly identified 66% of those who fell above the 50th percentile, and 69% of those who did not. The overall correct prediction rate was 67%.

Discussion

The current results show promise for the further development of benchmark instruments in the academic domains of reading comprehension, mathematics, and writing. The results for the reading comprehension benchmark test indicated very good levels of internal consistency for all grade-level passages within the two potential parallel forms. Further, the finding of no between-form differences on mean-level performance indicates that these two forms show promise for the future development of similar parallel forms.

As regards the predictive strength of the reading comprehension benchmark test, variability in performance related to the 6th-grade reading passage was found to have the strongest predictive strength to overall performance variability on both the SAT9 and HSEE reading sub-scales. One possible reason for this particular finding is statistical in nature. In the current sample, benchmark performance levels were optimally variable on the 6th-grade passage, thereby potentially increasing its predictive relation to the outcomes. A majority of students, on average, passed all items on the 4th-grade passage and comparatively fewer students passed items on the 8th-grade passage. It is quite possible that the performance distributions across grade levels, and thereby the predictive relations, may differ on other samples whose demographic characteristics vary from those of the current sample.

Another possible reason for the 6th-grade predictive relations relates to the purpose of the particular outcome measure being assessed. For example, on the HSEE in California, one purpose may be to ensure that students have a minimal level of literacy before graduating high school; a 6th grade reading proficiency level being a commonly accepted standard of literacy. The discriminant analyses support this contention in that, based primarily upon their reading level on the benchmark 6th-grade passage, 88% of the students who received a passing score on the HSEE reading sub-scale were correctly identified. Certainly, a worthwhile goal of education is to ensure that <u>all</u> students attain a certain level of reading fluency and comprehension before leaving school. The HSEE is given at the beginning of 9th grade, with opportunities given throughout the next four years to re-take the test. For those students who are reading at a proficient level in the 9th grade, the barrier to graduation is removed early in their high school academic careers. On the other hand, for those who do not pass early on, there is time to address the reading proficiency issues before these students either drop out or graduate from high school.

The purposes, however, are not the same for all high-stakes tests administered and, thus, different predictive relations to those outcomes may ensue. For the SAT9 reading, in comparison, students' scores on the 6th-grade passage of the reading comprehension benchmark test did not correctly identify those students who fell within the upper 50th percentile. The SAT9 is a nationally norm-referenced test that is used in the districts involved in the current study to assess relative performance across schools rather than across individual students. Thus, there may be higher performance expectations to fall above the 50th percentile relative to national norms than those put forth by the high-school exit exam.

The predictive relations of the reading comprehension benchmark test will, of course, need to be validated on future samples with different demographic characteristics and different high-stakes outcomes and purposes, in order to assess the generalizability of the current results. The current sample for the reading comprehension assessment was not necessarily a strong sample upon which to test the prediction to the upper 50th percentile on a national norm-based test. In the current sample, an average of 12% of the students fell within the upper 50th percentile with regard to SAT9 reading scores, thus potentially contributing to a restriction in the range of scores. Quite possibly, given more upward variability in both the benchmark measure of reading and the national norm-based outcome, stronger predictive relations would be obtained.

For the mathematics benchmark test results, internal consistency estimates fell within an acceptable range for the different sub-components of the test (i.e., open-response computation items; multi-step, multi-choice word problems; open-response, pre-algebra items); the highest sub-component reliability estimate being found for the more complex open-response questions. The individual item analyses on each sub-component indicated that the majority of students correctly answered the lower-level, single-step, computation items, with decreasing numbers passing the more difficult multi-step open-response items. This trend is what would be expected of a scale that incorporated a range of increasingly difficult skill-level items. The goal of developing a benchmark measure that tapped a wide range of skills was set forth so as to avoid potential floor and ceiling effects and have the ability to assess meaningful performance change in both low- and high-performing students.

The strongest predictive relations among the sub-components in the continuous variable regression analysis were found for the more complex open-ended questions covering content upwards to pre-algebra skills. This result was replicated in the larger estimated mathematics

sample as well. As noted in the Results Section, the increased variability within this subset of items, due to the larger number of items and total point potential, could have contributed to its increased prediction. However, the discriminate analysis at the item level revealed specific item patterns not impacted by the overall variability that successfully discriminated those who fell within the upper 50th percentile on SAT9 math as well as those who received a passing score on the HSEE.

