An open letter from the Center on Education Policy to the SMARTER Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers

May 3, 2011

To the member states of SBAC and PARCC:

The Center on Education Policy supports the efforts of the SMARTER Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers to develop high-quality common assessment systems aligned to the common core state standards. This is a positive and historic step that will go far in alleviating some of the problems associated with having many different state assessments and standards.

I would like to share some lessons we at CEP have learned after several years of studying student achievement trends and No Child Left Behind implementation. We appreciate that the technical aspects of your assessment designs are being handled by people with considerable expertise, so our intent is not to weigh in on the details of that process. Rather, we want to highlight a few issues for you to consider that address the most pervasive problems we encountered in our research and that have important implications for those who will use your assessments.

To make the new assessment systems as useful as possible to policymakers, educators, researchers, parents, and the public, we urge both consortia to consider the following suggestions:

- Routinely report mean (average) scores on your assessments for students overall and for each student subgroup at the state and local levels, as well as across the consortium. This should be done in addition to reporting the percentages of students reaching various achievement levels.

- Report achievement gaps between subgroups in terms of mean scores instead of in terms of percentages of students reaching various achievement levels. This avoids the problem of gaps appearing smaller or larger depending on where the cut score for proficient (or basic or advanced) performance is located on the scoring scale for a test.

- Establish linkages between the SBAC and PARCC assessment systems so results on select indicators can be compared across all states.

The rest of this letter explains our reasons for making these suggestions and provides supporting data from CEP's research.

Mean Test Scores

From our efforts to collect data for our achievement studies, we have learned that many states do not report mean scores on their tests, even though they are a simple and obvious way to summarize group performance. This omission is largely due to NCLB, which requires states to report percentages of students reaching three achievement levels but not mean scores.

Although percentages proficient are the "coin of the realm" in NCLB, they do not tell the whole story of what's happening with student achievement. The percentage proficient places an implied value on bringing students to a minimal level of proficiency but does not capture achievement gains above the proficiency cut score (or below it for students who have not yet reached proficiency). Mean scores, by contrast, capture performance across the achievement spectrum and place an implied value on improving performance at all levels, from the lowest- to the highest-achieving students. Mean scores offer a different way of looking at student performance that fills in the information missing from percentages proficient.

Therefore, we suggest that each consortium *routinely report mean scores on the common assessments for students overall and for each subgroup, at the state and local levels as well as across the consortium.* This should be done in addition to reporting the percentages of students reaching various achievement levels.

Mean scores have a particular advantage over percentages proficient in measuring and reporting on achievement gaps between subgroups. For that reason, we have a separate suggestion on this issue, explained below.

Reporting on Achievement Gaps

In our five years of studying state test score trends, we have continually come up against a problem in reporting on gaps between subgroups in percentages proficient—namely, that the apparent size of these gaps varies depending on where a state has set its cut scores for proficient performance on its tests. If a cut score has been set very high or very low on the scoring scale, so that almost everyone reaches it or almost nobody reaches it, the gaps between subgroups will appear to be small. But if the cut score is set closer to the average test score—and thus closer to where the bulk of students' scores are distributed along the scoring scale—then achievement gaps will appear to be larger. Paul Holland described this phenomenon in a 2002 study in the *Journal of Educational and Behavioral Statistics*, which used data from the 2000 National Assessment of Educational Progress.

Our own review of data from the 2000 NAEP also illustrates this point. That year, the average NAEP score in grade 8 math was 275 on a scale of 0-500. With a cut score of 262, somewhat below the average score, 77% of white students and 32% of African American students reached or exceeded this score, resulting in a gap of 45 percentage points. With a cut score of 299, somewhat above the average score, 35% of white students and 6% of African American students reached or exceeded the score, yielding a gap of 29 percentage points. But with a cut score of 333, far above the average score, 7% of white students and a rounded rate of 0% of African American students reached or exceeded this score, indicating a gap of only 7 percentage points. Thus, people can get a very different picture of the size of achievement gaps depending on where the cut score is set. In this example, the two cut scores that were somewhat below and somewhat above the average score revealed a larger gap than the cut score that was so high only a small percentage of students in either group reached it.

