Different But the Same:

Assessment "comparability" in the era of the Common Core State Standards

A White Paper developed for
The Council of Chief State School Officers

By
Brian Gong and Charlie DePascale
Center for Assessment

June 2013

Pre-publication version
A final version of this paper will be published by CCSSO

Acknowledgements

The Council of Chief State School Officers (CCSSO) is highly committed to supporting states implement college- and career-readiness systems of standards, assessment, and accountability. CCSSO is especially committed to implementation of the Common Core State Standards, which have been adopted by over 45 states, and whose development was co-sponsored by CCSSO in conjunction with the National Governor's Association. Part of CCSSO's strategy in supporting states' helping more students become more college- and career-ready is to support policy makers' and others' interpretation and use of assessment results. CCSSO is engaged in several ways to strengthen development of assessments for these purposes. CCSSO has commissioned this paper to help policy makers understand how test scores and interpretations may be compared for the assessments results available in 2015 and beyond. In addition, it serves to establish some shared understanding among those who may wish to engage in activities to try to link assessments or to interpret the empirical relationship of assessment results.

Score comparability and score comparisons are topics that have been widely discussed in the assessment community. This paper draws heavily on work done by many others, although the authors have not been able to acknowledge those connections throughout the paper. Any shortcomings are the responsibilities of the authors.

Table of Contents

Figures and Tables

Different But the Same:
Assessment "comparability" of in the era of the Common Core State Standards

Overview

To what extent can and should assessment results be "comparable"?

How and how much assessment results may be interpreted in relation to other assessment results are vitally important questions in the era of Common Core State Standards (CCSS) and assessments converging on the CCSS and other indicators of "college- and career-readiness." The answers to this question hinge on the intended purpose, interpretations, and uses of the assessment results as much as on the nature of the assessments themselves and the technical methods used to establish comparisons between assessment results. This document is intended to help policy makers anticipating adoption of assessments linked to the CCSS understand what is likely possible and not possible regarding comparison between assessment results, and to plan accordingly.

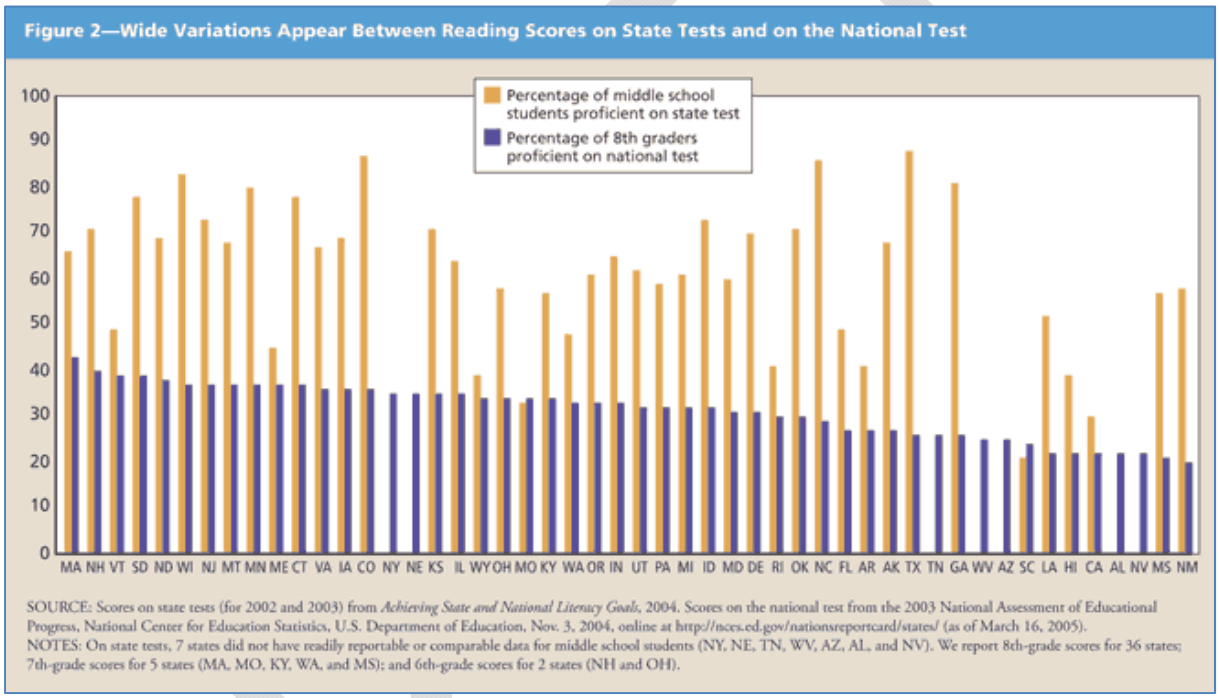
This document provides

- an overview of the reasons why score comparisons are especially important to consider now and the questions policy makers and others might consider
- a framework and examples of the types of "comparability" assessment results may have as tests and uses/interpretations interact
- a discussion with examples of the technical characteristics of assessments that influence comparability, including properties of the assessment scores, associated interpretations, test administration procedures, population, and opportunity to learn
- a summary of three levels of psychometric methods associated with producing assessment results that differ in degrees of "score interchangeability"
- a discussion of creating relationships and comparability interpretations among state assessments and other assessments that have significant differences but similar stated uses as indicators of college readiness
- an application to considering possible comparisons of PARCC and Smarter Balanced – two prominent CCSS-relevant assessments
- a summary of main points to help policy makers apply this information to making decisions about their assessment and accountability situations

Questions about assessment comparability in the era of the Common Core

It can be argued that the era of the Common Core State Standards (CCSS) began with questions about assessment comparability. If images can provoke action and define a movement then there are two related images that symbolize the issue of comparability and its impact on the development of the CCSS. Both images use the relationship between results on state assessments and NAEP to depict the wide variation in achievement standards across states. Figure 1 is an image which appeared in the 2005 RAND article by McCombs and Carroll, "Ultimate Test: Who is accountable for education if everybody fails?" The figure presents the difference between the percentages of students classified Proficient in reading on 2003 NAEP Reading test and 2002-2003 state assessments.

