Some Thoughts about Comparability Issues with "Common" and Uncommon Assessments

Scott Marion & Marianne Perie

Center for Assessment

March 25, 2011

Background and Introduction

There was little doubt that the push for common standards was intended only as a first step towards assessments or assessment systems where the performance of students might be compared across states or other units. Further, this comparability was an expressed goal of the Race-to-the Top assessment program and both general assessment consortia included comparability as a key component of their proposals. Politicians and policymakers have expressed a desire to be able to compare "their academic performance" to the performance of other states and to other nations (i.e., "internationally benchmarked"). Either people are not being clear or there is a lack of precise terminology about what people actually want compared. In many cases, it is both. For example, comparisons could range from the performance of individual students in different states (i.e., "interchangeable scores") to simply ensuring that achievement expectations across states are roughly the same across states. This is not the first time that the challenge of comparing performance across states has been considered. Braun & Qian (2007), Feuer, Holland, Green, Bertenthal, and Hemphill (1999), Koretz, Bertenthal, and Green (1999), and Mislevy (1992), have produced important publications documenting the challenges and potential approaches for comparing scores across states using NAEP, state tests, and other measures. The central challenges that Mislevy (1992) outlined in *Linking Educational Assessments* are still with us as we consider the issues in the current Race to the Top environment:

- discerning the relationships among the evidence the assessments provide about conjectures of interest, and
- figuring out how to interpret this evidence correctly" (p. 21).

It is important to shed more light on both the technical and policy concerns related to cross-state comparisons so that state leaders and federal policymakers can more fully understand the

ramifications of the choices they will make in the upcoming months and years. The take home message of this brief is that there are considerable trade-offs that policymakers will have to consider in order to achieve a high degree of technical comparability. This brief document is a first step in trying to help reveal and clarify many of the issues at play when it comes to comparing the performance of students, schools, districts, and states. Our goal is to help provide some guidance for policymakers and other stakeholders in thinking through these competing priorities.

The Common Content Standards

One of the purposes of the CCSSO/NGA Common Core State Standards (CCSS) initiative was to ensure that students in all states were expected to have the opportunity to learn meaningful content and skills. Implied in the "common" part of the "core state standards" is that this would be the foundational piece of the standards-based assessment and accountability system to facilitate cross-state comparisons and address the concern that some state leaders were not holding their students to appropriate expectations. Of course, if content-based cross-state comparisons are desired, designing tests from common content standards is an important first step[1].

Common Performance Level Descriptors

We assume that states within each consortium will administer a common summative tests and use a common standard setting process to establish common cutscores for proficient and other achievement levels, similar to the New England Common Assessment Program (NECAP) process. In doing this, the states within the same consortium would have to agree on common performance level descriptors (PLDs): those statements that translate the content and skill statements into descriptions of expectations of "how well" students need to demonstrate this knowledge. We argue that if there is a desire to compare the performance of states operating in different consortia, then performance level descriptors common to both consortia are an important starting point. Research on standard setting has demonstrated the considerable

[1] We recognize that there are many other reasons—at least as important as cross-state comparability—for creating common content standards. This paper is focused only on comparability issues. Further, proponents of CCSS should articulate a theory of action that lays out how and why comparable expectations across states will lead to the types of educational improvements the proponents seek. Again, this is beyond the scope of this paper and for now, we accept that some sort of comparability is required.

influence of PLDs on where in the score distribution that cutscores are established (Hambleton, 2001; Impara, Giraurd & Plake, 2000, Loomis & Bourque, 2001, Perie, 2008). Comparably can be achieved without common PLDs, but we argue that starting from common PLDs, based on the same content standards, will lead to much more meaningful comparisons of achievement at specific benchmarks. Therefore, while we and others have advocated that the organizers of the CCSS generate common PLDs, aligned with each grade and subject, it is clear at this point in time that they are leaving this task to the consortia. The respective consortia will need to develop grade and subject specific PLDs, but still broad enough to provide test developers and policymakers a target while also allowing them to supplement the PLDs with specific details after a standard setting when they know more about the types of items or performance demonstrated by students at each level.