For SAT9 success, the math skills found to be the best at discriminating those who were successful were more diverse than skills identified for HSEE success. The skills that best discriminated SAT9 mathematics success from non-success included (a) multiplication and division skills applied to word problems; (b) division skills embedded in multi-step complex word problems; (c) multiplication rules for opposite-sign numbers; and (d) computing the area of objects embedded in one another, then manipulating the area estimates. For the HSEE, as with the SAT9, computing the area of different objects was also an important indicator of success. However, this skill was paired with geometric understanding of degrees (e.g., how many degrees are in a circle) and understanding the rules for dealing with exponents in further predicting HSEE success. Unlike the differential prediction to outcomes found with the reading comprehension benchmark components, the mathematics benchmark items equally predicted success on the SAT9 and the HSEE, with a high rate of accurate prediction.

Interestingly, though the majority of students successfully completed the computation items related to multiplication and division skills on the math benchmark test, the application of these skills to multi-step word problems served as a strong discriminator for SAT9 success. Thus, students who possess the computation skills for correctly multiplying and dividing numbers might lack the ability to apply these skills to more complex word problems. This fact would contribute to an application deficit being an important discriminator for success. However, differences in reading proficiency may also play an important role. If students cannot proficiently read the mathematical situations presented to them, it may be more difficult for them to apply their computational knowledge to these word problems. This fact, too, would contribute to an important discrimination difference on the mathematics outcome measures, but would lead to a different set of remediation efforts relative to improving math performance.

Unfortunately, one problematic issue arising from the current mathematics study was the fact that only a subset of students completed the more complex open-response items. Thus, there

was a limited sample upon which to develop the item-level discriminant analysis results. Understandably, school personnel did not want to over-stress their at-risk and lower-performing students by repeated test administrations; however, the resultant higher-performing sample provides for a less generalizable statement regarding the skill-levels most predictive of success on high-stakes outcomes. As well, even within the larger-range mathematics sample, there was not enough variability within the lower-level computation items to allow for a sufficient examination of critical lower-level skills that may be necessary for passing the outcome assessments. Thus, future research is needed that will increase the variability in range of skill levels to assess the viability of the lower-level items as well as to create a larger sample completing all components of the benchmark test to validate the current discriminant results.

For the writing benchmark test results, the fact that this was the first attempt to utilize the developed scoring rubric for the writing samples supplied by students is worthy of note. Most certainly, the results from the current analyses are promising. Adequate levels of inter-rater reliability were established on the individual scoring criteria. As well, a strong, simple and meaningful factor structure was developed from the criteria, resulting in four scoring factors representing (a) writing production skills; (b) writing clarity skills; (c) technical writing skills; and (d) sentence structure skills. Lastly, strong continuous predictions and adequate discriminant capacity were obtained. Clarity of writing in combination with a high quantity of output appear to be the strongest contributors to success on both the SAT9 and the HSEE writing outcome measures. These factors correctly classified 77% of those who "passed" and "did not pass" the SAT9 writing outcome and 67% of those who "passed" and "did not pass" the HSEE writing outcome.

One problematic issue within the current scoring rubric is the fact that different scoring criteria had different scaling metrics, with different criteria scored on different portions of the writing sample. The majority of the 1-5 scaling metric criteria was scored on the first 50 words of the writing sample (e.g., usage, punctuation), the exception being the 1-5 scoring of clarity. The total-number scaling metric criteria (e.g., total correct sentences) were scored on the entire writing passage. Interestingly, the factors that contained elements of the criteria based on the total written sample (i.e., clarity and production) displayed the strongest prediction. The limitation of scoring some of the criteria only on the first 50 words was based on a desire to reduce the amount of time necessary to score the passages. However, the current results must be

viewed with an eye towards balancing the efficiency of scoring with the validity of the scoring results. The scoring procedures are now being revised and current performance data will be re-analyzed relative to these changes.