Figure 1: Percentage of students reported "Proficient" or above, State tests and NAEP, Reading, Grade 8, 2003-4

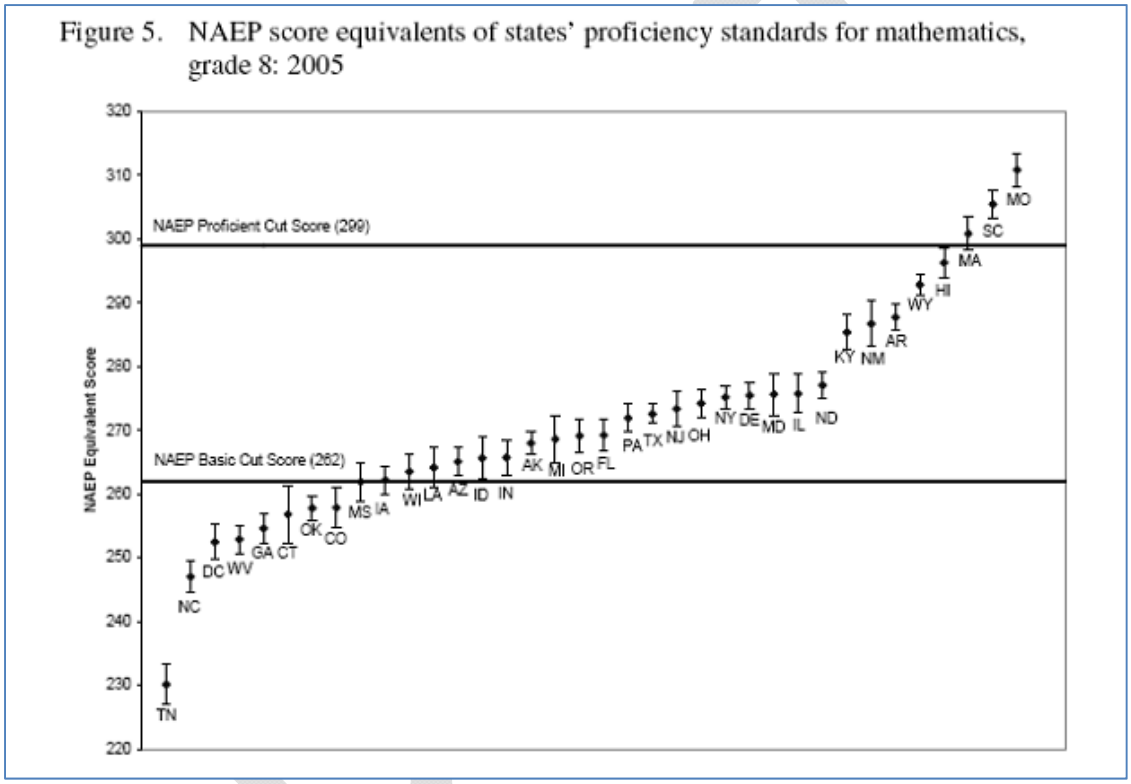


Observers noted of Figure 1that there was little agreement between the "percentage of students who scored proficient" reported by states on their own state assessments (gold bars), and the "percentage proficient" reported by NAEP (blue bars). In addition, there was much greater variation in the "percentage proficient" on the states' assessments than on NAEP; performance on NAEP was almost always lower than on the states' own assessments; and there was no consistent relationship between state and NAEP scores (e.g., it was not true that gap between NAEP and states' results was lower for states performing higher on NAEP).

Figure 2, which appeared in the 2007 NCES document, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales*, displays the results of an analysis to place each state's

proficiency scores onto the NAEP scale. The figures shows that three states' established proficiency cut scores for grade 8 mathematics were higher than the NAEP Proficient cut score of 299 on the NAEP scale; the other states ranged lower, with several states' proficient cut scores below NAEP's Basic cut score. The results displayed in Figure 2 along with those from several similar studies conducted between 2007 and 2009 made it clear that there was wide variation in the mean of proficiency across states and that the definition of proficiency by the vast majority of states was significantly different than the NAEP definition of proficiency.

Figure 2: NAEP scale score equivalents of states' proficiency standards for mathematics, grade 8, 2005



Both images have been replicated, widely circulated, and were widely cited as evidence of the need for common content standards, common achievement standards, and/or common assessments across states. The comparability arguments that fueled the development of the CCSS and subsequently the Race to the Top Assessment program are reflected well in U.S. Department of Education Secretary Arne Duncan's *Beyond Bubble Tests and Bake Sales* address to the 2010 National PTA Convention:

> For years, we have actually been lying to children and lying to ourselves by pretending that 50 different standards, in 50 different states, will make America

> competitive and help our children succeed in life. We have to stop pretending. We have to tell the truth. And we have to raise the bar for all children.
>
> Today, here in Tennessee—or in my home state of Illinois, and in too many other states around the nation-- a student that is deemed to be at grade level might be far behind grade level in Massachusetts.
>
> When you play basketball, the basket is always ten feet high. In football, the field is always 100 yards long. A 3-pointer is worth 3 points, and a touchdown is worth six points. Yet until now, we have resisted leveling the playing field in education. I'll put it plain and simple: When you tell students that they are ready for college and they are not, you are lying to children when they most need your candor and help. Thanks in part to the PTA's leadership and advocacy, we have 48 states working together on raising the bar, and seven states have already adopted these higher standards. This work, I'm convinced, is a game-changer.

Interpreting the images, arguments, and rhetoric from the years leading up to the release of the CCSS through the development of common core assessments we infer that to a large extent concerns about assessment comparability are, in fact, concerns about the comparability of achievement standards. To be more specific, questions about assessment comparability in the era of the CCSS are related primarily to the comparability of claims and inferences that can be drawn from assessment results. At the student level, when students in Louisiana, New York, Washington, and Maine are classified as *college-and-career ready* on the basis of their test performance, are we confident that the classification has the same meaning for each student? Similarly at the state level, when the percentage of *college-and-career ready* students in Louisiana, New York, Washington, and Maine is reported each year, how confident will we be that the problem depicted in Figures 1 and 2 has been solved? When multiple claims about the percentage of college-and-career ready students within a single state are made on the basis of different tests such as the state assessment, college admissions tests, and the national assessment, how are we to interpret consistencies and inconsistencies across tests?

In this paper, we will address critical factors that impact our ability to answer the comparability questions posed above. We will begin with discussing the extent to which comparability questions are impacted by whether we are comparing results across the same assessment, different assessments designed intentionally to produce the same results, and different assessments designed independently and for different purposes. Within that framework of comparability across tests, we will discuss several design, policy, and process decisions that impact the comparability of results across the same and different assessments. We will also provide an overview of methods for establishing links between assessments to support desired claims of comparability for assessments that are the same, are designed to be treated as the same, and assessments that are different. Finally, we will apply each of the topics described

above to a discussion of the common core assessments being developed by the Race to the Top assessment consortia: Smarter Balanced and the Partnership for the Assessment of Readiness for College and Career (PARCC). The discussion will address issues relevant to evaluating comparability now, in the initial year of administration, 204-2015, as well as presenting post-2015 scenarios and related comparability issues.

PARCC and Smarter Balanced: How comparable do they need to be? How comparable can we expect them to be?

Accepting the premise that the comparability of achievement level results across states is a primary comparability concern, attention naturally shifts to PARCC and Smarter Balanced and the development of the two multi-state assessment systems designed to measure performance against the CCSS.

Connection to PARCC/Smarter Balanced

PARCC and Smarter Balanced will probably not yield assessment scores that are completely interchangeable; in fact, there may be some significant differences.

Since PARCC and Smarter Balanced are still developing their tests and states will administer the new assessments for the first time in 2015, no empirical indication is available yet of how comparable the assessment results will be.

However, PARCC and Smarter Balanced have announced they are working together on several initiatives to support more comparable results, for example, working towards common definitions of permissible accommodations for students with disabilities and English language learners.