Some Basics of Comparability

Test score equating or linking is the most common way in which we address comparability goals in our current testing context. The goal of equating is to disentangle differences (across different forms or tests) in item or form difficulty from changes in actual student achievement. A common example involves ensuring that the scores on the state's 5th grade mathematics test in 2009 can validly be placed on the same scale as the 2008 scores. In this example, different students (the 5th graders in 2008 and 2009) will have completed tests containing a different set of items, except for a subset of items that were administered in both 2008 and 2009. It is this subset of items—assuming many conditions are met—that allows us to disentangle the changes in student achievement from the changes in the difficulty of the other (non-linking) items on the test. The challenge is ensuring that the assumptions are actually met. Most psychometricians are familiar with the often repeated Lord's Paradox that the only time all of the assumptions for test score equating are met is when the same students take the same test under the same conditions. Of course, when those conditions are met, we do not need to equate.

Following is a brief description, relying largely on Mislevy's (1992) framework, of three approaches to linking, from a traditional psychometric perspective (in descending order of equivalence). We present this as a context for many of the issues discussed in the remainder of the paper.

Equating. The linking is strongest (and simplest) if the two tests were designed from the same test blueprint to measure the same construct(s). Holland (2007) describes the purpose of equating is to be able to use scores interchangeably and results when the tests measure the same construct with the same intended difficulty and reliability. The most common example is two or more forms of the same test. "Under these carefully controlled circumstances, the weight and nature of the evidence the two assessments provide about a broad array of conjectures is practically identical" (Mislevy, 1992, p. 21).

Calibration. If the two (or more) tests were not designed from the same test blueprint, but both have been constructed to provide evidence about the same type of achievement (e.g., same construct), then the scores can be related through calibration. "Unlike equating, which matches tests to one another directly, calibration relates the results of different assessments to a common frame of reference, and thus to one another only indirectly" (Mislevy, 1992, p. 24). There are several situations in which calibration is used. The two most common are: (1) constructing tests of differing lengths from essentially the same blueprint, and (2) using IRT to link responses to a set of items (e.g., item bank) built to measure the "same construct" (Mislevy, 1992, p. 24). Holland (2007) describes this approach under an umbrella of "scale aligning" with the purpose of "transforming scores from two different tests onto a common scale" (p. 12). This might be applicable to reporting scores from two different common state assessments onto one scale.

Projection. "If assessments are constructed around different types of tasks, administered under different conditions, or used for purposes that bear different implications for students' affect and motivation, then mechanically applying equating or calibration formulas can prove seriously misleading: X and Y do not 'measure the same thing.' (Mislevy, 1992, p. 24). Mislevy's concern here is that the two assessments measure qualitatively different information. Oftentimes, projection is used to make statements like "a student who scores X on Test A would have a 75% probability of scoring

between Y and Z on Test B." While there are often very good reasons for administering different tests to different students, even though we might want to make statements that imply a certain level of comparability such when concordance tables are used to relate ACT and SAT scores, users need to understand that such inferences of comparability are much weaker than with equating or calibration.

Conditions for comparability

As seen in these definitions, Mislevy and Holland focus considerable attention on the match between the two test blueprints and for good reason, because this is a crucial concern if we are to compare student-level scores from two different sets of tests. But there is a long list of other standardization concerns as well. In fact, there is a good reason that most end-of-year state tests are referred to as "standardized tests."

The test must be administered and scored under standard conditions in order for the results across time, students, schools, or other units of interest are to be formally equating with the results placed on the same scale. Standard conditions almost always include the following:

- ✓ Specified test administration windows
- ✓ Common administration rules/procedures
- ✓ Standard accommodation policies
- ✓ The same inclusion rules
- ✓ Clearly specified and enforced security protocols
- ✓ Specific computer hardware and software (for CBT and CAT), or at least a narrow range of specifications.

