California Council for the Social Studies

www.ccss.org

OCCASIONAL PAPERS

Volume 2, Number 1 Fall 2013

The Case for Assessment Reform for California Schools

by

Jim Hill
College of Education
Program of Educational Administration (Ret)
Adjunct Faculty, Educational Leadership Doctoral Program
California State University San Bernardino

California is exploring new assessment systems that will substantially alter accountability processes and procedures, as well as testing itself. As a governing member of one of two large multi-state consortiums designing new assessments, California is hoping to see the first set of new tests in place by spring 2015. These assessments will test English/Language Arts (ELA) and math, in those grades required by federal law (grades 3-8 and one year in high school). But California is also looking to go much beyond that into a new and quite unique 'further generation' assessment design, surpassing those of the consortium assessments. Specifically, California is looking to include ELA and math grades not tested by the consortium tests - the grades in which science is not tested - as well as the rest of the school core curriculum. History/Social Science (also called history/social studies, or simply social studies outside of California) will be included in this 'further generation' assessment program, which will take shape and begin to be put in place after 2016. In the California 'high stakes' testing world, the goal is to have the 'thinking-based' multidisciplinary assessment program promote a 'thinking-based', multidisciplinary instructional program. History/Social Science educators and organizations can and should be involved in creating these assessments.

Recent legislation (Assembly Bill 484, signed into law on October 2, 2013) has started the process of change. First, the law cancels almost all of the existing state standards tests (California Standards Tests, or CSTs) effective spring 2014. Second, the law authorizes field testing and implementation of the new Common Core State Standards-based assessments in ELA and math for grades 3-8 and 11, which are being created by the consortium of states called the Smarter Balanced Assessment Consortium (SBAC). In California, this assessment program will be called the California Measurement of Academic Performance and Progress for the 21st Century, with the acronym CalMAPP21, or simply MAPP. And thirdly, where California looks to go further, the law requires the California Department of Education (CDE) - the original author of much of

AB 484 - to present a plan that would create an additional assessment program by March 2016. As the law states, the plan is to

> "…include additional assessments….The recommendations shall consider assessments in subjects, including, but not necessarily limited to, history-social science, technology, visual and performing arts, and other subjects as appropriate, as well as English language arts, mathematics, and science assessments to augment the [SBAC] assessments….over several years, the use of matrix sampling, if appropriate, and the use of population sampling." (1)

These new tests will go beyond SBAC assessments in scope and design. These 'further generation' assessments, in addition to SBAC's, are the ones that will promote classroom instruction that encompasses '21st century' skills (critical thinking and problem solving, communication, collaboration, creativity and innovation) while at the same time, reducing and eliminating many of the negative consequences of the state's current accountability system. They are necessary to support the kind of educational program called for by U.S. Secretary of Education Arne Duncan, among many others, who proclaimed in 2010 that

> "…I reject the notion that the arts, history, foreign languages, geography, and civics are ornamental offerings that can or should be cut from schools during a fiscal crunch. The truth is that, in the information age, a well-rounded curriculum is not a luxury but a necessity….A well-educated student, in other words, is exposed to a well-rounded curriculum. It is the making of connections, conveyed by a rich core curriculum, which ultimately empowers students to develop convictions and reach their full academic and social potential….There is no doubt that math, reading, writing, and science are vital core components of a good education in today's global economy. But so is the study of history, foreign languages, civics, and the arts. And it is precisely because a broad and deep grounding in the arts and humanities is so vital that we must be perpetually vigilant that public schools, from pre-K through twelfth grade, do not narrow the curriculum." (2)

ACCOUNTABILITY

Assessment and especially accountability have been the watchwords in public education for decades. The 1983 report entitled *A Nation At Risk*, produced by the National Commission on Excellence in Education, resulted in public demand for more school accountability and higher academic expectations for all students. The then current California Assessment Program (CAP) of the 1970s and 1980s was not 'high stakes' because it did not include accountability. CAP was eliminated and was destined to be replaced by the California Learning Assessment System (CLAS) in the mid 1990s. CLAS was projected to make assessment more student performance- based and rely less on multiple choice tests typical of CAP. In fact, the California Department of Education's roll out conference for CLAS in 1989 was entitled 'Beyond the Bubble', and had many illustrations of performance-based types of assessment. Rather suddenly, however, CLAS was eliminated for political reasons by then Governor Pete Wilson in 1994. As a result, and in response to growing public and policymaker demands for accountability, the California state legislature quickly created a new testing system called the Standardized Testing and Reporting (STAR) in the latter 1990s with several bills requiring testing of all students in most grades and some subjects, beginning first with the use of 'off the shelf, nationally-normed' general multiple choice achievement tests. The national tests were replaced by state-created California academic content standards-based multiple-choice, or almost totally multiple-choice, achievement tests a few years later. The legislature next invented a school accountability system by creating

the Academic Performance Index (API), which used test scores to create a three-digit number between 200 and 1000 that became the measure of school quality. With API, the testing became 'high stakes' indeed. Schools were to be held accountable for student learning, and from early on, ultimate sanctions included replacing school staff. While the state tests were based primarily on the state academic content standards, those portions of the test that were aligned to the accountability system became the drivers of curriculum and instruction throughout California's K-12 public education system.