It is also true that some design decisions and operational considerations indicate that the assessments from PARCC and Smarter Balanced will probably not yield assessment scores that are completely interchangeable; in fact, there may be some significant differences. The same is true for comparisons of PARCC or Smarter Balanced with other "college-readiness" assessments that are commercially available or are custom developed for particular states or the nation.

The question of how comparable the results of PARCC and Smarter Balanced need to be is also still open. Although the scores may not be interchangeable, it is still to be determined how comparable the scores must be across states within and across the consortia assessments to have sufficiently improved upon the results in Figures 1 and 2.

Framework for considering "comparability": Tests and uses

Sometimes the question is asked, "Are these *tests* the same?" when the real question is, "Can I *use* the results from these tests interchangeably?" The similarity of tests and the interchangeability of scores for specified uses are related, but often not quite the same. Tests might be quite similar but factors related to their use may limit the extent to which their scores are interchangeable. Conversely, test results might come from somewhat dissimilar test instruments, but be used in quite similar ways.

As an example of the former, consider the case in which the same test is administered in multiple states, but there is wide variation among states in policies related to administration procedures, use of accommodations, or scoring procedures. Perhaps the test is used as the basis for student promotion and graduation decisions in one state, but not in the others. Any one of those factors could impact the degree to which we regard the results across states as comparable.

As an example of the latter, consider the comparisons between NAEP and the state assessment results shown in Figures 1 and 2. The many different state tests are used in a similar way to inform policy makers "how many students scored proficient" to support evaluation of the academic attainment of states. A policymaker in Missouri or South Carolina might have little trouble using the NAEP and state assessment results interchangeably to inform a policy decision. Policymakers in many other states, however, would be left with more questions than answers after receiving the results from the two assessments.

From the first example it is clearly implied that administering the same test is not a sufficient condition to support the claim of comparable results. But is it a necessary condition? The examples of the similarity of the NAEP and state assessment results in Missouri, South Carolina, and a handful of other states suggest that the answer to that question is no, tests do not have to be the same to produce comparable results.

Thus, it is essential that the comparability question be framed specifically: is it about tests, score interpretations and uses, or both? This section provides a framework for discussing aspects of test similarity, of use similarity, and of the interactions between them. (See Figure 3.) The examples described above touch on three of the four cases described in Figure 3: Same Interpretation/Uses and Same Tests; Same Interpretation/Uses and Different Tests; and Different Interpretations/Uses and Same Tests. However, the fourth cell, Different Interpretation/Uses and Different Tests is also central to a discussion of score comparability. When considering among options for adopting tests, designing a comprehensive assessment, or interpreting disparate results from two similarly named tests it is important to be able to identify different tests designed for different purposes.

Figure 3: Combinations of Test and Use Similarity and some Examples

Combinations of Test and Use Similarity		
	Test Similarity	
Test Score Interpretation and Use	Different Test	Same Test
Same Score Interpretation and Use	• One state administers PARCC and one state administers Smarter Balanced in 2014-2015 for school accountability • One state administers Smarter Balanced in 2014-2015 and 2015-2016 for school accountability	• Two states administer Smarter Balanced in 2014-2015 for school accountability • Two states administer PARCC in 2015-2016 for school accountability and high school graduation
Different Score Interpretation and Use	• One state administers the high school Smarter Balanced test and the SAT as indicators of college-and-career readiness • One state administers the Smarter Balanced high school test for school accountability and another state administers the PARCC end-of-course high school tests for school accountability and high school graduation	• One state administers PARCC in 2014-2015 for school accountability and another state administers PARCC in 2014-2015 for school accountability and high school graduation • One state administers Smarter Balanced in 2014-2015 for school accountability and another state administers Smarter Balanced in 2014-2015 for school and teacher accountability

Despite the clear categorization in Figure 3, it should be apparent from the examples provided that neither "test similarity" nor "test use similarity" is a matter of a simple "yes/no" or "same/different" dichotomy. Tests may be more similar in degree over a single dimension, or, in fact, over a number of dimensions. For example, one would certainly expect two forms of a test within a single assessment program to be more similar than test forms from two different assessment programs. Even within a single program, one might expect alternate forms administered within a single year (with the same set of administration and scoring conditions) to be more similar in many ways than two forms administered in different years. In each of those cases, however, the forms are different tests that must be linked to produce results that can be compared across forms. Similarly, although there may be clearly different test uses between two states (e.g., one state uses a test for high school graduation and another does not), there are also

degrees of similarity and difference between states in the way tests are used for the same purposes such as school accountability or teacher evaluation. The degree of similarity in the tests themselves and the ways they are used impacts the extent to which *scores* are comparable across two assessments.

Factors that impact the comparability of scores

In this section we will identify several factors that are likely to impact the degree to which we can claim that two tests are similar and are being used for similar purposes. Using the four cells in Figure 3 as an organizer we will begin with factors that may impact the comparability of scores from "same tests with the same uses" and proceed through "same tests with different uses," "different tests with the same uses," and "different tests with different uses." Each of the factors presented for one cell will also apply to subsequent cells.

Same Tests and Same Uses

To ground the concept of same tests and same uses, we will use two states administering the Smarter Balanced tests in 2014-2015 and using the results for school accountability. As a starting point, let us stipulate that a computer adaptive test produces comparable results across students even though students are not assessed on the same set of items.

We can begin by identifying the key factors that would make the results of that test administration comparable across the two states. Of course, the two most important factors are that the content is the same, and the cut scores used to report achievement level results are the same.

Factors that impact the comparability of scores include purpose/use, construct/content, administration conditions, scoring, population of students tested, and opportunity to learn.

What then are the factors that could reduce the comparability of scores across the two states? The traditional criteria for standardization (i.e., content, administration, and scoring) provide two major threats to the comparability of individual student scores: *administration conditions* and *scoring*. Moving from the interchangeability of individual students' scores to the comparability of aggregate scores we add the *population of students tested*. Finally, adding another layer to the interpretation of aggregate scores, we consider *opportunity to learn*.

Administration conditions

Administration conditions include a multitude of factors across several dimensions. On one level there are factors related to test security, including scheduling requirements that minimize the opportunity for students and teachers to share information about the test. On another level there are factors such as testing time (e.g., timed, untimed, loosely timed); testing window; policies related to the use of tools such as calculators or dictionaries; and procedures for test administrators. Finally, there are policies related to the use of accommodations including not

only which accommodations are allowed, but also which students are allowed accommodations and the policies for matching students with appropriate accommodations.

A key administration condition that deserves considerable attention during this transition to the common core assessments is mode of administration. Within an assessment program or within a state, to what extent will some students take computer-based versions of the assessment while others take paper-and-pencil forms? Although previous research has shown that mode effects can be minimal when transferring the same test content from paper-and-pencil to computer, it would be unreasonable to expect little impact when producing a paper-and-pencil version of a test form with content designed specifically to take full advantage of enhanced technology. Within a computer-based administration, to what extent will the test experience vary among students based on the technology used by the particular contractor selected to administer the test or the particular devices used by the students taking the test? For many states and assessment programs, these are new questions yet to be answered.