Robert Linn of the University of Colorado, who is a member of a panel of expert advisors to CEP, encouraged us to explore this phenomenon using state test results from our comprehensive dataset from all 50 states. Specifically, we looked at the size of African American-white and Latino-white achievement gaps on state tests in two groups of states: 1) those in which more than 90% of white students reached the proficient level; and 2) those in which less than 80% of white students reached the proficient level. We did a similar analysis of gaps between low-income and non-low-income students, grouping states with proficiency rates for non-low-income students that were 1) over 90%; and 2) under 80%. This analysis assumed that states in the first group were likely to have easier tests and/or cut scores than states in the second group. At the same time, we recognize that some of the differences between states in percentages proficient reflect real differences in student achievement.

The table on the next page shows the results of our analysis based on state test data for school year 2008-09. In general, we found that gaps for African American, Latino, and low-income students were often more than twice as large in the second group of states (those likely to have higher cut scores) than in the first group of states (those likely to have easier cut scores). In other words, when the proficiency cut score is easy enough that the vast majority of students in both the lower-performing and the higher-performing group reach it, there is not much of a gap between the two groups. When the proficiency cut score is set higher—but not so high that it's out of range for all but the most advanced students—many more students will have scores clustered right above or below the cut score. But more students in the higher-performing group will achieve above the cut score, while more in the lower-performing group will achieve below it, producing a larger gap.

Consider the African American-white gap in high school reading. As the table shows, this gap averaged 11 percentage points in the first group of states, those with white percentages proficient above 90%. Nebraska, for example, reported a proficiency rate of 94% for the white subgroup and 83% for the African American subgroup—a gap of 11 percentage points, the same as the average gap for the first group of states. Tennessee, another state in this group, reported a proficiency rate of 98% for white students and 95% of African American students—a gap of only 3 percentage points. But in the second group of states, those with white

percentages proficient below 80%, the African American-white gap in high school reading averaged 29 percentage points. In Oregon, for example, 73% of white students scored proficient in high school reading, compared with 43% of African American students—a gap of 30 percentage points.

Relationship between location of cut score and size of gap in percentages proficient, 2009

Subgroup	Average size of gap in terms of percentage points					
	Grade 4 reading	Grade 8 reading	High school reading	Grade 4 math	Grade 8 math	High school math
African American & white gap						
Size of gap in states with white percentage proficient above 90%	14	12	11	16	24	13
Size of gap in states with white percentage proficient below 80%	25	26	29	25	30	33
Correlation between white percentage proficient and size of gap	-0.45	-0.53	-0.52	-0.47	-0.33	-0.31
Latino & white gap						
Size of gap in states with white percentage proficient above 90%	11	11	10	11	14	7
Size of gap in states with white percentage proficient below 80%	23	23	24	20	22	23
Correlation between white percentage proficient and size of gap	-0.47	-0.44	-0.48	-0.54	-0.37	-0.42
Low Income & non-low income gap						
Size of gap in states with non-low income percentage proficient above 90%	14	13	10	13	16	11
Size of gap in states with non-low income percentage proficient below 80%	25	26	24	24	25	23
Correlation between non-low income percentage proficient and size of gap	-0.62	-0.58	-0.46	-0.66	-0.28	-0.12

The table also shows the results of our tests of correlation. (As you know, tests of correlation indicate whether two measures—in this case, the white percentage proficient and the size of the gap—tend to move together in the same direction or different directions. A positive correlation, represented by a number from above 0 to 1, indicates that one measure increases as the other increases. A negative correlation, represented by a number from below 0 to -1, indicates that one measure increases as the other decreases.) We found moderate to strong negative correlations between the white percentage proficient and the size of the achievement gap. That is, the higher the white percentage proficient, the smaller the apparent gap.

The problem is that most people would simply infer that certain states have larger achievement gaps than others without realizing the location of the cut score was a factor. But as the above analysis illustrates, there is no one “true” gap between two different groups—rather, different cut scores yield different snapshots of the gap.