UNINTENDED CONSEQUENCES

Schools respond very precisely to that for which they are held accountable. The more sanctions for low performance that threaten staff job security, the stronger this focus. Indeed, one of the results of accountability has been to narrow school curriculum to what the accountability system measures and reports. Because English/Language Arts dominates both state and national accountability, followed by math at both levels, curriculum has narrowed to English and math. Teachers have commented in Common Core State Standards (CCSS) surveys and anecdotes that schools with relatively lower API scores narrow even more, especially at the elementary grades where the standards-based tests contain mostly basic skills questions or items, thus, resulting in a stronger focus on just developing basic skills. Even U.S. Secretary of Education Arne Duncan pointed out during his national tour speech in 2010 that, "...Almost everywhere…I heard people express concern that the curriculum had narrowed as more educators 'taught to the test,' especially in schools with large numbers of disadvantaged students." (3)

Assessment linked to accountability frames what happens in curriculum and instruction in schools. Teachers complain in these same surveys that they are required to engage in elaborate test preparation work to the exclusion of all else; an assessment system based almost totally on recall-type tests which produces a recall-based instructional program wrapped around whatever content is to be tested, which, in turn, becomes 'practice' for the expected yearly tests. Questions of equity, or access to the full curriculum, emerge as schools with lower scores, mostly those with students from lower-income families, are forced by prescribed remediation interventions to shorten or even eliminate instructional time on content outside of basic skills in reading and math. In essence, they become test preparation centers addressing basic skills. Schools with higher overall scores preserve more of the enriched curriculum content. Therefore, students emerging from the lower-scoring schools are less able to enter the career market of the 21st century than their peers who have participated in performances, applications of learning, large-scale academic competitions such as science fairs, history days, as well as music and art experiences, as these students have learned to use the skills their compatriots have not. (4)

THE NEED TO CHANGE ASSESSMENT AND ACCOUNTABILITY

The California Department of Education's *Report to the Governor and the State Legislature: Recommendations for Transitioning California to a Future Assessment System* of January 2013, and the memo entitled *A Long-Term Plan for the California Assessment System* of June 2013 to the State Board of Education from Educational Testing Service (ETS), propose a new approach to assessment that is bold and massive. Both claim that the current testing system is not adequate for a number of reasons. The current testing system has not advanced learning in the direction that students need to go. The weakness of the California Standards Tests and the role of assessment in driving instruction are clearly articulated by State Superintendent of Public Instruction (SSPI) Tom Torlakson in his introduction to the CDE *Report*:

> "The ability to engage in critical thinking and solve complex problems cannot be reliably assessed with the kinds of multiple-choice tests that are the centerpiece of our current system. The Common Core State Standards ask students to acquire deeper knowledge of the subjects they study and be able to perform more complex tasks using what they have learned….I believe this work provides us with the opportunity to develop new assessments that serve as models for the kind of high-quality teaching and learning necessary for a world-class education. The concept is simple but powerful: if our assessments require students to use problem solving and critical thinking skills to perform well, those same skills are much more likely to be taught in our classrooms day in and day out. The goals we set for our assessment system have profound implications for our students and our schools….California must plan for and develop a cohesive and adaptable assessment system that prepares its students for college and careers in the 21st century by focusing attention on building and assessing critical thinking skills across all subjects." (5)

The *Report* goes on further and states:

> "The current assessments have been criticized for not measuring students' achievement of the standards in sufficient depth. This is a fair criticism….The multiple-choice format also precludes measuring academic content standards that call for students to demonstrate more complex processes, such as critical thinking and problem solving, or application of knowledge in real-world settings. A legitimate concern is that when multiple-choice tests are used, in-depth understanding of subject matter is devalued because it is not easily measured. Likewise, critical thinking and complex problem-solving skills have the potential to become devalued…
>
> Assessing more complex instructional concepts would require different types of test items or questions that ask students to provide more complex responses and/or respond to more complex stimuli than the current assessments allow….Performance tasks are even more involved items that require students to complete a multifaceted assignment or project that demonstrates competence in a variety of areas and demonstrate the application of knowledge..." (6)

The *Report* criticizes the 'narrowing' of the curriculum, and points to the need to include other subjects, including History/Social Science, in a new assessment system.

> "The current system of assessments has also been criticized for negatively influencing instruction through the narrowing of the curriculum to only those subjects that are tested, certainly an unintended consequence. Currently, ELA [English language arts] and mathematics are tested at every grade from two through eleven. In the elementary grades, science is tested less than either of these subjects, and history–social science is tested even less. (7)
>
> While [the new Common Core Smarter Balanced Assessment Consortium] SBAC assessments will be an integral part of California's future assessment system, the system must expand beyond SBAC by providing assessments of subjects other than ELA and mathematics (e.g., science and history–social science). (8)
>
> To achieve these benefits across the curriculum (e.g., science and history–social science), the state will need to invest resources to administer these types of assessments." (9)