The greater the extent to which the consortium has established a clear set of administration policies to be adopted and implemented by all test users the greater the likelihood of maintaining comparability of scores. Of course, as with any assessment program, the enacted administration conditions are as important as the adopted administration conditions.

Scoring

Accuracy and consistency in scoring, particularly the scoring of constructed-response items, is another threat to the comparability of scores from the same test. In current state assessment programs, responses within and across students are routinely scored by different scorers, often in multiple sites; and with distributed scoring models there may not be any common scoring site at all. States and contractors have established scoring procedures to enhance the likelihood of accurate and consistent scoring and to monitor accuracy and consistency throughout the scoring process.

The possibility of multiple scoring contractors being used across states administering the same test introduces a new source of variability into the scoring process that could impact the comparability of scores. The potential for variability exists whether student responses are scored by human scorers or different automated scoring engines.

Population of Students Tested

Even if scores are interchangeable at the student level, the comparability of aggregate scores at the state level can be impacted by differences in policies related to which students are tested and included in the reporting of test results. Federal assessment and accountability regulations have standardized participation requirements across states to a large extent, but interstate variations in participation policies must be considered when comparing aggregate results.

Opportunity to Learn

In the current accountability era, opportunity to learn emerged as another factor impacting the comparability of scores at the individual student and aggregate levels. Conceptually, opportunity to learn is related more to the fair use and interpretation of test scores than to their accuracy and comparability in a technical sense. That is, even if all other conditions are identical the manner in which test scores are interpreted and used for accountability purposes may vary based on an evaluation of students' opportunity to learn.

Same Tests and Different Uses

An example of two states administering the same test for different uses might be two states administering the PARCC high school tests in 2015-2016 with one state using the tests as a high school graduation requirement and one using the tests only for school accountability. Variations in intended test use, particularly differences between high-stakes and low-stakes uses of test results, tend to impact the comparability of results in two ways. The first, *motivation*, impacts the comparability of results at the student level as well as at aggregate levels. The second, *unintended consequences*, can have a significant impact on comparability with respect to the generalizability of results.

Motivation

It is generally accepted that higher stakes on an assessment are directly related to higher motivation. Although it is clear that differences in motivation might impact the comparability of results, it is less clear how motivation impacts the validity of inferences drawn from results. That is, it is not clear whether higher test scores on a test used for student graduation are an indicator of greater overall proficiency in the content area or simply better test performance.

Unintended consequences of high-stakes testing

Each of the large-scale assessments being considered here is designed to be an indicator of a student's level of proficiency in the content domain being assessed. That is, generally we are more interested in the level of proficiency suggested or indicated by performance on the test than we are in the actual test score itself. One of the unintended consequences of high-stakes testing, however, is that it has the potential to negatively impact our ability to generalize from the test score to inferences about performance in the broader domain. This may be due to narrowing of the curriculum, greater emphasis on improving test-taking strategies than increasing content knowledge, or more inappropriate approaches to improving test scores. The result is that variations in stakes associated with testing can impact the comparability of scores.

Different Tests and Same Uses

One state administering the Smarter Balanced test and one state administering the PARCC tests in 2014-2015 for school accountability is a prime example of different tests being used for the same purpose. In addition to all of the threats to comparability discussed in the first two cells, we must now add the two key factors that we eliminated from our consideration of the same

tests: possible differences in test content and achievement levels (i.e., achievement level descriptors and/or achievement level cut scores).

Content

The PARCC and Smarter Balanced tests are each derived from and designed to measure student performance against the Common Core State Standards. Differences in how the two consortia interpret individual standards, determine which standards to include on their assessment, and the balance of representation of those included standards on the test are all factors that could impact the comparability of results across the two assessments. Closely related to what content is assessed, the manner in which particular content is assessed could also impact the comparability of results across the assessments. This includes the impact of variations in factors such as test length, test format, and item format which might impact content-related factors such as depth of knowledge and rigor of the assessment.

Achievement Levels

Given that all other conditions are met, the comparability of results across assessments still could be impacted by differences in the definition of achievement levels and/or the cut scores used to classify student performance into a particular achievement level.

As a starting point, we know that the number of achievement levels used to report results will vary between PARCC (5) and Smarter Balanced (4). The extent to which there is a clear mapping across tests between final achievement level descriptors for one of more of the achievement levels remains to be seen. Additionally, as we have seen with current state assessments, consistency in high-level achievement level names such as "Proficient" and even high-level achievement level descriptors is not sufficient to ensure comparability of achievement level results across tests. It will be necessary to validate that the cut scores established for corresponding achievement levels across tests require the same level of performance for the results to be comparable.

Different Tests and Different Uses

The administration of the Smarter Balanced, PARCC, ACT, and SAT high school tests as indicators or college-and-career readiness might be considered as an example of different tests designed for the same purpose. Of course, the intended uses vary significantly across those tests. The ACT and SAT are used for college admissions. PARCC and Smarter Balanced may be used for college placement, but are unlikely to be used for college admission. Each of the four tests may be used for school accountability, but it is unlikely that the ACT and SAT will be used for high school graduation. Clearly, although the tests may share the same high-level use as an indicator of college-and-career readiness, the specific intended uses of the individual tests impacts the comparability of scores across the various tests.

The good news is that there are not really any additional threats to comparability introduced in this cell. All of the threats previously described apply here as well.

Test Information Beyond Scores

To this point, we have focused on scores resulting directly from the assessment such as scaled scores or achievement level classifications. However, the use of large-scale assessments in recent years has included information in addition to those scores. Some information related to test results uses test data to create other statistics. We refer to this as *auxiliary test-referenced results*.

Some examples of *auxiliary test-referenced results* are results from growth models that manipulate test scores, and results from accountability models that qualify, combine, weight, judge, and otherwise bring to bear criteria and rules from outside the test results themselves.

It should be obvious that it is not possible to discuss whether *auxiliary test results* are "comparable" or "similar" without considering the specific rules and situations used to generate those results. One implication is that creating "comparable scores" between two assessments does not necessarily mean that will result in similar auxiliary test results (e.g., value-added growth or school accountability results).

Perhaps less obvious, however, is the question of whether it is possible to create comparable *auxiliary test results* such as growth scores or accountability ratings from assessment scores that are not comparable. That is, can we begin with assessment scores that are not comparable and embed them in models that produce scores which are functionally similar? The answer to that question most likely is yes.

Psychometric Methods for Establishing Links between Assessments

Psychometric criteria for establishing links between assessments that are the same, similar, and different are well established and widely used. As suggested at the beginning of this paper, multiple efforts to link the NAEP tests with 50 different state assessments may have played a role in the effort that led to the CCSS. Links between college admissions tests such as the ACT and SAT are widely known and accepted and there are also efforts to link state assessments with one or both of those tests. Of course, within states hundreds of analyses are conducted each year to link multiple grade-level test forms within and across years as well as to link test forms across grade levels to form vertical scales.