This is just a minimum set of conditions. More stringent requirements could specify the arrangement and number of students in the testing setting as well as narrowing the testing window such that all students test on the same day at the same time like the SAT.

Two other considerations are not often addressed explicitly in psychometric discussions of comparability. The first involves the population of test takers, especially in terms of trends over time and the second involves the learning opportunities experienced by the students. Part of the population issue may fall under inclusion conditions, but post-Katrina analyses conducted by Richard Hill demonstrated, albeit in an exaggerated case, the effect of changing populations on

assessment and accountability results. Interpreting cross-state assessment trends would require a thorough understanding of the inclusion and population dynamics. In other words, even if we assumed that we had met all critical conditions for comparability for any given year, we know that policymakers are interested in evaluating score trends over time. Variable shifts in populations from one state to another can make interpreting these trends far less comparable than one might think even when other conditions for comparability are met. At a minimum, even if the trends could be psychometrically calculated, interpreting such trends would require an understanding of the population shifts behind the scores.

Many argue that having common content standards and common assessments are sufficient for ensuring comparability. At the very least, many argue that standardization conditions, such as those listed above are also needed. Furthermore, assuming many of the conditions described above are satisfied, differences in curriculum and instruction—the second condition—may lead to much less sophisticated inferences about score differences. Policymakers and others might think that by having students with different curricular backgrounds participate in a common assessment it could provide good evaluative information about the effectiveness of certain curricular materials. However, this condition would tempt people to search for easy explanations (different curricula) for score differences when there is likely much more going on. For example, most test items, especially those measuring sophisticated knowledge and skills

Non-psychometric considerations

It is important to note, however, that in only one of Mislevy's three categories above are students taking essentially the same test. In both the calibration and the much weaker projection case, students are not taking the same tests, but there are methods to relate the scores from one assessment to another. In other words, students do not need to take the same test in order for us to make comparisons about their relative performance. This is important to keep in mind in the current push for common assessments.

In fact, many other countries do not use conventional psychometric approaches at all to ensure that students completing different examinations are being held to the same standard. Many senior secondary examinations (e.g., "A Levels") such as those governed in England by the

Qualifications and Curriculum Authority (QCA) or examinations used for determining "awards" (course "grades) in many Australian states use approaches that rely on the expertise of the personnel involved (e.g., markers/scorers) to establish equivalent standards of performance from one year to the next. Most of these systems also use some sort of statistical moderation procedure (e.g., normative comparisons of results across years) as a check on the expert judgments (Newton, et al., 2007). Systems such as those used in the Australian state of Queensland go further still and use an intensive moderation process to establish the comparability of results from school-based assessment systems. In the Queensland case, comparability is established first within schools and then across schools through a process whereby peers and experts examine systematic samples of student work to ensure that students are held to comparable standards. The exams in England and Australia described here have significantly higher student stakes than most U.S. exams, yet procedures that many in the U.S. would find "soft" are held in higher esteem in other countries than more psychometric methods. These international examples illustrate that there could be a larger set of methods for establishing the types of comparability desired by many policymakers than the narrow psychometric approaches generally used in this country.

Even with the U.S. context, there are other defensible approaches to linking test scores at certain points, such as a judgmental linking study to link cut scores that could be valid for comparing the performance of students or other units across different tests or other sources of evidence (e.g., Burton & Linn, 1993). Of course, there must be a basis for such comparisons, such as common expectations. For example, if there were reasonably common expectations for college and workforce readiness, students could complete different assessments or present different sets of evidence to demonstrate that they have met these common expectations. This does not mean that students' scores from different assessments or collections of evidence can be interchanged as in the strict equating sense, but these different sets of scores can be compared against a common threshold (e.g., college readiness) to judge students' performance. For example, we cannot directly compare the scores of students who complete International Baccalaureate (IB) or Advanced Placement (AP) exams, but as long as students meet expectations on either assessment, our inferences about college readiness would likely be comparable.