This criticism of California tests, as well as the use of test scores to measure school quality, echoes statements from a number of assessment experts. The existing tests provide only limited useful information about whether students are really learning more, or that 'achievement gaps' are or are not narrowing - certainly not that students are able to use what they are learning in significant and meaningful ways. Too many unfounded conclusions or inferences (using testing lingo) are being made about test scores. Daniel Koretz (Professor of Education at Harvard University, and a nationally recognized expert of educational testing), summarizes both issues of testing and the uses to which test scores are put:

> "A test…covers a small sample [of knowledge] of the domain [the content area]….The accuracy of a test depends of a careful sampling of content and skills.…The accuracy of a test score depends on seemingly arcane details about the wording of items [questions], the wording of 'distractors' (wrong answers)…the difficulty of items….the attitudes of test takers…the behavior of others….If there are problems with any of these aspects of testing, the results from the small sampling [of the domain, or content area]…will provide misleading estimates of students' mastery of the larger domain…..A failure to grasp this principle is at the root of widespread understandings of test scores….[especially if] instruction is focused on the small sample actually tested rather than the broader set of skills…the test is supposed to signal." (10)

Tests have to be reliable and generally understood as producing roughly a similar result to those of many other test takers over time. Statisticians have many types of reliability measures, that is, tests also have to be valid, meaning that a given test actually must test what it says it tests. A math test that includes complicated verbal test questions might be testing reading skills as much, if not more than math; thus, information about actual math learning from such a test would be limited. The California tests are statistically reliable in a number of ways, and do test annually about a third of the grade level standards for that given year. Conclusions, however, about test results are often too far reaching. Too much weight is given to the precise score a student receives on these annual tests, which results in too many ill-founded conclusions.

A student receives a 'scaled score' in each subject tested each year. This scaled score shows how well a student did in comparison to other students taking that same test. Most well-known national tests report scores in this manner, for example, the college entrance Scholastic Achievement (formerly Aptitude) Test (SAT) of the College Board reports a scaled score somewhere in the range of 200 to 800 for each of its three (formerly two) test sections. The California Standards Tests use a scale of 150 to 600 for each content area test. A score of 500 on any of the SAT tests communicates that the student hit the overall mean (or 'average', as commonly used) score of all the test takers. For the California tests, the mean changes a little each year, for each grade level and each subject tested. The California State Board of Education has set a 'performance band' range of 300 to 350 on the scale as 'Basic', which is defined by the Board as meaning the student has not quite mastered the content standards for that year in the tested subject. The range of 350 to approximately 400 (the number varies a bit by grade level and test) is 'Proficient', meaning the student has mastered the material; and 400 and above is considered 'Advanced'. Schools are expected to have an increasing percentage of their students score at the 'Proficient' or 'Advanced' levels each school year, using complex formulas for state and federal accountability purposes. (Note that although California has two additional performance bands - 'Below Basic' and 'Far Below Basic' - this article is looking just at the higher three performance bands). (11)

Test makers and experts know that scores are not consistent over time. The same student taking the same test on different days and times will score differently, even on the most valid and reliable test. So scores on such tests are reported both as 'score points', the actual scaled score achieved, and also as a score range, the

statistically calculated range that the student would score within, were the student to take the test numerous times (12). This range is calculated to correct for various reasons that might cause a test taker to score differently, should they take the same test several times. In the case of the SAT, the range is a bit more than 50 scaled score points ranging both ways from the score point, or about eight percent of the scale range each way from the score point. The test taker is told that with 95% certainty, were s/he to take that SAT again and again, the score point would fall within that range.

If the California Standards Tests used a score range similar to that of the SAT, then this range would be about 35 scaled score points above and below the score point. A student with a score point of 375 and defined as 'Proficient' would have a range from about 340 to roughly 415. This student could be expected to score somewhere between the upper end of 'basic', anywhere in 'proficient', and the lower end of 'advanced' performance bands were the student to take the same test multiple times. That is a lot of variation, and drawing precise conclusions from the score point about the student's mastery of that year's content standards is not defensible. For example, the controversial measures of school and teacher quality now in use ignore the reality of range, and use only the score point to determine how many students are 'proficient', or have moved from one performance band to another. In an even more controversial use of score point data or in the case of those districts extrapolating teacher quality from test scores, they are looking to see how many students who were 'basic' last year are 'proficient' this year. Such precise interpretations of what scores mean are quite misleading and produce significant misunderstandings.

In August 2013, newspaper headlines announced 'School Test Scores Fall…Statewide' (Sacramento Bee, August 9), or 'Academic Performance Drops Statewide' (Los Angeles Times, August 29). These reports indicated a decline of .8 percent in the percentage of students proficient in English/Language Arts. The two-point drop in statewide Academic Performance Index was widely touted. API is calculated from the percentages of students in each performance band, as described above. In most years English score averages did decline, the largest of these being four scale score points. In grade seven, for example, the mean (or average, as most people would use the term) scale score went from 367 in 2012 to 363.5 in 2013. Because the cut line for proficiency is 350, a larger percent of seventh graders in 2013 scored just below 350 than was the case in 2012. But these mean scale differences reflect very little actual change in the number of correct answers on the seventh grade test. On the 75 question English test, the average 2013 seventh grader answered correctly about one-and-a third questions fewer than did the 2012 seventh grader. An average difference of just over one question answered correctly from one year to the next does not strike testing experts as being all that momentous, and certainly nothing like those blazing news headlines indicated. (13)

