

THE NEW HAWAII CONTENT AND PERFORMANCE STANDARDS II STATE ASSESSMENT

MICHAEL HEIM

"Assessment and accountability is where we say that our expectations for the students and for ourselves are serious, and that we will take them seriously."

"It is inconceivable that we would ask teachers to teach to the standards and then assess on something else. Yet that is a very realistic and all too common example of the lack of coherence that we seek to correct..."

Dr Paul G LeMahieu; Superintendent's Education Leadership Conference; August 10, 1999

This article describes the design, development and uses of the new Hawaii Content and Performance Standards II State Assessment, the Department of Education's statewide assessment that measures, in part, student achievement relative to selected portions of the Hawaii Content and Performance Standards II (HCPS II). The article also presents some of the key processes used in the test's construction, provides details about its major characteristics, and outlines the steps that remain to be taken to complete it.

THE NEED FOR A NEW ASSESSMENT

The new assessment is an integral part of the effort to transform Hawaii's traditional K-12 public education system into a standards-based system. In early 1999, the State Board of Education adopted a Comprehensive Needs Assessment (Hawaii Department of Education; April 1999) to establish goals and priorities for the whole of the system. Following the needs assessment, a Strategic Plan for Standards-based Reform (Hawaii Department of Education; September 1999) was developed to address the priorities established by the Board and build upon the "Images of Success" articulated in the needs assessment. The images of success are consistent threads that connect the needs assessment, the Board's priorities, the strategic plan and the specific actions within the task areas delineated in the strategic plan. Thus, much of the impetus for the student assessment and accountability systems outlined in the strategic plan can be traced back to the "Standards-Based Learning" images of success developed in the needs assessment.

In order to improve student assessment, the strategic plan outlined two broad strategies and a corresponding two-tiered assessment system: (1) a redesigned statewide assessment program called the Hawaii Assessment

Program, and (2) school/classroom assessment programs along with supports for those local activities from the state program. Together, both state and classroom assessments are seen as forming a balanced and comprehensive assessment system. As classroom assessment expert and advocate, Rick Stiggins (1999), has noted: "If assessment is not working effectively in our classrooms every day, then assessment at all other levels (district, state, national, or international) represents a complete waste of time and money" (p. 193). No single assessment can adequately provide the range of assessment information needed by students, teachers, counselors, principals, support personnel, policy makers and public.

The statewide tier of the assessment system, the Hawaii Assessment Program, is composed of two parts. The first part, the most publicly visible portion, is the new HCPS II State Assessment. The second part, the new School Assessment Liaison Program, is designed to help schools improve their "assessment literacy" efforts, particularly in the use of sound classroom assessment practices. This program and the Assessment Matters website (<http://assessmentmatters.k12.hi.us>) are the primary means of state support for schools' assessment programs and classroom assessment practices. I will devote the remainder of this article, in spite of interesting developments in the School Assessment Liaison Program, to a description of the first part, the HCPS II State Assessment.

A NEW TEST FOR HAWAII'S STUDENTS

For more than two decades, until 2000, the Stanford Achievement Test © Harcourt Educational Measurement, an "off-the-shelf," norm-referenced achievement test series served as the Department's primary student achievement measure. The HCPS II State Assessment, which incorporates a small portion of the Stanford series, shares some of the characteristics and purposes of the former, but extends those in ways appropriate to standards-based education. In broad terms, the new assessment has been designed to provide:

- Annual data on student, school, and system performance at benchmark grade levels;
- Fair, technically rigorous, and adequate measurement of performance against standards; and,
- Measures of student achievement relative to both the HCPS II and national norms.

These changes are in line with the Board of Education's policy that, "The Department of Education shall establish a statewide assessment program that provides annual data on student, school, and system performance, at selected benchmark grade levels, in terms of student performance relative to the Hawai'i Content and Performance Standards and relative to nationally representative norms" (#2520, State Assessment Program Policy). In addition, there was a strong desire to have an assessment that would be experienced by students and school staff as one coherent, seamless whole, rather than a series of disparate pieces. The new assessment was, therefore, deliberately designed as an integrated package to provide both norm-referenced and criterion-referenced (standards-based) information.