Score Linking and Comparison

Holland (2007) provides one framework for considering the relations between scores and aligning tests to make their scores more inter-changeable. Holland distinguishes between *prediction*, *scale aligning (*or *linking)*, and *equating*. Prediction methods are used to predict a student's score on Test Y using the student's score on Test X. Prediction, through methods such as linear regression, is insufficient to create comparable scores or scales. For example, regressing X on Y often produces a different function than regressing Y on X. *Linking* and

equating are more commonly used in large-scale state assessments to produce comparable scores. Holland distinguishes between six types of linking and equating by using the four key aspects of: Construct, Reliability, Difficulty, and Population. Each of these key aspects is classified according to the extent that they are the same, similar, or different across assessments to be linked. Holland's framework, using these four aspects to classify different methods and types of linking tests, is summarized in Figure 4. Note that test use is not one of the aspects included in the Holland table, although its impact on linking has been established in other frameworks.

Figure 4: A Framework of Score Linking and Comparison (Holland, 2007)

Three Overall Categories of Test Linking Methods and Their Goals Adapted from Holland (2007)		
Linking Test X to Test Y		
Predicting Y from X	**Scale Aligning X and Y**	**Test Equating X to Y**
Goal is the Best Possible Prediction	*Goal is Comparable Scales*	*Goal is Interchangeable Scores*
Weakest form of linking with tests of different constructs, reliability, difficulty, and populations.	Most common form of linking. Encompasses a variety of linking approaches based on the similarity of constructs, reliability, difficulty, and population: • Battery scaling • Scaling to an Anchor Test • Vertical Scaling • Calibration via common population or anchor measure • Concordance via common population or anchor measure	Strongest form of linking requiring tests of the same construct and same intended difficulty and reliability

Cautions about Establishing Links between assessments

An unintended consequence of the ubiquitousness of studies linking assessments over the last decade is a misguided belief that the results of any two assessments can be linked and that the results of linking any two assessments will provide useful information that can be easily interpreted. There is an expectation that simply administering different tests to a common group of students (or a randomly equivalent group of students) or embedding a set of common items on tests administered to different groups of students is sufficient to establish a link and produce comparable scores. Unfortunately, this expectation is too simplistic and optimistic. The *Uncommon Measures* (1999) report, examining techniques for establishing equivalence among different educational tests paints a clear picture of the complexity in trying to establish links between different tests, particularly when the goal is to produce more than high-level results

primarily for descriptive purposes. Recent linking studies with NAEP, state assessments, and international assessments provide clear examples of all of the factors that one must try to control or account for to create a solid link between assessments. Finally, although the field has advanced since 1999 in terms of the technical aspects of linking assessments, little is still known about the appropriate use and interpretation of scores resulting from linking different assessments. In many cases (e.g., college admissions), the appropriateness of particular uses of scores from different tests are determined case-by-case on the basis of specific experiences accumulated over time.

Placing Comparability in Context: State and Other Assessments

We began this document with the premise that the comparability of achievement level results across states was perhaps the preeminent question related to assessment comparability in the era of the Common Core State Standards. Additionally, given that the vast majority of states have chosen the assessments under development by either the Smarter Balanced or PARCC consortium as their state assessment, we have focused on those two assessments in providing examples throughout the document. Of course, there are states not participating in either PARCC or Smarter Balanced that may be developing their own custom state assessment or adopting a common core assessment developed by a commercial vendor. All of the same concerns and issues regarding comparability of scores raised here apply to those assessments as well; and any one of the states and their assessment could be placed into one of the cells in the Test and Test Use table in Figure 3. Although there may not be the same implicit assumption of comparability of state scores with the stand-alone state assessments as there is with the consortia assessments, there almost certainly will be an expectation that the scores can be made comparable.

The context of the Common Core State Standards also raises questions about the appropriate external criterion against which to judge the comparability of state assessment results. As reflected in Figures 1 and 2, NAEP has been widely regarded as the gold standard for comparing results across state assessments. With nearly 50 different state assessments each based on a unique set of content standards and each with their own achievement standards, NAEP was a logical choice as a national yardstick. Although like each of the state assessments, NAEP was a unique assessment with its own set of content standards (i.e., frameworks) and its own achievement standards, it was a common measure across states. Now, however, with virtually all state assessments based on the same set of content standards, and a very limited number of assessments and achievement standards across states there may be less appeal and relevance to an external criterion which is a unique assessment with a unique set of content and achievement standards. At a minimum, the case for NAEP as the yardstick for comparability must be made again in this new era of common standards and assessments before figures corresponding to Figures 1 and 2 are produced for tests administered in 2015, 2017, or 2019.

Finally, the shift from "Proficient" to "College-and-Career Ready" as the achievement goal brings an additional set of questions and assessments into the comparability discussion. There are established assessment programs, such as those developed and administered by ACT and the College Board, that produce a variety of widely used indicators of college readiness for a variety of very specific purposes (e.g., college admissions, course placement, awarding of course credit). NAEP, too, may produce scores that are an indicator of the level of college readiness at the aggregate state or large district levels. What is the expected or required level of comparability among the "college readiness" scores produced by the state assessments and these tests? Beginning with possible differences in their definitions of "college readiness" and moving through similarities or dissimilarities in content, format, administration conditions, populations taking the tests, intended purposes, and actual uses there will be factors that support or detract from the comparability of scores across these assessments. It will be necessary for test users to fully understand how to use the set of assessment results that will be placed in front of them for an individual student, school, district, or state. Which results are intended to be measures of the same aspect of college readiness, which results are intended to measure a unique aspect of college readiness, and which, if any scores can be used interchangeably?

Conclusion: Why is now an especially important time to consider score comparability?

In summary, there are several reasons that score comparability may be more extensively possible now than at any other time in U.S. history. There may be reasons that score comparability is more important now than at any other time in U.S. history (e.g., greater mobility at the postsecondary and K-12 levels; equity across states; and global competitiveness). And policy makers may be committed to certain uses and interpretations of scores now that require score comparability in new or more extensive ways . Some specific aspects why now it is appropriate to consider score comparability include:

- *Common content standards across many states:* More than 45 states have adopted or adapted the "Common Core State Standards," which provide a common definition across states of what students should know in mathematics and English language arts from grades K-high school. Since most states have adopted common learning targets, many state policy makers hope for assessment results that could be compared across as well as within states.
- *Common assessments across many states:* More than 40 states have joined at least one of the two multi-state consortia funded by the U. S. Department of Education to develop mathematics and English language arts assessments of the Common Core State Standards for the general student population. Through the PARCC and Smarter Balanced consortia's assessment programs, there likely will be extensive student assessment data available across many states, providing within-state and cross-state data unlike any available to this point.

- *Common post-secondary uses:* PARCC and Smarter Balanced high school tests aim at reporting the "college-and-career readiness" of high school students, inviting comparisons with existing assessments used in college and work settings.
- *On-going issue of making comparisons across states:* While many states are members of the PARCC and Smarter Balanced assessment consortia, there are other states that have not joined those two consortia, but are using their own state custom assessments and/or commercially available assessments including the ACT and SAT at the high school level. There will be continuing interest on the part of policy makers and others in how the results from all state assessment programs compare, and might be compared. There may be an interest in CCSSO providing support to member states to make and interpret these comparisons.