The accountability system leads to overstated conclusions about the precision of test score points, and uses scores as if they were absolute fixed measures of learning. Students, teachers, schools, districts, and the entire state are often stigmatized by these ill-founded conclusions, based on how many students' specific score points are above—or are not above—the proficient line. Even the very setting of the cut lines for performance bands raises questions. The cut lines are not derived from any scientifically-developed statistic; the cut lines are judgments made by committees (the process is explained in the yearly *Technical Report* referenced in note 8) about how many questions a student 'should' answer correctly to be judged 'basic', 'proficient', 'advanced', whether realistic or not. Daniel Koretz explains that

> "Many current testing programs are designed in part to determine whether students have reached…the 'proficient' standard mandated by NCLB….The complication is that 'proficient' is merely an arbitrary point on a continuum of performance; it does not indicate mastery of all of a discrete set of skills….An even larger issue is deciding where to put the cut score [of

> proficient]….What level of performance is required to be called…proficient…remains a matter of judgment…the judgment…is not…[based on] some scientifically validated criterion." (14)

There are few published studies that compare a 'proficient' score on any of the California tests to nationally normed tests, where 'grade level' is generally defined as the 'average' score (or the mean). Available comparisons indicate that 'proficient' is well above statistical definitions of 'grade level'. (15)

Yet another distortion is what experts call score inflation. This is a serious weakness of high stakes testing. When instruction focuses on those pieces of content expected on the test, mastery is undermined and scores are 'false positives' in that they give scores that do not relate to mastery of the content domain as a whole. The domain is not being learned - only that part expected to be tested. Thus, California-scaled score averages have risen over the past 12 years, and California-scaled score averages on the National Assessment of Educational Progress (NAEP) show very little change. Experts question whether increased learning has actually taken place. Robert L. Linn, head of the Center for Research on Evaluation, Standards and Student Testing of the American Educational Research Association, and one of the most widely acclaimed national assessment experts, explains:

> "The biggest question of all is whether the assessment-based accountability models that are now being used or being considered by states and districts have been shown to improve education. Unfortunately, it is difficult to get a clear-cut answer to this simple question. Certainly, there is evidence that performance on the measures used in accountability systems increases over time, but that can also be linked to the use of old norms, the repeated use of test forms year after year, the exclusion of students from participating in accountability testing programs, and the narrow focusing of instruction on the skills and question types used on the tests…. Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high stakes are attached to them. The unintended negative effects of the high-stakes accountability uses often outweigh the intended positive effects. It is worth arguing for more modest claims about uses that can validly be made of our best assessments and warning against the over-reliance on them that is so prevalent and popular." (16)

Experts note that data from multiple choice tests can provide useful information about student learning, but they only provide some information. Even the College Board says that its SAT tests should be used in a larger context with other sources of information to evaluate student learning. In the case of the CSTs, the yearly administration of the tests and the use of the score point and not score range have led to unwarranted conclusions.

Professor Emeritus James Popham of UCLA calls reliance on test scores to measure school quality 'misguided'. 'Score spread' on these tests is created more by socioeconomic differences than by learning differences.

> "…The preoccupation with raising test scores has become dominant throughout most parts of the country. ...The preoccupation was with test-score raising, not necessarily with teaching kids the things that children ought to be learning….The classroom becomes a drill factory, where relentless pressure, practice on test items, may raise test scores -- but may end up having children hate school....

> At the very beginning of the accountability movement, I don't believe the policy makers really understood what kinds of measures should be used to judge schools; the policy makers stipulated that student test scores would be the prime determiner of educational quality. They were nationally standardized tests. They were produced by reputable companies. So the belief was these will be the appropriate tests to use. The fact is, however, these are not the right kinds of tests to use to judge the quality of schooling….
>
> The common belief that schools that score high on a standardized achievement are effective and that schools that score low are ineffective is simply misguided. It reflects ignorance about the nature of the test being used, because… tests, frankly, in many years measure the kind of conduct, knowledge and skills that children bring to school -- not necessarily what they learn at school. What you want to judge the quality of schooling is the test that measures how well children were taught, not whether they come from a ritzy background.
>
> Traditionally constructed standardized achievements, the kinds that we've used in this country for a long while, are intended chiefly to discriminate among students ... to say that someone was in the 83rd percentile and someone is at 43rd percentile. And the reason you do that is so you can make judgments among these kids. But in order to do so, you have to make sure that the test has in fact a spread of scores. One of the ways to have that test create a spread of scores is to limit items in the test to socioeconomic variables, because socioeconomic status is a nicely spread out distribution, and that distribution does in fact spread kids' scores out on a test. An example I often use is a question that involved a child's familiarity with fresh celery. There are actually questions on one of the currently used standardized achievement tests where you have to know what fresh celery looks like. But kids from upper-class homes, middle-class homes, where they buy fresh celery all the time, have a much better shot at that question than do kids from families where they're getting by on food stamps.
>
> Now, there are many such questions in a test. You wouldn't think there would be. Why would they have them? But those tests spread out examinee performances very well….[Another is} one that's currently used right now, where the emphasis was on the youngster's being able to tell what the word "field" meant. "In which field do you plan to work after you graduate?" Well, children from families where a mother or father has a professional field, like a lawyer or a dentist or a physician, they're going to be more familiar with the world "field" in that connection than would be a child from a family where a mom is a grocery store clerk or a dad who works in a car wash. So the kids from the middle- and upper-class families, where they have fields of occupation, will clearly have a better shot at that item than will kids from disadvantaged families." (17)