The new assessment measures student achievement relative to the HCPS II and provides a view of Hawai'i's students' mastery of the content standards through their performance on criterion-referenced (standards-based) items. Thus, student performance is compared to the criteria given in the standards. It also ensures consonance between what is tested and what is taught. Later, when statewide student assessment data is used in accountability systems, only the standards-based assessment information will be used.

Achievement on the Stanford Achievement Test's norm-referenced items provides a view of Hawai'i's students' performance compared with students in a nationally representative norm group. It also provides an external perspective to partially corroborate performance against the HCPS II. However, given the imperfect alignment of Stanford items with the HCPS II, the norm-referenced information is not to be used in Hawaii's accountability systems.

INTENDED USES

Validity – the most important quality of a measure – is often mistakenly thought of as a technical characteristic inherent in the measure or measuring instrument itself. Actually, it is the soundness of the inferences or interpretations and uses of the measure that is the focus of validity. Thus, one cannot adequately judge the validity of a measure without a clear and explicit understanding of the purposes for which the measure was designed. The HCPS II State Assessment was designed for the following uses:

- To monitor student achievement relative to the HCPS II and national norms;
- To inform systematic improvements to curriculum and instruction (at the programmatic level) and schools' Standards Implementation Design (SID) plans and strategies;
- For school accountability, as an indicator for initial classification;

- For student accountability (rewards/recognition, assistance); and
- As an high school diploma requirement.

Additionally, the state assessment serves as an operational standard for defining an "equivalent" alternative assessment. I will say more about this use later, but for now I will offer some more details to clarify and expand upon these intended uses.

It is important to stress that the HCPS II State Assessment is not a classroom assessment system and should not be used in that way. Classroom or school assessments are valuable in guiding teachers' ongoing instruction and in assisting students to meet the standards. Results of school and classroom assessments at every grade level play a critical role in knowing how well students are learning the standards in the intervals between benchmark points and how to help with appropriate instruction. The state assessment operates at a different level by providing valuable information for conducting, for example, large-scale reviews (including school-wide reviews) of curriculum and instruction at the programmatic level. But when it comes to informing day-to-day classroom instruction and curriculum, decisions must be served by classroom assessments. Given these limitations, the standards-based or criterion-referenced results from the HCPS II State Assessment might, nevertheless, be a useful trigger to school-wide programmatic and professional development discussions. For example, they may be used to identify needs for staff development in specific instructional strategies or to review the articulation and coherence of curriculum across grade levels and within grade levels or departments.

The use of state assessment results for accountability is a complex and extensive topic that warrants a separate paper. Here, I can provide only a broad-brush sketch of school and student accountability plans in order to convey the intended uses of the state assessment in that context. Additional details can be found in the strategic plan (Strategic Plan for Standards-based Reform, Hawai'i Department of Education, September 1999). What follows is a tentative outline of school and student accountability intents, conceptual plans, and timelines.

As far as school accountability is concerned, the HCPS II State Assessment is anticipated to serve as an initial indicator of school improvement. The orientation is constructive, not punitive, and focuses on continuous improvement with a full range of consequences for observed performance. A major challenge in designing the implementation steps (which are still under development) is to balance fairness with complexity. Initial indicators will focus on students learning the standards, and, possibly, on students' safety and well-being. SAT-9 norm-referenced

scores will not be used for school accountability. Rather, standards-based proficiency scores, derived from those items that assess the HCPS II and show the proportion of students meeting or exceeding standards, will be used as indicators of student learning. Schools that have not shown agreed-upon progress on the initial indicators, as validated during an on-site visit by a School Review Team, might be classified as "assistance schools." By 2005, schools would be classified into one of three categories: reward, recognition, or assistance. A tentative schedule for the implementation of school accountability follows:

Proposed School Accountability Implementation Schedule

- 1999-2000: Complete assessment development (except for setting performance levels).
- 2002-2003: Establish baseline for initial school accountability indicators.
- 2004-2005: Establish first comparison point. Corresponding accountability consequences would include rewards and assistance only (with no sanctions at this point).
- 2006-2007: Establish second comparison point with a full range of consequences – rewards, assistance, and sanctions.