Future research will continue the iterative examination of these benchmark tests. Plans for upcoming data collection are specifically geared towards addressing the weaknesses inherent in the current study. A large data collection effort is now in progress in which parallel forms of the reading benchmark test will be administered to the same sample of students in order to lessen the potential of cohort effects. Additionally, relative to the reading benchmark test, a 10^{th}-grade reading passage is being included to upwardly increase the skill level tapped. The current sample may have been restricted downward relative to their levels of reading comprehension, as indicated by their SAT9 performance, and thus future samples incorporating a fuller range of skill levels will require this upward expansion in order to tap the progress of higher-performing students.

With regard to the mathematics benchmark test, efforts are being currently made to align the content across the different sub-components. The goal is to reduce duplication in the assessment of specific skills and formats so as to shorten the test to be more amenable to administration during one class period. Further, the scoring of the open-response pre-algebra items, where item scores are based on the number of correct steps followed in solving the problems, is somewhat cumbersome and may reduce teachers' willingness to administer this portion of the measure. Thus, future work will involve investigating the feasibility of adapting these open-ended response items to a multi-step, multiple-choice format.

Relative to the writing benchmark test, the writing passages were scored by research project staff that had extensive training and communication regarding scoring criteria before actually scoring the research data. The goal, however, would be to have a rubric that teachers could score with minimal effort. Thus, as the scoring rubric is being refined, a manual for teachers is being developed utilizing examples and non-examples from the current writing samples. Upcoming examinations will include assessing whether teachers can attain inter-rater reliability simply by using the scoring guide provided with the writing benchmark test.

Finally, in upcoming assessments, all benchmark tests will be administered to the same sample of students in order to assess the potential of crossover effects between different academic domains. Specifically, as noted above in the discussion of the application of

computation skills to more complicated word problems, the amount of unique variability in performance on mathematical word-problems accounted for by the general reading comprehension level of the student must be determined. Collecting benchmark data on all students across all benchmark exams will allow an examination that, first, controls for the impact of general reading comprehension proficiency on mathematics performance before assessing the unique explanatory power of the students' mathematics knowledge.

In addition to general reading comprehension deficits, the potential impact of more specific content-area reading proficiency on the ability to successfully complete mathematical problems is also interesting. To this end, efforts are being focused on developing another sub-component of the mathematics benchmark test to include an assessment of more directed content-area reading skills. Such a focus has been noted by others as an important contributor to success within the area of mathematics (Jones, 2001) and would help to clarify whether an application deficit or a reading proficiency problem should be the primary focus of remediation. These types of crossover examinations, with their associated results, will inform schools relative to the interactive impact of skill deficits and assist in better specifying the remediation needs of students.

References

AERA (2000). American Educational Research Association Position Statement Concerning High-Stakes Testing in PreK-12 Education. *AREA Statements and Reports,* July, 2000, 1-5.

American Federation of Teachers. (1998). *Redesigning low-performing schools: Its union work,* American Federation of Teachers: Washington D.C.

Anderman, C., Cheadle, A., Curry, S., Diehr, P., Shultz, L., & Wagner, E. (1995). Selection bias related to parental consent in school-based survey research. *Evaluation Review, 19*(6), 663-674.

California Department of Education (2001). *California High School Exit Exam*. Sacramento, CA: Standards and Assessment Division.

Carmines, E.G., & Zeller, A.Z. (1979). *Reliability and validity assessment.* Sage University Paper series on Quantitative Applications in the Social Sciences. Beverly Hills and London: Sage Publications.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of test. *Psychometrika, 16*: 297-334.

Dent, C.W., Galaif, J., Sussman, S., Stacy A., Burtun, D., & Flay, B.R. (1993). Demographic, psychosocial and behavioral differences in samples of actively and passively consented adolescents. *Addictive Behaviors, 18*(1), 51-56.

Ellickson, P.L., & Hawes, J.A. (1989). An assessment of active versus passive methods for obtaining parental consent. *Evaluation Review, 13*(1), 45-55.

Harcourt Educational Measurement (2000). *Stanford Achievement Test-9th Edition.* San Antonio, TX: Psychological Corp.

Howell, K. W., Zucker, S.H., & Morehead, M.K. (1994). *Multi-level academic skills inventory (MASI).* Paradise Valley AZ: H & Z Publications.