Finally, it should be no surprise that the State Superintendent calls the CSTs 'outdated' as they measure mostly recall types of information, and not the types of thinking needed in an increasingly 'knowledge-based' economy. Assessment experts show that 'outdated' is too mild a term to use in criticizing the California assessment/accountability system. These very people who know more than anyone how tests are made, how they are used, and how they should be used, agree that the tests are being used incorrectly and are being given too much credence.

THE NEW ASSESSMENT VISION

The new SBAC assessments planned for 2015 will be in ELA and math, and limited to only grades 3-8 and 11 which meet the ELA and math requirements of the federal Elementary and Secondary Education Act (ESEA) - the latest reauthorization of what is called 'No Child Left Behind.' California will also test science at three grade levels, required by the same federal law, and will soon revise the science tests to fit with the new 'next generation' science standards. The SBAC tests will include portions that look at much more than 'recall of information', and will have students respond to writing tasks, solve problems, and do at least one more elaborate response in ELA and one in math. But how much SBAC tests will differ from the cancelled CSTs depends on total time schools must allot for testing, how well the technology will work, and the financial support for scoring non-traditional tests. SBAC assessments are a work still in progress.

In addition to SBAC, the California State Superintendent proposes 'un-narrowing' the curriculum, knowing fully well that what is tested (and counted in accountability) is what is taught. Among the twelve recommendations of his report was the inclusion of non-ESEA content areas in a future assessment system. His Report calls for additional assessments in ELA, math, and science in those grades not tested by SBAC and developing a wide ranging assessment program that tests the rest of the core curriculum. The Report includes:

> **"Recommendation 7 – Assess the Full Curriculum Using Assessments that Model High-Quality Teaching and Learning Activities:** Over the next several years, consult with stakeholders and subject matter experts to develop a plan for assessing grade levels and curricular areas beyond those required by the ESEA (i.e., ELA, mathematics, and science) in a manner that models high-quality teaching and learning activities. Areas for consideration should include the visual and performing arts, world languages, technology, science, and history–social sciences…." (18)

This broadening of assessment will go way beyond the ELA and math assessments that are being created by SBAC. By 2016, the state is to have a plan to design additional assessments that will include even more focus on student analysis based writing, student projects, student produced work including student application and analysis of knowledge. AB 484 says that the state is to create assessments that go beyond MAPP.

> "Exclusive of those assessments established by a multistate consortium, produce performance standards to be adopted by the state board designed to lead to specific grade level benchmarks of academic achievement for each subject area tested within each grade level based on the knowledge, skills, and processes that pupils will need in order to succeed in the information-based, global economy of the 21st century….The system includes assessments or assessment tools for multiple grade levels that cover the full breadth and depth of the curriculum and promote the teaching of the full curriculum." (19)

Some of these assessments will be created at the local level, and some at the state level. In all cases, '21st century' skills will be the basis for assessment. School quality will be measured by student performance of these skills as well as by some traditional recall-oriented test items. History/Social Science as a content area will be included, as stated by the law, and History/Social Science (the term used in CA AB 484) analysis and reasoning skills will be practiced in performance assessments.

The vision for these assessments was presented to the State Board of Education in July 2013 in a report from

the Educational Testing Service (ETS), as stated in the CDE's contract with ETS and described as part of the Superintendent's Report. ETS outlined the steps necessary to enact this entire revolutionary change over the next several years. Significantly, Educational Testing Service is one of the giants of the testing industry. The company has been around for more than half a century, and manages a variety of testing programs that have impacted many, if not most, Americans in some way. It is the company that brings us the Scholastic Achievement Test, Graduate Record Exams, the Advanced Placement exams, and is the prime contractor for the now mostly defunct California Standards Tests. Even more significantly, ETS explained that assessment needs to move away from relying just on multiple choice tests! This would be something akin to the Association of Horse and Buggy Makers of the United States, if there had been such a thing, calling for the creation and expansion of the automobile industry in 1903.