Note that the above schedule includes a baseline, a first comparison point, and a second comparison point. Each is two years in length. Thus, "2002-2003" denotes two school years, those that end in the calendar years 2002 and 2003. Lessons learned from school accountability efforts in other states argue strongly for two-year rather than one-year data cycles. This is due to confounding factors such as cohort effects (variation in year-to-year school-aggregate achievement results attributable solely to differences in the groups of students assessed each year). It also avoids the necessity of developing a separate methodology for small schools, especially those with less than 25 students enrolled per grade level.

It may be noted that school year 2001 does not appear in the above schedule, a result of the April 2001 teachers' strike and the cancellation of what was to be the first "live" or operational administration of the HCPS II State Assessment. Consequently, it has been necessary to reschedule all the activities that depended upon the 2001 statewide assessment.

The state assessment is expected to serve as an indicator to identify students with non-proficient performance who may need extended learning opportunities and to identify, for the purposes of recognition and reward, students with proficient performance. A tentative schedule for student accountability implementation is as follows:

Proposed Student Accountability Implementation Schedule

- 2005: "Mid-stakes" accountability - Students would be expected to demonstrate proficiency in order to participate in the state assessment at the next benchmark level.

- 2008: "High stakes" accountability - Students would be expected to demonstrate proficiency on the state assessment administered in high school in order to receive a High School diploma (pending Board review and action).

Standards-based proficiency scores are to be the basis of "mid-stakes" accountability for students as of 2005. SAT-9 norm-referenced scores will not be used to show student mastery of standards. Rather, standards-based proficiency levels, derived from the criterion-referenced portions of the assessments, will be used for that purpose. A student in Grade 5 may not take the Grade 5 benchmark assessment if proficiency on the Grade 3 assessment has not been demonstrated. Similarly, students in Grades 8 or 10 must demonstrate proficiency on the prior benchmark assessment to be eligible to take the Grade 8 or 10 assessment. This does not mean necessarily that the student is retained in grade. However, the student will not take the next benchmark test until proficiency in the prior one is demonstrated. Voluntary or mandatory learning opportunities (a wide range is possible, and preferred options and decisions are best made at the school level) must be provided in order for students to be held accountable for meeting state standards.

DEVELOPMENT PROCESS AND KEY REVIEWS

The following chart summarizes the key tasks for developing the new assessment through completion of the final assessment forms. Since the development of the new state assessment is an enormous project requiring multi-faceted capabilities and expertise, Harcourt Educational Measurement, publisher of the Stanford Achievement Test series, was contracted to assist. Harcourt will serve as the Department's developer and publisher for the HCPS II State Assessment, and they will also provide related scoring and reporting services.

Task	Work Period or Completion Date
Develop assessment "blueprint." Develop draft specifications for the HCPS II based assessments in reading, writing, and mathematics at grades 3, 5, 8, and 10.	Sept. 1999
Develop items as specified in the assessment blueprint.	Sept. - Dec. 1999
Conduct item review sessions. (Review criteria: Congruence with standards, instructional sensitivity, absence of out-of-school factors, absence of bias) Conduct field test (including user surveys, focus groups). Administer SAT-9 Abbreviated and the standards-based field-test forms to students.	Jan. 2000 May 1-15, 2000
Construct final forms. Construct final draft forms (2 per grade level in reading and in mathematics) for the operational or "live" HCPS II State Assessment. Construction of the final assessment forms included an item data review using data from the field test, as well as a forms bias review by a community panel.	Sept. 2000

Several points about the assessment blueprint may be of general interest. The blueprint required that all items for the standards-based segments of the assessment be newly

written. They were not to be obtained from previously developed assessments or from item banks. In the reading section, the blueprint called for reading passages of authentic literature: previously published pieces, often (but not exclusively) by Hawai'i authors. Given the very short item development timeline, this specification proved particularly challenging. Future editions of the assessment will use commissioned reading passages, too.