States and other organizations need to start now *to design and implement studies to inform score comparability actions and interpretations for the next 3-6 years.*

- *Current opportunity to promote score comparability between PARCC and Smarter Balanced programs:* PARCC and Smarter Balanced are planning to administer their assessments operationally for the first time in spring 2015. CCSSO and other entities are quite interested in what might be done to help promote more useful assessment results from PARCC and Smarter Balanced, including work prior to 2015 to design the two assessment programs to yield more comparable scores where appropriate, and in studies to be conducted starting prior to 2015 and extending beyond to analyze the relationships of scores in the future. Now is a good time to plan that design work and those studies.
- *Current opportunity to advise states on score comparability within their own assessment programs, and interactions with accountability and other uses:* States that have adopted the Common Core State Standards (CCSS) need to administer new assessments designed to measure the CCSS. Most of those states need to consider how to transition across new assessment programs while maintaining accountability programs, such as student graduation and/or promotion and scholarship qualification; school accountability; and educator effectiveness evaluation. A key consideration in that transition is what scores will be kept comparable over time, and which will change, when, and how that will be accomplished.
- *Now is the time to design score comparability studies:* As a practical matter, states and other organizations concerned about score comparability need to start now to design and implement score comparability studies to inform actions and interpretations for the next 3-6 years.

References

Duncan, A. (2010). Beyond Bubble Tests and Bake Sales: Secretary Arne Duncan's Remarks at the 114th Annual National PTA Convention. Washington, DC: U.S. Department of Education. Transcript retrieved 4/3/13 from http://www.ed.gov/news/speeches/beyond-bubble-tests-and-bake-sales-secretary-arne-duncans-remarks-114th-annual-nationa

Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W., & Hemphill, F. C. (Eds.). (1999). *Uncommon Measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Holland, P.W. (2007). A Framework and history for score linking. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and Aligning Scores and Scales*. NY: Springer.

McCombs J.S. & Carroll, S.J. (2005). Ultimate Test: Who is accountable for education if everybody fails? Santa Monica, CA: RAND. Report retrieved 4/3/13 from http://www.rand.org/publications/randreview/issues/spring2005/ulttest.html

National Center for Education Statistics. (2007). Mapping 2005 State Proficiency Standards onto the NAEP Scales. (NCES 2007-482). U.S. Department of Education. Washington, DC: Author. Report retrieved 4/3/13 from http://nces.ed.gov/nationsreportcard/pubs/studies/2007482.asp

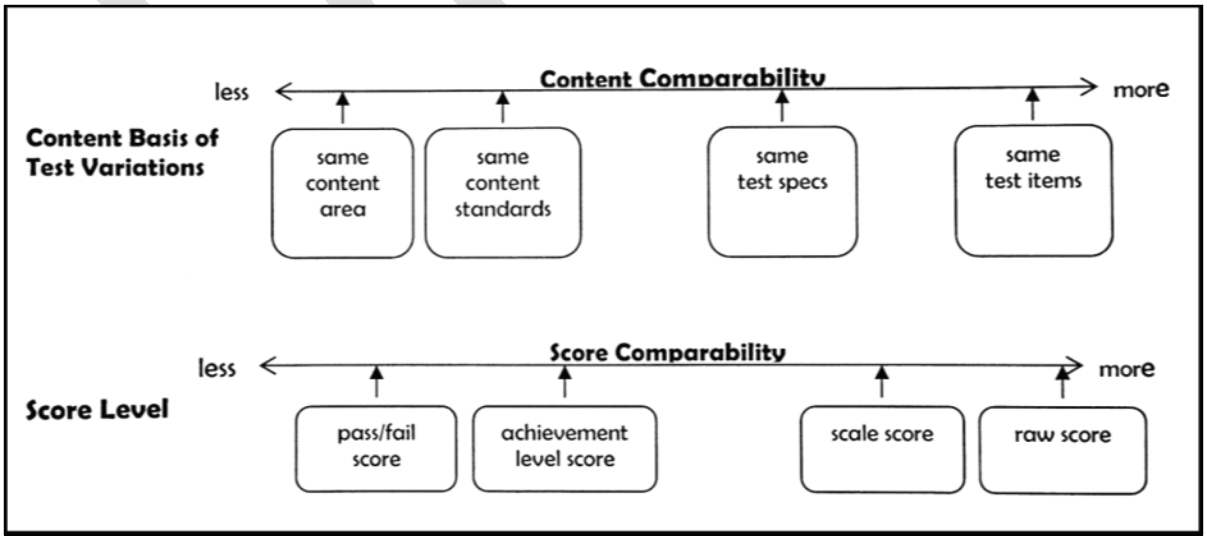
Winter, P. C. (2010). Introduction. In P. C. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations.* Washington, DC: Council of Chief State School Officers.

Appendix A

PARCC and Smarter Balanced: Level of Comparability

In terms of construct similarity, PARCC and Smarter Balanced share much, but also have significant differences. It is likely too late for PARCC and Smarter Balanced to make major changes in their test designs. It is clear that possible score comparability across the two assessments will be less than absolute score interchangeability. Determining the appropriate level of content and score comparability, however, will depend upon empirical analyses and also definitions of intended use. While construct similarity has been recognized as very important, there are few methods to analyze construct similarity, let alone quantify or evaluate how similar two tests are in terms of constructs. One recent approach, however, is presented below and applied to the case PARCC and Smarter Balanced.

"The comparability of test scores is a matter of degree" (Winter, 2010, p. 5). "How comparable scores need to be for a specific test variation depends on how the test scores will be interpreted and used" (ibid, p. 5). Winter (2010) provides two interrelated continua that help evaluate similarity of tests in terms of content and score comparability. Winter's work is informative for our purposes, although Winter was concerned about describing tests and "variants" that were in some way alternative forms of a general state assessment, such as accommodations for students with disabilities to the regular administration procedures; simplified language for English language learners of the test items; administration of a paper-and-pencil version of the test compared with computer-administration of the same version; and the "AA-MAS assessment" intended for persistently low achieving students (capped for accountability purposes at 2% of the population). The same principles apply given that we are interested in the comparability of results from alternative assessments (i.e., PARCC and Smarter Balanced) designed to measure achievement against the same set of content standards (i.e., the CCSS) and to produce comparisons against the same criterion (i.e., college-and-career readiness).



The first of Winter's continua describes the content basis of variations between the alternative tests and addresses the extent to which the alternative tests might be measuring the same construct. The four broad "buckets" placed on the continuum from less comparable to more comparable range from tests which purport to measure the same content area in name only (e.g., the grade 4 NAEP mathematics test and a current grade 4 state mathematics test) to tests based on the same set of content standards to tests based on the same content standards and built to the same test specifications to tests based on the same content standards, test specifications, and containing the same test items (e.g. tests in which the only differences might be due to accommodations offered or the format of administration).