> "The use of performance tasks in large-scale assessments introduces the potential to enhance the assessment experience for students, expand the wealth of information on student understanding that could be accessed by educators and other interested parties, and influence in positive ways the direction of instruction and learning in the classroom. Performance tasks can take on a variety of forms….Standards documents such as the Common Core State Standards (for English language arts and mathematics), the Next Generation Science Standards, and the National Curriculum Standards for Social Studies [these documents are listed in the Reference section. This ETS memo was written 6 months prior to the publicaton of the national 'C3' social studies framework, mentioned below.] all clearly communicate the importance of well-developed reasoning, analytical, and research skills, in addition to strong discipline-based content knowledge and competence. And, more generally, the Partnership for 21st Century Skills promotes "fusing the 3Rs and 4Cs (Critical thinking and problem solving, Communication, Collaboration, and Creativity and innovation)" (http://p21.org/). These standards documents along with others suggest a potentially significant role for performance tasks in the larger assessment picture. Additionally, when thoughtfully designed into an assessment, the combination of short and extended performance tasks with discrete items and smaller item sets can support the efficient assessment of a wide range of content along with the more targeted assessment of particular aspects of disciplinary habits of mind…." (20)

ETS says large-scale performance testing is viable and recommends moving to it!

The ETS memo continues and concludes:

> "Assessment of student skills and knowledge and of their use of content in analysis and application will drive curriculum and instruction to do the same. The limits of multiple choice types of tests and even more, the uses to which test scores are used, drives the move to new, broader assessments that examine student skills. The value of a student performance assessment connected to a new accountability system that looks at school quality in a different way will lead to student performance-based curriculum and instruction. This refocus on learning will also tend to level the playing field among students as a whole.
>
> [Tests could in] 30 or more minutes, ask the student to…analyze particular aspects of several literary works or historical pieces…. More extended performance tasks, however, offer greater opportunities to assess students' capabilities to think deeply and may reveal new insights into their critical and creative thought processes. Consider, for example, a performance task that spans

> a period of several days or even weeks in which a student is required to provide interim products at specific milestones and a final product. A possible valuable by-product of such a task is that it creates a path of observable behaviors from which data may be collected for later analysis.
>
> Additionally, certain kinds of extended performance tasks might introduce opportunities for small groups of students to collaborate over a period of days or weeks toward a common goal, such as the submission of a product prototype that they have developed to satisfy a particular set of design requirements. Part of such an exercise might involve not presenting the student or group of students with all of the information and resources at the outset that they will need to achieve their end goal, but instead having them decide what is needed initially to carry out their task and then deciding how to utilize those materials and resources most efficiently…. Additionally, when thoughtfully designed into an assessment, the combination of short and extended performance tasks with discrete items and smaller item sets can support the efficient assessment of a wide range of content along with the more targeted assessment of particular aspects of disciplinary habits of mind....***One additional benefit related to the inclusion of performance tasks on large-scale assessments is the impact on classroom learning and instruction. If there is even a grain of truth to the statement that "what gets assessed is what gets taught," then the need for presenting students with opportunities to demonstrate their academic competence in more real-world settings (and that demand the integration of knowledge, skills, and thought processes consistent with those required in university-level studies and in their careers) would seem to support the inclusion of a range of well-designed performance tasks in large-scale assessment."*** [emphasis added] (21)

Both the SSPI report and the ETS memo, taken together, propose a different future direction for assessment inclusive of all core subjects, as well as different kinds of accountability. The SSPI Report talks of using performance assessment, possibly in part at local levels, as a piece of any new accountability system. The ETS memo gives examples of assessment activities that include individual and group projects, possibly spread out over time, including such things in the History/Social Science or social studies area such as History Day projects, various civic learning activities, mock trials/moot court activities, geography 'story maps' and use of mapping technology to analyze everything from immigration patterns to major historical events, or activities in which students practice cost/benefit analysis or examine the importance of human capital, that could well be part and parcel of regular instruction. (22) The SSPI report - and more specifically the ETS report - speak to the need to have assessment guide or model the kinds of engaging instructional practices that have students write, create presentations, engage with each other in debates and simulations, and other activities.

The assessment and accountability system following it, per ETS, would consciously model the kind of learning that is based on best instructional practices, as described in the Common Core State Standards and other new standards documents. Importantly, because a considerable (and a growing body) of research showing that reading comprehension is significantly connected to subject content, students need content knowledge to comprehend what they are reading, and they need content knowledge to write analytically. History/Social Science instruction that includes analysis and argumentative writing will help both history and general ELA comprehension in any assessment system.

Assessment of student skills and knowledge and of their use of content in analysis and application will drive curriculum and instruction to do the same, as the Consortium tests ideally will do in English and math. The

limits of multiple choice types of tests and even more, the uses to which test scores are used, drives the move to new, broader assessments that examine student skills. The value of a student performance assessment connected to a new accountability system that looks at school quality in a different way will lead to student performance-based curriculum and instruction. This refocus on learning will also tend to level the playing field among students as a whole.

HISTORY/SOCIAL SCIENCE AND A NEW ASSESSMENT SYSTEM

The California Department of Education's plan is to create this new assessment system over the next several years. Some will be local in nature, and will be connected to the new school funding system just enacted by the California legislature and signed by the Governor. Some of the new system will be 'matrix' or 'sample' testing; not all students at a given school would take all the subject tests each year, and some students would not get the same test questions or projects in a given subject. Not knowing in advance which students would be tested on what parts of the curriculum, schools would need to teach all the curriculum to all the students; the curriculum would be 'un-narrowed' and all subjects would be included at the school site. History-Social Science or social studies would not only return, but would be better than before. The new assessment system would inspire it.