After years of using various ways to report on gaps in our achievement studies, we’ve come to see that using mean scores for gap reporting is preferable to using percentages proficient because mean scores avoid the problem described above. We’ve also found that across the

board—for all subgroups, in both reading and math, and at grades 4, 8, and high school—mean scores give a less rosy and more accurate picture of progress in narrowing gaps than percentages proficient do. For example, in our 2010 report, *State Test Score Trends through 2008-09, Part 2: Slow and Uneven Progress in Narrowing Gaps*, we aggregated trend lines between 2002 and 2009 across all states with sufficient data in both subjects and at three grade levels. We found that the gap between low-income and non-low-income students narrowed in 72% of the trend lines analyzed using percentages proficient but just 57% of trend lines using mean scores. Similarly, the African American-white gap narrowed in 78% of trend lines using percentages proficient but just 61% using means. A particularly telling example occurs in grade 4 reading, where the gap for low-income students narrowed in 74% of the trend lines analyzed using percentages proficient but less than half of the trend lines (44%) using mean scores.

An example helps to illustrate why this occurs. Between 2002 and 2009, large numbers of African American students improved their state test performance enough to move from just below the proficiency cut score to just above it or higher. During the same period, many white students were already scoring above proficient, so improvements in their scores did not affect the white percentage proficient. Thus the African American-white gap in percentages proficient narrowed while the mean score gap stayed the same or sometimes even widened.

Undoubtedly both consortia will need to set cut scores to determine whether students have reached various achievement levels, and we recognize this is a complex process. (We assume the states in each consortium will use common cut scores; if this has not been decided, it is very important that they do so.) But we do not think the measurement of gaps should be so dependent on where these cut scores are set. We urge you to *report achievement gaps between subgroups in terms of mean scores instead of in terms of percentages of students reaching various achievement levels.* Separating decisions about cut scores from decisions about gap reporting will allow you to set cut scores for the various achievement levels as defensibly as possible, without being concerned about how this will affect people's perceptions of achievement gaps.

<u>Comparability between SBAC and PARCC Assessments</u>

Assuming your work results in two common assessment systems, and assuming all states in each consortium agree to use the same cut scores, then all results *within* each consortium will be comparable. So, parents in Missouri could compare their child's performance on the SBAC test to averages for the school, district, and state. They could also interpret their child's performance relative to any other school, district, or state that is part of SBAC.

But many parents want to reference their children's achievement nationally, not just to students who live in states that happen to belong to their consortium. Comparisons across states would also be of great interest to the research and policy community, as well as to the states themselves. In the 2010 report, *Designing Common Standards and Assessments*, the National Governors Association and Council of Chief State School Officers stated their intention "to lead a joint effort to enable test scores to be compared across the summative assessments

being created by the two consortia." They consider this "a top priority goal" and will engage testing experts to help participants understand the many possible methods for promoting comparability across assessments. For instance, one approach is to embed a set of common questions across the two different tests. At the same time, the NGA and CCSSO acknowledge that the capability to make some types of comparisons is lost with multiple consortia.

We do not want to minimize the difficulties of establishing a linkage between the two consortia's assessments, which are envisioned quite differently. In its 1999 report, *Uncommon Measures: Equivalence and Linkage among Educational Tests*, the National Research Council elaborated on the numerous factors that must be considered to create such a linkage, such as similarity of content, item formats, and administration conditions. But only with a certain level of comparability will we fulfill a promise of common standards—the ability to tell parents and others with a stake in the educational system how students in their community are performing compared with those in other states and the nation as a whole.

Toward that end, we suggest that both consortia *establish linkages between the SBAC and PARCC assessment systems so results on select indicators can be compared across all states.* For instance, a link should be established so that the proficient score on the SBAC is comparable in difficulty to the proficient score on the PARCC; this would enable one to compare the percentages proficient across the two assessment systems. For these types of linkages to be created successfully, developers of both assessments will need to communicate with each other early in the design process rather than after the fact.

The work of the consortia offers exciting possibilities for a more coherent, effective, and cost efficient approach to assessment for the nation. We would be happy to provide more information about our work that can help with this endeavor. Please contact us if you have any questions.

Sincerely,

Jack Jennings
President and CEO

cc: Technical advisory committees, SBAC and PARCC
Laura Slover
Jeff Nelhaus
Doug Sovde
Joe Willhoft
Sue Gendron
Linda Darling-Hammod