Of course, the blueprint's specifications involved practical compromises. Most notably, these related to the level of detail to be derived from the scores. While a strong argument can be advanced for designing the assessment so that it produces a highly reliable score for each of the standards measured, that approach would have led to an unacceptably long and time-consuming assessment for the three reading standards, two writing standards, and 14 mathematics standards measured. In addition, not all of the six reading and six writing content standards can be appropriately assessed in a large-scale assessment setting. For example, the reading standard "Students will read a range of literary and informative texts for a variety of purposes," can be assessed using a series of classroom assessments during the year, but not by means of a once-a-year, on-demand, pencil-and-paper assessment. In reading and mathematics, the compromise was to design the assessment to provide highly reliable "standards-referenced" proficiency level scores for the content area along with moderately reliable "indicators" or sub-scores for each of the three reading content standards assessed and for each of the five mathematics strands. These latter included the main categories into which the mathematics content standards are organized: numbers and operations; measurement; geometry and spatial sense; patterns, functions, and algebra; data analysis, statistics, and probability.

In January 2000, the sets of draft items prepared by Harcourt were subject to an intensive item review by the Department. Eighty-three (83) Hawai'i public school educators served as panelists. Panelists were teachers (regular education, special education, English for Second Language Learner, Title I), School Renewal Specialists, and school administrators drawn from all levels (elementary, middle/intermediate, high school) and geographic regions (Oahu and neighbor islands). Over the course of three intensive day-long sessions, eight item review panels (one each per benchmark grade level – 3, 5, 8, and 10 – for reading and for mathematics separately) reviewed draft items using four key criteria that Dr W James Popham had developed for our use. They included standards congruence; instructional sensitivity; absence of out-of-school factors; and absence of bias. Dr Popham also kindly helped with training panelists in the use of the criteria for the review. For each item reviewed, the panels had the authority to accept the item, to accept the item with an "on

the spot fix" or with specific recommendations for later modification, or to reject the item.

Only non-rejected items were used in the field test. Following the field test, an in-depth item data review was conducted in September 2000. Thirty-one (31) Hawai'i public school educators, with characteristics similar to the panelists who conducted the item review, met for four consecutive days in four item data review groups. They reviewed the field test data and the field test items for compliance with item/form specifications, for statistical quality, for adherence to criteria for writing multiple choice and constructed response items. They also reviewed, once more, for congruence of the items with the standards assessed and the benchmarks corresponding to those standards. The item data review groups decided whether to accept or reject items (and, for reading items, the associated reading passages) for inclusion into the final test forms.

Immediately following the item data review sessions, drafts of the final forms were assembled. The draft forms were then reviewed for overall "forms bias" by a community review panel. Forms bias, while similar to item bias, can occur at the level of a complete form and may not be detected during an item review. For example, it may take the form of gender stereotyping where girls are generally portrayed, over the course of an entire form, in passive roles and boys in active roles.

The community review panel was composed of twelve (12) members who met for two consecutive days. They represented a variety of stakeholder sectors, e.g., higher education, Board of Education, Superintendent's Education Cabinet, Office of the Attorney General, Hawaii State PTSA, The Hawaii Chamber of Commerce/Small Business, military community, Filipino Chamber of Commerce, Hawaiian language advocate/specialist, parent/City & County of Honolulu official, teacher/HSTA member. Most panelists had multiple work and community roles. As a group, the community panelists represented a variety of demographic characteristics (ethnicity, gender, island residence) and a full range of expertise with large-scale student assessment. They reviewed each final test form as a whole in terms of bias or stereotyping and recommended accepting or rejecting each final test form.

As a result of the above development processes, two final forms were produced for each grade level benchmark in reading and in mathematics. (A somewhat different but roughly parallel set of processes was used in the development of the writing assessment prompts, rubrics, and sets of training and anchor papers.) For actual use, one final form was to be used in operational administrations and the second form held in reserve as a "breach" form in the event of assessment administration or security problems.