The 'College, Career, and Civic Life' ('C3') social studies document (23), under development for three years as a project of the Council of Chief State School Officers and published by the National Council for the Social Studies (NCSS) in September 2013, provides a framework for the analysis and evaluation skills that use social studies content in the disciplines of history, geography, economics, and civics. These disciplines can shape the design of assessment of student learning and practices of various social studies skills, and fit well with the Common Core ELA standards; the History/Social Science or social studies educators in California need to participate in the design of the new assessment system, at both their local level, as the local accountability design takes shape, and at the state level, by volunteering for committees, by encouraging continual support from their local state legislators, and by supporting the work of social studies organizations such as the California Council for the Social Studies, which will join with other social studies organizations in supporting and advocating for the new assessment system. History/Social Science educators will need to be the ones who define and determine the shape and content of these assessments.

Education in California is on the edge of a whole new way of assessing student learning that will both improve instruction and change school accountability. Schools will be accountable for learning what really counts, not just something that can easily be counted. And most importantly, the public schools will become places that nurture students, and actually motivate a life-long curiosity and commitment to learning.

References

Bracey, Gerald W. *Reading Educational Research: How To Avoid Getting Statistically Snookered*. Portsmouth, New Hampshire: Heinemann, 2006.

California Department of Education. *Common Core State Standards** (Sacramento, CA, CDE. Modified March 2013 edition), Available at www.cde.ca.gov/re/cc/index.asp.
*[*Note from author: These are slightly modified from the national Common Core State Standards published by the Organization of Chief State School Officers. See also www.cde.ca.gov/pd/ca/sc/ngssstandards.asp for the Next Generation Science Standards. All these standards have been adopted by the State Board of Education.]*

National Council for the Social Studies. *National Curriculum Standards for Social Studies: A Framework for Teaching, Learning, and Assessment.* September 2010. Available at www.socialstudies.org/standards.

Popham, W. James. *The Truth About Testing: An Educator's Call to Action.* Alexandria, Virginia: Association for Supervision and Curriculum Development, 2001.

Ryan, Katherine and Lorrie A Shepard, editors. *The Future of Test-Based Educational Accountability*, New York: Routledge, 2008.

Notes

1) California Legislature, Assembly Bill 484, Section 15:5:F, Chaptered 2 October 2103. Available at www.leginfo.com.

2) United States Secretary of Education Arne Duncan, "The Well Rounded Curriculum; Remarks at the Arts Education Partnership National Forum", 9 April 2010. Available at www2.ed.gov/news/speeches/2010/04/04092010.html.

3) Arne Duncan, "Beyond the Bubble Tests: The Next Generation of Assessments…Remarks to State Leaders", 2 September 2010. Available at www.ed.gov/news/speeches/beyond-bubble-tests-next-generation.

4) The sanctions applied to low scoring schools and districts have required an increase in time spent on reading and/or math remediation. Usually the number of minutes per day of remediation from specific programs is part of the intervention. Often the lower scoring students are mandated to have additional remediation minutes. The result in elementary schools is deletion of those content areas not tested at elementary school; ultimately everything except English language arts and math. If the school scores do not increase adequately, the number of minutes required for remediation increases yet again. In middle schools, non-tested subject area courses are eliminated, or reserved for the higher scoring students. Sometimes, non-tested subject areas, or those that are not included in federal legislation for Annual Yearly Progress, are made into electives, and even made into 'after school' voluntary programs. Thus the curriculum narrows, and for children from lower socioeconomic backgrounds it tends to narrow even more. Some scholars argue that denying the full curriculum in essence penalizes the latter children even more than those not disadvantaged. See especially E.D Hirsch, Jr., *The Knowledge Deficit*, (Boston: Houghton Mifflin, 2006; Hirsch explains how the acquisition of vocabulary is best made when the learner is provided the context of the vocabulary, meaning the full curriculum of humanities, sciences, arts, and history/social science.

5) California Department of Education, "*California Department of Education Report to the Governor and the State Legislature: Recommendations for Transitioning California to a Future Assessment System"* (January 8 2013). Available at http://www.cde.ca.gov/ta/tg/sa/ab250.asp., iv.

6) *Ibid*., 22-23.

7) *Ibid*., 24.

8) *Ibid*., 36.

9) *Ibid*., 37-38.

10) Daniel Koretz. *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, Mass. Harvard University Press, 2008), 19-22.

11) Full descriptions of California Standards Tests are posted each year on the California Department of Education website. These include an annual *Technical Report* (650 pages) available at www.cde.ca.gov/ta/tg/sr/documents/cst12techrpt.pdf for the 2012 report, the most recent complete report as of September 2013. At www.cde.ca.gov/ta/ac/ap/ are located the yearly *Information Guide* (which for 2013 runs to 84 pages). An *Overview of Accountability* is 5 pages. The *Parent Guide* is 2 pages.

12) This range is called a 'confidence interval' by statisticians, and is a specific calculation giving a margin of error for tests, opinion polls, and surveys, among other things. How measurement error and margin of error are calculated is a chapter in most

regular overview textbooks on statistics (for example: David S Moore, *The Basic Practice of Statistics* (New York: WH Freeman, 1995), 324-349). For the purposes of this essay the important idea is that scores will vary within predictable ranges when the same student takes the same test several times.