A Theory of Action or Why are We Interested in Comparisons?

In most of the discussions about the need for cross-state comparability, there is little attention to the explicit rationale for how increased cross-state comparability will improve the educational system. Clearly, there are several implied reasons for needing increased comparability, but it is critical to make these rationales explicit before making firm design commitments. Creating a theory of action can be a useful exercise to help with this process. A theory of action helps to connect the specific, often policy-related, inputs (e.g., increased comparability) with the intended outcomes (e.g., increased rates of college and career readiness) through a type of logic model where the specific mechanisms that enable the system to move from the inputs to the outcomes are revealed. The rationale and coherence of these theories of action can be evaluated both logically and empirically. In this case, in addition to specifying the type of comparability desired (e.g., interchangeable student scores across states), a theory of action would reveal the processes and mechanisms for how achieving the required comparability will lead to the intended outcomes. Statements such as, "we want to be able to compare our performance across states to ensure that we maintain rigorous expectations" offer an explanation for how this increased comparability might actually improve the educational system in any of the states being compared. Specifying a theory of action can help guide design decisions such that if comparability was a desired goal, we would then have the specificity necessary to create a system to achieve the desired goals.

What Do We Want to Compare?

The previous discussion focused largely on comparing scores of students completing either the same or different assessments. Current policy discussions have stressed the desire for comparability, but the conversations have been remarkably non-specific. As noted above, a theory of action can help reveal the "why" and even the "how." A well-specified theory of action should also explicate the "what," but given the importance of this topic, we devote a separate section to this discussion. The "what" should be driven by the theory of action such that the mechanisms and processes should inform us whether interchangeable student scores or NAEP-like state comparisons are necessary to achieve the intended outcomes. There are a range of possible comparisons in the current policy context, some of which include comparing:

- ✓ Student scores within the same multi-state consortium using the same form of a test