Iverson, A. M., & Cook, G. L. (1994). Guardian consent for children s participation in sociometric research. *Psychology in the Schools, 11*, 108-112.

Jones, C.J. (2001). CBAs that work: Assessing Students Math Content-Reading Levels. *Teaching Exceptional Children,* Sept/Oct, p. 24-28.

Linn, R.L. (2000). Assessments and Accountability, *Educational Researcher, 3*, 4-16.

Madaus, G.F. (1985). Public policy and the testing profession: You ve never had it so good? *Educational Measurement: Issues and Practice, 4*(4), 5-11.

Mayer, R.E. (1993). Understanding individual differences in mathematical problem solving: Towards a research agenda. *Learning Disabilities Quarterly, 16*, 2-5.

Meyer, R.H. (1996). Comments on chapters two, three and four. In H.F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp.137-145)l Washington D.C.: The Brookings Institution.

Noll, R.B., Zeller, M.H., Vannatta, K., Bukowski, W.M. & Davies, W.H. (1997). Potential bias in classroom research: Comparison of children with permission and those who do not receive permission to participate. *Journal of Clinical Child Psychology, 26*(1), 36-42.

Novick, M. & Lewis, G. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1-13.

Robbins, M. (2001). The failure of testing. *Salon.com.* www.salon.com/mwr/feature/2001/05/11/test_revolt/print.html, 1-9.

Severson, H.H, & Ary, D.V. (1983). Sampling bias due to consent procedures with adolescents. *Addictive Behaviors, 8*(4), 433-437.

Thorndike, R.L., & Hagen, E. (1961). *Measurement and evaluation in psychology and education.* New York: Wiley & Sons, Inc.

Wise, L.L., Sipes, D.E., George, C.E., Ford, J.P., & Harris, C.D. (2001). California high school exit examination (CAHSEE): *Year 2 evaluation report (HumRRO Interim Evaluation Report IR-01-29r).* Alexandria, VA: Human Resources Research Organization.

Xin, Y.P, & Jitendra, A.K. (1999). The effects of instruction in solving mathematical word problems for students with learning problems: A meta-analysis. *The Journal of Special Education, 32*, 207-225.

Table 1

Sample Descriptives: Performance Levels

	Mean (sd)	Range	Skew(se)	Kurt(se)
Math Sample 1:				
SAT9 Math	18 (8.23)	0-41	.29 (.20)	.44 (.40)
HSEE Math	331 (24.0)	283-399	.57 (.20)	.24 (.39)
Math Sample 2:				
SAT9 Math	20 (11.0)	0-41	-.20 (.26)	-.51(.51)
HSEE Math	343 (26.0)	287-399	.22 (.26)	-.45(.52)
Reading Form A Sample:				
SAT9 Reading	43 (15.0)	0-76	-.36 (.23)	.60 (.46)
HSEE Language Arts	354(25.0)	282-409	-.43 (.24)	.75 (.47)
Reading Form B Sample:				
SAT9 Reading	44 (13.0)	0-74	-.26 (.23)	.09 (.45)
HSEE Language Arts	351 (25.0)	279-413	-.08 (.24)	.19 (.47)
Writing Sample:				
SAT9 Writing	25 (9.0)	0-43	-.52 (.16)	.21 (.33)
HSEE Language Arts	352 (25.1)	279-413	-.26 (.17)	.40 (.34)

Table 2

Reliability Estimates for the Reading Comprehension and Mathematics Benchmark Tests

Scale Name	Alpha
Reading Comprehension	
Form A: Grade Level	
4th Grade Passage	.88
6th Grade Passage	.97
8th Grade Passage	.98
Total Form A	.97
Form B: Grade Level	
4th Grade Passage	.89
6th Grade Passage	.97
8th Grade Passage	.98
Total Form B	.97
Mathematics	
Computation	.73
Multiple-Choice Word Problems	.75
Open-Ended Response Problems	.91
Total Math .93	