MAJOR CHARACTERISTICS

Listed immediately below are the major characteristics of the new assessment. An outline style is used to facilitate reading and for reference.

Content areas:

- Reading
- Writing
- Mathematics
(Other areas, such as science and social studies, may be phased in over the years.)

Benchmark grade levels:

- Grade 3 -Assesses mastery built from Grades K-3
- Grade 5 -Assesses mastery built from Grades K-3 plus 4-5
- Grade 8 -Assesses mastery built from Grades K-3 plus 4-5 plus 6-8
- Grade 10 -Assesses mastery built from Grades K-3 plus 4-5 plus 6-8 plus 9-12

(Note: Students may begin "challenging" the high school assessments in Grade 10.)

Standards-based content measured directly:

- Reading: Response to Text, Comprehension Processes, Conventions and Skills
- Writing: Rhetoric, Conventions and Skills
- Mathematics: Numbers and Operations; Measurement; Geometry and Spatial Sense; Patterns, Functions, and Algebra; Data Analysis, Statistics, and Probability

Norm-referenced content measured via the SAT-9 Abbreviated:

- Reading Comprehension (30 items)
- Mathematics Problem Solving (30 items)

Note: Selected SAT-9 Abbreviated items also contribute to the standards-based Reading and Mathematics scores.

Number of "Double Duty" SAT-9 Items		
Grade	Reading	Mathematics
3	8	29
5	9	26
8	10	18
10	16	8

These "double duty" SAT-9 items were identified by a panel of curriculum and measurement specialists as satisfying the requirements of the assessment blueprint and the criteria of "standards congruence."

Item formats:

- For the standards-based Reading and Mathematics segments
 - Multiple-choice items: 1 score point each
 - Constructed response items of various lengths: 2, 3, or 4 score points each
- For standards-based Writing assessment (essay writing to a prompt)

- Scored on five dimensions (Meaning, Voice, Clarity, Design, Conventions) using rubrics with scales of 1-5 points for each dimension
- For SAT-9 Abbreviated Reading and Mathematics segments
 - Multiple-choice items: 1 score point each

Types of scores:

- Standards-based proficiency scores: Exceeds proficiency, meets proficiency, approaches proficiency, well below proficiency
- Norm-referenced scores (e.g., stanines, national percentile ranks)

Number of administration sessions:

- Reading: 3 sessions
- Writing: 1-2 sessions (1 session for Grade 3 and 5)
- Mathematics: 3 sessions

Total time requirements:

- Total time required: 6 hrs. 51 min. (grade 3) to 8 hrs. 46 min. (grade 10), which includes time for distribution and collection of materials, reading directions to students and, for the writing assessment in Grades 8 and 10, a 10-minute rest break between sessions.
- The standards-based segments have suggested administration times but, strictly, those segments are not timed. The SAT-9 Abbreviated segments must follow the publisher's standardized time requirements in order to produce valid norm-referenced scores.

(Note: The total time requirements reflect, in part, the design characteristics of the standards-based segments of the assessment and are larger than what would be expected based on experience with multiple choice tests. On the standards-based segments, students often produce a response rather than only select a response. The standards-based segments are intended to function as "power" rather than timed tests. For Grades 8 and 10, the writing assessment includes a revision process, and that takes more time than a first-draft only writing sample would require).

Provisions for special populations:

- IDEA-eligible & Section 504 students
 - SAT-9 accommodations
 - Standards-based segment accommodations
 - Alternative Assessment
- ESLL students
- Hawaiian Language Immersion Program students
- Students with serious disciplinary action status
- Students on home/hospital instruction

(Note: Appropriate accommodations and alternatives to the HCPS II State Assessment must be provided in order to enable all students to participate).

Special materials and equipment needed:

- None

Materials packaging:

- Student assessment materials will be packaged as consumables

- Reading (both standards-based and norm-referenced segments) and writing will be packaged as one language arts assessment booklet at the elementary level and as two booklets (reading, writing) at the secondary level.
- Both mathematics segments (standards-based and norm-referenced) will be packaged as one mathematics assessment booklet.
- A separate scannable answer document will be used at all grade levels.