The magnitude of the gap between "same content standards" and "same test specs" on the continuum, although arbitrary, is intended to indicate the tremendous extent to which test design decisions can impact the comparability of test results even if those tests are derived from the same set of content standards. At this time, it is safe to conclude that the Smarter Balanced and PARCC assessments, although built on the CCSS will differ in terms of key test specifications. A first step in evaluating the comparability of results from the two assessments, therefore, is to develop a solid understanding of the similarities and differences between the two assessments in terms of what they actually measure and how they measure it. Those are the issues raised in the body of this document.

The second continuum, score comparability, attempts to relate content (or construct) comparability to the level at which scores might be interchangeable between two assessments. At the "more comparable" end of the continuum we have the cases in which two tests with the same items are administered (perhaps in a different format) and it may be possible to claim that raw scores are interchangeable. The next highest level of score comparability in which scaled scores are considered interchangeable refers to cases such as annual forms of state assessment built to the same test specifications and linked in some manner to produce scores on the same scale. Note that this is a higher level of score comparability than results from attempts to produce a concordance table relating scaled scores from two assessments that may or may not be built to the same content standards or test specifications.

As we move further to the left on the score comparability continuum, our goal is to compare high-level judgments of the overall quality of a performance against an established and common criterion (e.g., college-and-career readiness). It can be argued that this is the most relevant and appropriate level of comparison between scores on the PARCC and Smarter Balanced tests – tests that are derived from the same content standards for the purpose of producing an indicator of performance on the same criterion (i.e., college-and-career readiness). What is missing from the two continua presented, however, is a representation of the importance of a) a common and well-established definition and understanding of the criterion and b) common achievement standards when considering score comparability. The importance of a shared understanding of

the criterion or "standard of excellence" is well documented by Wiggins (1991) in his article calling for increased attention to standards, not standardization[1]. As we will discuss in the next section, this is a starting point for evaluating the comparability of results from PARCC, Smarter Balanced and other college readiness tests.

How can we evaluate the conceptual similarity of PARCC, Smarter Balanced, and other college readiness tests?

With a clearly defined and common definition of college-and-career readiness accompanied by rich, descriptive exemplars of the performance of college and career ready students as our standard, achieving high-level score comparability across Smarter Balanced, PARCC, and other college readiness tests without the standardization of a common test would be a challenging, but achievable goal. To the extent, however, that there is no shared common definition of college-and-career readiness and that performance level descriptors vary across the consortia, even high-level claims of score comparability may be difficult to support.

A first step, therefore, in evaluating the score comparability across the PARCC and Smarter Balanced assessments is to examine their definitions of college readiness and their descriptions of college-ready performance. That is not to say, however, that the two assessment programs will not be able to produce results that are equally useful even if they are not strictly comparable. As an example of this, we have the college readiness benchmarks established for two existing assessments, the ACT and SAT. As is easily seen in the statements below, the college readiness benchmarks established by ACT and the College Board are fundamentally different on multiple levels, but both provide useful and interpretable information about student readiness for college.

ACT Information Brief (2013) What are ACT's College Readiness Benchmarks?

> *ACT's College Readiness Benchmarks are the minimum ACT® Test scores required for students to have a high probability of success in credit-bearing college courses – English Composition, social sciences, College Algebra, or Biology…Students who meet a Benchmark on the ACT Test or ACT Compass have approximately a 50 percent chance of earning a B or better and approximately a 75 percent chance of earning a C or better in the corresponding college course or courses.*

College Board (2011) SAT Benchmarks (research report 2011-5)

> *The college readiness benchmark was calculated as the SAT score associated with a 65 percent probability of earning a first-year GPA of 2.67 (B-) or higher. The SAT benchmark determined in this study was 1550 for the composite [critical reading, mathematics, and writing SAT score].*

[1] Wiggins, G. (1991). *Standards, Not Standardization: Evoking Quality Student Work.* Educational Leadership, 48 (5), pp18-25, Feb 1991.

For some, the predictive, outward-looking definitions of college-and-career readiness described above may not be appealing or may not seem sufficiently standards-based. Those people may prefer to stake their claims of comparability on the assumption that sufficient achievement of the knowledge and skills contained in the Common Core State Standards (CCSS) constitutes college-and-career readiness and to the degree that both the PARCC and Smarter Balanced Assessments are designed to measure attainment of the CCSS results from the two assessments can be comparable. This inward-looking premise is valid. However, it merely shifts our attention from the score comparability continuum to the content comparability continuum and all of the issues impacting comparability that were discussed in the opening of this section and presented in the body of this document. Specifically, a determination of the conceptual similarity of PARCC and Smarter Balanced would be influenced by factors such as

- the manner in which they interpret individual standards and clusters of standards,
- the balance of representation of standards on the assessments,
- the way in which content is assessed (i.e., test formats and item types), and
- the ways in which items are scored and overall scores and performance ratings are determined.

A timeline for comparing PARCC and Smarter Balanced

As development and implementation of the Smarter Balanced and PARCC assessments continues, we can divide the activities related to evaluating the comparability of PARCC and Smarter Balanced into three distinct phases: Pre-2015 Questions and Comparisons, Comparisons in 2014-2015, and Post-2015 Scenarios and Issues. In this section, we provide a brief overview of major tasks that can be accomplished and questions answered in each phase.

Pre-2015 Questions and Comparisons

Several key questions related to the comparability of the assessments can be addressed in the next year – before their initial administration in 2014-2015. In the preceding section we identified several questions related to the comparability of the consortia's definitions of college-and-career readiness and description of college level performance that one can begin to answer immediately. Also, although our focus for that discussion was on the high school assessments and college readiness, comparisons of the claims and performance level descriptions at all grade levels can provide valuable information about the potential comparability of the results from two assessment programs.

In the next year it will also be possible to examine in detail the test specifications for the two assessments. Historically, there have been few methods for systematically analyzing the content of tests, comparing and evaluating tests in terms of construct or conceptual similarity. One approach that has emerged in the past decade is "alignment methods" that quantify content similarity. Alignment methods typically involve classifying, counting, and checking whether the quantitative ratios are "good enough." Widely cited alignment methods include those developed

by Webb, Porter-Smithson, Schmidt et al., and Achieve. Significant variants have been developed by WestEd and NCEO. The complexity of the CCSS combined with the complexity of the planned assessments, however, makes it clear that current alignment methods based on classifying and counting are not likely to be sufficient to evaluate tests designed for the CCSS. We anticipate that modified and new alignment protocol will emerge in the next year as the need to evaluate the alignment of tests to the CCSS increases.

By the time that the PARCC and Smarter Balanced assessments are first administered in 2014-2015 key differences in design, format, and specifications and the impact of those differences on the comparability of assessment results should be fully understood.

Comparisons in 2014-2015

Following the 2014-2015 school year, it will be possible to begin comparing results from PARCC and Smarter Balanced. An important factor in interpreting differences in results across the two assessments will be to understand the characteristics of the states participating in the two consortia. This includes understanding their prior performance (e.g., recent NAEP results), planned uses of the assessment results (e.g., school, teacher, and student accountability), and level of implementation of the CCSS. By the beginning of the 2014-2015 school year, it should be clear which assessment is being administered by each state, and it will be possible to begin this analysis. Then when the tests are administered, the focus can shift for the first time to the actual results of the assessments.