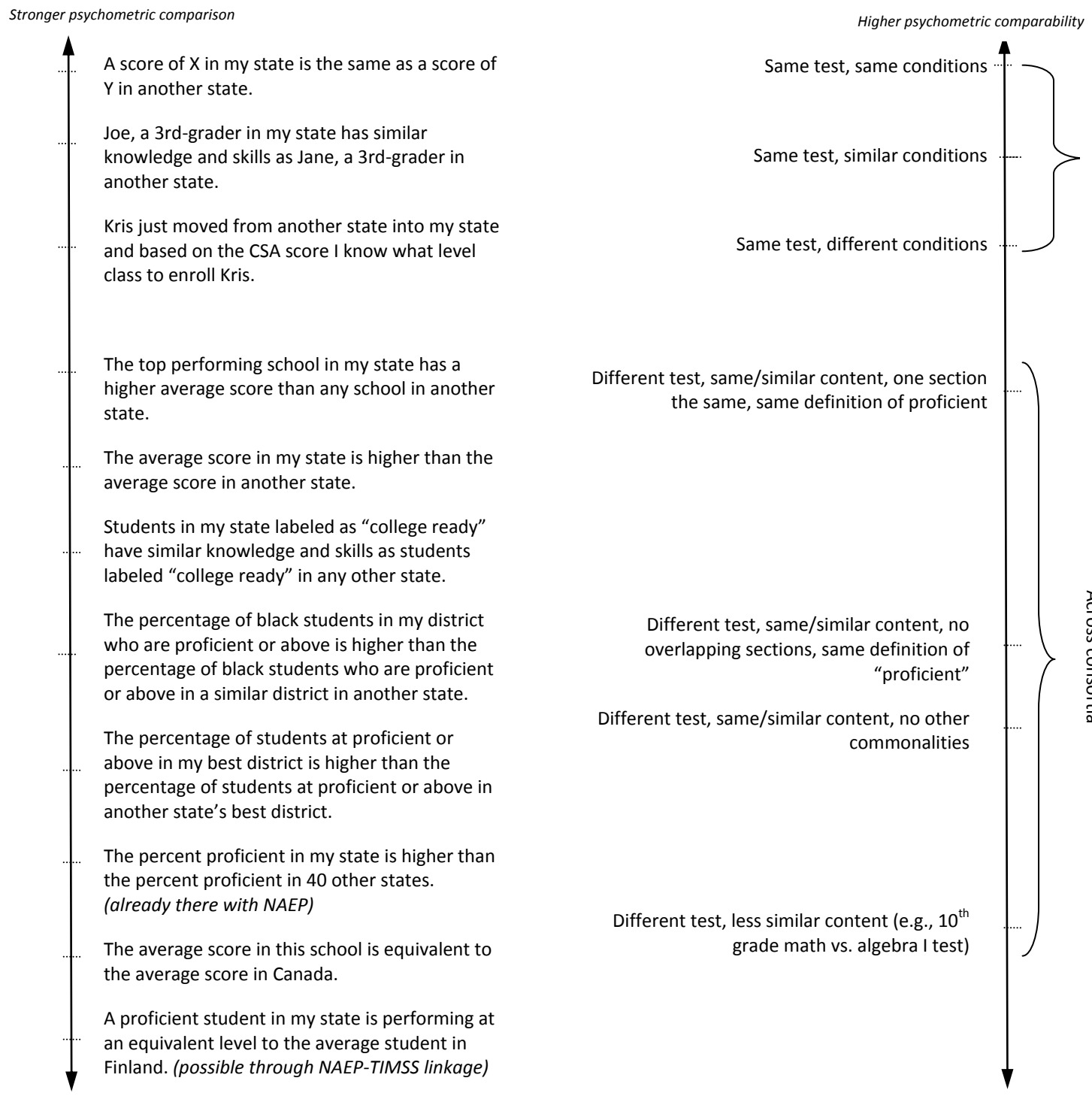
- ✓ Student scores within the same multi-state consortium using adaptive tests
- ✓ School performance within the same multi-state consortium
- ✓ State performance within the same multi-state consortium
- ✓ Student scores across different multi-state consortia using different tests
- ✓ School performance across different multi-state consortia using different tests
- ✓ State performance across different multi-state consortia using different tests
- ✓ Performance of students completing different complex assessments and/or courses
- ✓ Performance of schools where students complete different complex assessments and/or courses

What are our "leads"?

To help think through the questions of "what", we thought it would make sense to think through the types of statements that policymakers would like to make and that might serve as the "leads" for media stories about assessment results. Figure 1 below presents a selection of these statements graphically. As can be seen from the figure, the strictest student-level comparability requires the same test to be administered under the same conditions. Importantly, there are many statements listed in this figure of the sort that we have heard advocated that do not require the same tests. We acknowledge that this figure likely oversimplifies the ordering of the statements and assessment conditions. There could be instances where the ordering can actually flip. This is because the figure collapses several factors (e.g., content, test, many conditions) onto a single outcome. But it is possible, for example, for different tests/similar conditions to be more comparable than same tests/different conditions if the two different tests were measuring the same math content under very similar conditions while the same math test was administered to students using different testing windows and accommodation policies[2]. The figure provides a good general overview of the trade-offs between comparability statements and design and administration conditions, but we urge the reader to not view this as a specific formula or menu.

[2] Thanks to Brian Gong for this example.

Figure 1. Continuum of Comparative Statements and Types of Comparability*



*Assuming no additional external linking/equating studies.

Note that no inferences should be made about the differences in the level of the comparisons based on the distance between comparisons.

We outlined above some important conditions to consider when comparing assessment results. It is useful to return to this discussion in the context of this figure. We noted that the strictest, equating-type, comparability requires the "same test administered under the same conditions." However, we think it is useful to unpack this phrase. At a minimum, the same conditions must include the same:

- ✓ test administration windows (or at least the same number of school days prior to administration)
- ✓ administration rules/procedures
- ✓ accommodation policies
- ✓ inclusion rules/rates(?)
- ✓ security protocols
- ✓ computer hardware and software (for CBT and CAT), or at least a narrow range of specifications.