Scoring:

All scoring will be conducted externally by Harcourt Educational Measurement. Harcourt has been contracted to serve as the Department's developer and publisher for the HCPS II State Assessment and also provides related scoring and reporting services.

Reports:

- For the 2002 assessment, two "waves" of reports will be provided.
 - Interim consolidated reports that contain norm-referenced SAT-9 scores and raw (total points) scores for the standards-based assessments. These reports are "interim," pending the completion of work to establish cut-scores and the associated proficiency levels. (Aug. 2002)
 - Final reports to the schools, district, and state. Contents of these updated reports for the 2002 assessment will show proficiency scores for the standards-based assessments, as well as the previously reported SAT-9 scores. These reports will become the "model" for future assessment reports. (Nov. 2002)
- 2003 & thereafter: One set of consolidated reports annually (August)

STEPS REMAINING

Results from the first "live" administration of the HCPS II State Assessment were to be used in setting proficiency levels (exceeds, meets, approaches, and well below proficiency) for the standards-based reading and mathematics assessments. The April 2001 teacher's strike, however, necessitated canceling the assessment. Consequently, the remaining development work, shown in the chart below, had to be deferred one year. A revised schedule is provided below.

Task	Work Period or Completion Date
Administer the HCPS II State Assessment.	Apr. 2002
Score the assessment and produce interim reports. Distribute interim reports to the schools, districts, and state offices. Contents of these interim consolidated reports will contain norm-referenced SAT-9 scores and raw scores (total points) for the standards-based assessments.	Aug. 2002
Set proficiency levels. Recruit school staff and community members to serve on proficiency level setting panels.	Aug. 2002
Secure the Superintendent's and Board of Education's review and approval of the proficiency levels recommended by the panels.	Oct. 2002
Produce final reports. Generate updated reports for the 2002 assessment and distribute reports to the schools, districts, and state offices. Contents of these updated consolidated reports will contain norm-referenced SAT-9 scores and proficiency scores for the standards-based assessments. (Begin establishing baseline data for the school accountability system.)	Nov. 2002

WHAT COMES NEXT?

Work has already started on developing "Generation II," the next edition of the state assessment which is scheduled for operational use in 2004. Initial item reviews for that edition will be held in fall 2001. Items for Generation II will be field tested by embedding and spiraling them in student's booklets in the 2002 administration, obviating the need for a separate stand-alone field test administration. Work is also underway to develop "alternative" assessments to the HCPS II State Assessment. The HCPS II State Assessment (Generation I), now almost completed, serves as the operational standard – in terms of breadth of coverage, depth of academic rigor, and technical quality – which proposed "equivalent" alternatives to the state assessment will be evaluated. The HCPS II State Assessment should not be confused with the Alternate Assessment, which is used for students for whom the HCPS II is inappropriate. Work in two areas of alternative assessment is in development. First, planning is being conducted for a feasibility study to assess the prospects for developing Hawaiian language versions of the state assessment. Secondly, an exploratory study has been started that is looking at "equivalent" alternative ways by means of which students could fairly, credibly, and validly demonstrate their attainment of the standards.

We are, finally, near the point where it will be truly "... inconceivable that we would ask teachers to teach to the standards and then assess on something else" (Dr Paul G LeMahieu; Superintendent's Education Leadership Conference; August 10, 1999). But this is only one part of the Hawai'i Assessment Program. The other essential part will require enhancement of teachers' assessment literacy and their use of sound classroom assessment practices. Both parts are vital to the success of standards-based education – in every classroom, for all students, every day.

REFERENCES

Stiggins, Richard (1999). "Assessment, Student Confidence, and School Success," *Phi Delta Kappan*, November, p 193).

Michael Heim is director of Planning and Evaluation Branch Planning in the Budget and Resource Development Office of the Hawai'i Department of Education.