13) The 7th grade mean scale score declined from 367 to 363.5. According to the Technical Report for 2012 cited above, the 2012 7th graders got on average a bit more than 53 questions right on the 75 question English test, and the 2013 7th graders got on average 52 and a fraction right. This caused .8% fewer students statewide to be in the 'Proficient' category in English in 2013 than was the case in 2012. *Technical Report for 2012, 559.)*

Teresa Watanabe, "Academic Performance Drops Statewide, But LA Unified Improves," *Los Angeles Times*, (August 29, 2013):1.
Loretta Kalb, Phillip Reese and Brittany Torrez, *"School Test Scores Fall in Sacramento Area, Statewide"*, (August 9, 2013):1b.

14) Koretz, 184.

15) "Interestingly, little test comparison, the way new tests are often validated, has been published for the CSTs. The public and educational community has been told that a scaled score of 350 is a proficient score, with the unspoken implication that this is grade level. It appears that a proficient scaled score of 350 is in fact much higher than normal grade average, and much more demanding than is commonly understood.
In 2001, the 9th grade English Language Arts scores of the national norm referenced test (SAT-9) still being given as part of the California state system, were correlated with the English Language Arts and reading scores of the new state standards test for 9th graders at a large suburban high school at which this writer was working. 740 9th grade students completed both tests in the same week of the school year. These students had the same degree of motivation (or lack of it) for the two tests, as both were used as part of the state measurement. Of the 740 students, 609 scored at or above the 50th%ile on the normed SAT-9 test. On norm referenced tests, generally the 50th%ile is understood to be 'grade level'. On this measure, then, these students were at or above grade level on the normed test. Of the 740 students, 268 scored as proficient or above on the CST. Obviously it was much more difficult to be proficient on the CST than to be above the 50th%ile on SAT-9. How much more difficult becomes clear when comparing achievement on the SAT-9 to the CST. The median percentile of those CST proficient students on the SAT-9 was 81%ile! This means that the average of proficient students was almost one full standard deviation above the mean on the SAT-9, a very significant difference. None of the CST proficient students scored below the 65th%ile on SAT-9. So, to be proficient on the CST was to be scoring higher than two thirds of all students nationally on the SAT-9! The standard deviation of the CST proficient students on the SAT-9 was 8%ile points; two thirds of the proficient students scored between the 74th and 89th%iles on the SAT 9. Demanding that students be proficient in California demands that they be above the 65th%ile nationally, or in the top one third of all students nationally, with most of them scoring in the highest fourth nationally.
This test comparison was one time, one year. The difference between SAT-9 grade level and CST proficiency may or may not be the same for other grades or for other years. Comparisons need to be repeated, and in a number of grade levels, as the CST change half their items each year. But for this comparison, proficiency was way, way above accepted definitions of grade level. This issue must be examined and re-examined and CST proficiency re defined."
Jim Hill, "'Value-Added' History-Social Science Effectiveness," *Social Studies Review* 50, (2011): 24.

16) Robert L. Linn, "Assessment and Accountability", *Educational Researcher*, vol 29, no 2, (2000): 4-16. Available at http://pareonline.net/getvn.asp?v=7&n=11.

17) James Popham, "Interview: James Popham," *Frontline* Public Broadcasting System, April 25, 2001. Available at www.pbs.org/wgbh/pages/frontline/shows/schools/interviews/popham.html.

18) Report, 43-44.

19) Assembly Bill 484, Section 4:a:1.

20) *A Long-Term Assessment Plan for the California Assessment System*, Education Testing Service, June 2013. Available at www.cde.ca.gov/be/pn/im/documents/memo-dsib-adad-jun13item01.doc. 77.

21) *Ibid*., 78-79.

22) In addition to the California History/Social Science Subject Matter Project, Constitutional Rights Foundation, Center for Civic Education, Stanford History Education Group, and other websites such as SCORE and CLRN, the California Geographic Alliance

and the California Council on Economic Education post performance-based lesson activities. The California Council for Social Studies publications regularly feature articles about performance-based lessons in Geography, Economics, Civics, as well as History. See in particular the *Social Studies Review* issue for Spring/Summer 2003 (Vol. 42, No. 2), *A Passion for Geography*, and *Annual Issue 2103*, (Vol. 52), *Economic Education in the 21st Century: How Can Students Develop Economic Reasoning.*

23) The *College, Career, & Civic Live 'C3' Framework for Social Studies State Standards* developed by a collaborative of national social studies organizations, for a time under the auspices of the Organization of Chief State School Officers, published in September 2013, Available at http://www.ncss.org/system/files/c3/C3-Framework-for-Social-Studies.pdf

CCSS peer-reviewed **Occasional Papers** editors are Maggie Beddow, Ph.D., CSU Sacramento [beddow@csus.edu] and Emily Schell, Ph.D., San Diego State University [eschell@mail.sdsu.edu]. If you are interested in submitting an article on a research-based topic of importance to K-12 social studies curriculum leaders, pre-service teacher candidates, university educators, or classroom teachers, please contact one of the editors for submission specifications.