Additionally, as we noted earlier, clear documentation of the population to which the assessment results are intended to generalize is an important condition for understanding the extent of the achieved comparability. For example, if one state has a 30% in/out migration rate over three years and the students who move into the state are lower achieving than those already in the state, then maintaining the same achievement level would actually be an improvement in performance. But if this state was being compared to a state whose performance actually increased with only a 5% in/out migration rate over three years, we might incorrectly infer that the second state is performing much better than the first state. Of course, growth and/or value-added models (VAM) can help ameliorate some, but not all, of these concerns. That is, if policy makers want strong enough comparability to be able to say that student scores are interchangeable, they must be willing to agree to some very stringent design and administration conditions.

We suspect that much of the interest in comparability is focused on state level comparisons rather than interchangeable student scores. With that in mind, we created Figure 2. The conditions are similar to those outlined in Figure 1—and the same caveats apply—but the statements in each of the boxes are focused on state-level statements.

Figure 2. State-to-State Comparisons

Compared to another state, my state…

…uses a different assessment with no overlapping section but assesses the same content standards and uses the same definition of proficiency.

Statements I may be able to make:

- ✓ A similar percentage of students demonstrate proficiency in my state as in the other state.
- ✓ The black-white achievement gap in terms of percent proficient is larger/smaller in my state than in the other state.
- ✓ More/Fewer Hispanic students moved from not-Proficient to Proficient over the last two years in my state than in the other state.
- ✓ Approximately the same percentage of students in our two states reached the "college-ready" bar.

…uses a different assessment to assess the same content standards but includes one overlapping section with the other state's test and uses the same definition of proficiency.

Statements I may be able to make:

Everything in the previous box, and:

- ✓ Students in my state labeled as "college ready" have similar knowledge and skills as students labeled "college ready" in the other state.
- ✓ Black third-graders in my state are outperforming black third-graders in the other state.
- ✓ The black-white achievement gap, in terms of percent proficient, has decreased more in my state than in the other state over the past 2 years.
- ✓ If a student from the other state moved into my state, I could estimate their score on my state test based on their score on their state test.

…uses the same assessment but under very different conditions (e.g., different inclusion rates/testing windows).

Statements I may be able to make:

Everything in the previous boxes, and:

- ✓ Fifty percent of fourth-graders in my state answered this item correctly compared to 40% of students in the other state.
- ✓ The black-white achievement gap in terms of scale score differences appears to be slightly larger/smaller in my state than the same gap in the other state.
- ✓ A greater percentage of Hispanic students in my state are college ready than in the other state.

…uses the same assessment but under slightly different conditions (e.g., a <u>slightly</u> different accommodations policy or testing window).

Statements I may be able to make:

Everything in the previous boxes, and:

- ✓ If a student in my state has the same score as a student in another state, their level of knowledge and skills is about the same.
- ✓ The average score of my best school is higher/lower than the average score of the other state's best school.
- ✓ The percentage of students reaching the "college ready" benchmark has grown at a faster rate over the past three years than the percentage in the other state.

…uses the same assessment under the same conditions.

Statements I may be able to make:

Everything in the previous boxes, and:

- ✓ Student scores have the same meaning for both of our states and are virtually interchangeable.
- ✓ Over the last two years, the black-white achievement gap, in terms of scale score differences, decreased more/less in my state than in the other state.

International Benchmarking

In addition to state-level comparisons, we have heard many state leaders and other policy makers indicate that they want to make sure that their state (or consortium) assessments are "internationally benchmarked." This phrase suffers from the same lack of specificity as many other comparability statements. We suspect that most consortia will not build strict psychometric comparability for international comparisons into their assessment designs, and in this case, the U.S. policymakers can only control one of the two tests they are interested in linking. However, we wanted to provide an overview of types of statements that might be made and the associated design conditions required to make such statements.