Table 3

Mean Differences for Expository and Narrative Writing Prompts

Scoring Criteria	Prompt A Mean	Prompt B Mean	t-value	p-value
# Of Correctly Structured Sentences	5.3	4.0	5.0	.000
Total Number of Words	153.2	130.7	6.5	.000
Total Number of sentences	9.8	7.1	9.7	.000
Overall Clarity	3.3	2.3	7.1	.000
Run-On Sentences	3.8	3.0	4.6	.000
Word Usage	2.3	1.7	3.5	.001
Punctuation	2.4	2.1	1.6	.101
Spelling	3.3	3.4	-.32	.749
Sentence Fragmentation	4.2	4.5	-2.1	.038
Capitalization	3.9	2.2	8.2	.000

Table 4

Factor Analytic Results for Writing Scoring Rubric

Scoring Criteria	Factor Loadings			
	1	2	3	4
Number of Correctly Structured Sentences	.75	--	--	--
Total Number of Words	.79	--	--	--
Total Number of sentences	.93	--	--	--
Overall Clarity	--	.74	--	--
Run-On Sentences	--	.70	--	--
Word Usage	--	.62	--	--
Punctuation	--	--	.69	--
Spelling	--	--	.73	--
Sentence Fragmentation	--	--	--	.52
Capitalization	--	--	--	.87

Table 5

Descriptive Statistics for the Total Benchmark Scores

Variable Name	Mean (sd)	Skew (se)	Kurtosis(se)	Range
Writing Benchmark Factors:				
Quantity: Production	155.0 (62.89)	.40 (.19)	-.49 (.38)	40.0-306.5
Quality: Clarity	1.6 (.58)	.27 (.19)	-.24 (.38)	.40-3.0
Quality: Technical	1.1 (.47)	-.24 (.19)	-.58 (.38)	.00-2.0
Quality: Sentence Structure	1.5 (.37)	-.32 (.19)	.04 (.38)	.40-2.0
Total Writing Benchmark	159.2 (61.88)	.41 (.19)	-.49 (.38)	45.3-312.4
Mathematics Benchmark Scores:				
Mathematics True Score	1.6 (.33)	-.30 (.36)	.19 (.71)	.74-2.3
Mathematics Estimated Score	1.5 (.39)	-.45 (.20)	.00 (.39)	.49-2.2
Reading Benchmark Total	76.3 (19.37)	.08 (.45)	-.22 (.22)	33.0-107.0

Table 6

Correlations Among Benchmark Test Scores and Outcome Scores

	SAT9 %	Exit Exam
Benchmark Measure*:	Reading	Reading
Reading Benchmark	.38***	.49***
	Mathematics	Mathematics
Mathematics True Score	.80***	.78***
Mathematics Estimated Score	.59***	.54***
	Writing	Writing
Writing Benchmark	.39***	.39***

*Reading N=235;

Mathematics True N=89

Mathematics Estimated N=154

Writing N=141

*** p<.001

Table 7

Stepwise-Regression Results for the Prediction of High-Stakes Outcomes

	§	t (p)	Cumulative % Variance
Outcome: SAT9			
Reading Benchmark:			
Grade 6 Reading Passage	.57	5.7 (.000)	13%
Math Benchmark True Sample:			
Open-Response Word Problems	139.28	4.3 (.000)	53%
Multiple-Choice Word Problems	53.23	2.2 (.031)	57%
Math Benchmark Estimated Sample:			
Open-Ended Word Problems	149.99	8.7 (.000)	34%
Writing Benchmark:			
Production Factor	.17	5.6 (.000)	12%
Clarity Factor	15.18	4.9 (.000)	25%
Sentence Structure Factor	10.00	2.0 (.046)	26%
Prediction Outcome: HSEE			
Reading Benchmark:			
Grade 6 Reading Passage	.07	8.0 (.000)	24%
Math Benchmark True Sample:			
Open-Response Word Problems	134.25	4.6 (.000)	56%
Multiple-Choice Word Problems	43.79	2.1 (.047)	59%

Table 7, continued

	§	t (p)	Cumulative % Variance
Math Benchmark Estimated Sample:			
Open-Ended Word Problems	93.29	2.7 (.001)	42%
Multiple-Choice Word Problems	44.01	2.2 (.026)	44%
Prediction Outcome: HSEE			
Writing Benchmark:			
Clarity Factor	.11	4.0 (.000)	11%
Production Factor	.02	3.9 (.000)	20%
Technical Factor	.06	2.1 (.041)	22%

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
X	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").