The earliest comparisons might be made between Smarter Balanced and PARCC assessment results will be fall 2015 or winter 2016.

Smarter Balanced and PARCC are scheduled to have assessments available for operational administration in grades 3-8 and high school in spring 2015. Smarter Balanced is scheduled to set achievement level cut scores in 2014 following a large field test. PARCC plans to set proficiency level cut scores in 2015 following the first operational administration, and so it is likely that final PARCC scaled scores and proficiency level results will be available in fall 2015, with detailed comparisons of the achievement level results between the assessments possible in fall or early winter 2015.

Previously we have discussed the impact of the performance level descriptors and the test themselves on the comparability of the results. To fully evaluate the relationship between achievement level results, it will also be necessary to understand how standard setting is conducted for each of the assessments. We have already discussed the importance of defining college readiness at the high school level and the impact that this definition is likely to have on standard setting. At grades 3-8, the approach each consortium uses to vertically articulate the standards across grades and create a link to the high school achievement standards will be particularly important in evaluating the comparability of the assessment results. Different

models of the relationship across grades of the meaning of "on track to college-and-career readiness" could impact the comparability of results at each grade level.

It will also be valuable to analyze the technical characteristics of the assessments and results within each of the assessment programs. Smarter Balanced scores and achievement level results should be available in fall 2014 or winter 2015 to analyze the characteristics of the Smarter Balanced assessment, and to provide comparisons of student achievement across states, districts, and schools that administer Smarter Balanced assessments. PARCC results should be available a year following, in fall 2015 or winter 2016. Thus, the earliest comparisons might be made between Smarter Balanced and PARCC assessments based on large-scale operational empirical data would be fall 2015 and perhaps winter 2016.

Post-2015 Scenarios and Issues

Beyond the initial administration in 2014-2015, the comparability discussion can be expanded to include issues related to growth (either growth computed centrally by the consortium or individually by states), improvement, stability of results, and eventually, empirical studies of the relationship between assessment results and college-and-career readiness.

Another key to long-term evaluation of the comparability of scores after 2015 from PARCC, Smarter Balanced, and any other assessments focused on college- and career-readiness and/or the Common Core State Standards will depend on the similarity and stability of the testing programs. Changes in any factors discussed in this paper—the tests, test administration procedures, tested populations, or other conditions such as motivation due to changes in accountability—may affect the score comparisons.

One issue is whether test administration conditions will change over time. Individual states have taken considerable care to exercise some control over test administration policies and practices to provide acceptable standardization to support making desired score comparisons across tested units and over time. Multi-state assessment programs have paid similar attention to unified test administration policies and practices—examples are the NECAP program (involving Maine, New Hampshire, Rhode Island, and Vermont), the shorter-lived ADP Algebra I and Algebra II end-of-course exams, and the earlier New Standards Reference Exam.

Smarter Balanced has announced a policy that states will be responsible for contracting administration and scoring of Smarter Balanced tests with centralized guidance and quality control; Smarter Balanced assessments will not be centrally administered through a single organization. It is not yet clear how much cross-state standardization in test administration states will strive for that administer PARCC assessments. If there are changes in the tests or other conditions over time, it may be necessary to periodically redo analyses to establish score comparisons within and between the different assessment programs.

Comparison Template

Comparing PARCC/ Smarter Balanced

Evaluating Smarter Balanced-PARCC Conceptual Similarity

Many people assume that Smarter Balanced and PARCC assessments are a case of triangulation: that the tests are intended to measure the same content standards and should provide two measures of the same thing.

In fact, using Winter's continuum of Content Comparability, the two assessment consortia differ in ways that may prove substantial challenges to making interpretations about score comparisons. As more becomes known about the design and content of the two assessments, completing the information in tables such as those provided below can provide valuable information to help policymakers understand similarities and differences between the assessments and the level of comparability to expect from the two assessments.

	Content Comparability	
Content Basis of Test Variations	Smarter Balanced	PARCC
Same general content standards	CCSS	CCSS
Same content area *(reporting areas)*	ELA: Reading, Writing, Research	ELA: Reading, Writing, Research, Listening
Same (individual) content standards		
Same test specifications *(Balance of emphasis across standards, focus for individual standards, depth of knowledge as revealed in claims/PLDs/ALDs, item format, integration of skills/practices, etc.)*		
Same test items and item scoring		

	Score Comparability	
Score Computation	Smarter Balanced	PARCC
Method for computing an overall composite score across the performance and end-of-year components		
Score Level	Smarter Balanced	PARCC
Pass/fail score		
Achievement level score		
Scale score		
Raw score		

Appendix B

Policy Maker's Summary

- The intended uses and interpretations of assessment results drive the type and degree of score comparisons the assessment should be designed to support and report. Conversely, once an assessment is designed and implemented, it will only support well certain types of score comparisons and interpretations.
- Assessments may be viewed on a continuum along the dimensions of being conceptually similar and functionally similar. The more two assessments are conceptually and functionally similar, the more their scores are interchangeable. Absolute interchangeability is the highest degree of score comparison.
- Much is known about requirements for making tests and scores more conceptually similar. These procedures include attending to conceptual similarity in construct, reliability, difficulty, tested population, test design, test administration, test scoring, test scaling, and test reporting. Some procedures are emerging to design and analyze the relationships of tests that are designed to be conceptually similar. As multi-state consortia developing but not administering new assessments, PARCC and Smarter Balanced have challenges in meeting the requirements so that their tests will produce interchangeable scores across states and over time within their own program; the challenges are even greater for supporting score comparisons between programs. Some procedures are commonly used to analyze the functional ("use") similarity of different tests, but there are few established guidelines for interpreting the results of such analyses.
- There are several assessments that may be related to claims of assessing college- and career-readiness. These include ACT, SAT, AP, CTB's "*TerraNova* Common Core" assessment, College Board's *ACCUPLACER*, ACT's *Compass*, ACT's *WorkKeys*, the GED, and state assessments intended to be fully aligned with the Common Core State Standards (CCSS) such as Kentucky and New York. Assessments that are under development to make claims about college-readiness include PARCC, Smarter Balanced, ACT's announced *ACT Aspire* product system, the transformed GED, and perhaps NAEP. It is likely additional CCR tests will be made available by 2015 by other developers. These tests may be analyzed for type of degree of possible score comparisons using the framework introduced in the white paper.
- In terms of conceptual similarity, PARCC and Smarter Balanced share much, but also have significant differences. It is likely too late for PARCC and Smarter Balanced to make major changes in their test designs. Possible score comparisons will be less than absolute score interchangeability in terms of conceptual similarity. Functional similarity will depend upon empirical analyses and also definitions of use.
- Policy makers and others may wish to know more about the conceptual and functional similarity of several of the available (and intended) tests, and how these tests may support the policy makers' intended uses and interpretations.

- CCSSO and many others are planning activities to learn about conceptual and functional similarity and possible score comparisons, and to make that information available in a variety of ways. It might be mutually beneficial for these organizations and individuals to be informed of each others' efforts and to coordinate and/or collaborate as appropriate.