Regardless of the number of consortia or the number of states taking the same assessment, in order to make statements about student or state performance compared to the performance of students in other countries, we will need to link or otherwise evaluate the connection between state/consortium assessments and international assessment (e.g., TIMSS, PISA, or PIRLS). Depending on how we make that linkage, we can make different statements.

Using a systematic content-mapping/alignment analyses that evaluates and compares the items on consortium and international assessments, we can say:

- ✓ The 5th grade mathematics test used in our consortium is holding students to the same approximate level of expectations as the 5th grade mathematics test in Germany

Through international benchmarking (post hoc statistical analyses with no overlapping content, we can say:

- ✓ Massachusetts students performed at a similar level to Finland students, on average.
- ✓ The cut score for Proficient in Massachusetts is equivalent to the 60th percentile score in Finland.

Through embedding items from international assessments into state assessment, we can say:

- ✓ Students in Massachusetts answer this item correctly 65% of the time, while students in Finland answer the same item correctly 44% of the time (+/- 5%).
- ✓ Students in Massachusetts do better, on average, on items related to literary texts, while students in Finland do better, on average on items related to informational texts.

Through having the same students take both tests in a balanced linking design (a very unlikely case), we can say:

- ✓ A score of X on the Massachusetts state assessment is equivalent to a score of Y on the international assessment.
- ✓ Half of students in Massachusetts outperform the top 20% of students in Finland.

Discussion and Recommendations

We have attempted to outline some of the considerations and trade-offs associated with trying to ensure that consortium assessments can allow for comparability both within and across consortia. We argued that interchangeability of student level results or even school or state level results cannot be assumed simply because states are using a common assessment within the same consortium. Ensuring strict psychometric comparability requires adherence to a fairly strict set of administration and design conditions. If states are unwilling to agree to these conditions, then they might want to consider a more flexible design where certain types of comparability might be achieved, but where innovation might be encouraged.

We strongly suggested that states/consortia develop a theory of action that articulates how and why designing assessments to produce a certain level of comparable results will lead to intended educational outcomes. This exercise should start with explicating the types of statements state policymakers want to make about their students and schools compared to students and schools in other states. It would also be worthwhile to consider design alternatives that allow for comparability approaches other than strict psychometric ones (that might, for example, encourage the use of performance assessments) in a theory of action to see if these alternatives produce a more viable theory of action than ones based on strict comparability.

These two major points were intended to lead readers to recognize that there are always trade-offs in design activities. This is true when designing for comparability and it is important for policymakers to understand the ramifications of designing for strict comparability. We fear that a strict psychometric approach toward cross-state comparability may constrain assessment innovation. We urge policymakers to conduct a type of "cost-benefit analysis," in the context of a theory of action, before making design decisions that make strict comparability the highest priority.

However, we are not blind to the fact that some level of comparability is desired. State and national leaders have indicated that the variability in content and achievement standards across states was unacceptable. We suggest, therefore, that consortia employ a "threshold" approach where all states agree to adopt the common content standards, create common performance level descriptors, and employ defensible standard setting methods that incorporate external empirical information as part of the process. This minimal design would at least allow state leaders to be able to say "the expectations for students in my state are similar to those in State X."

The other issue to consider is how the content standards are translated into test specifications. Of course, for states within the same consortium, we would expect to see the same test blueprints and either the same tests or multiple forms with embedded linking item sets. For states in different consortia that are interested in ensuring that students are being help to similar expectations, there is a range of potential linking designs, all of which would require some significant trade-offs depending on the differences in the base assessment designs. Starting by communicating intended weighting or enacting of the content standards across consortia would help maintain a level of construct comparability. If state leaders wanted to evaluate the content and performance expectations across the two or more consortia, a content-mapping/alignment study similar to the one suggested for the international comparisons could provide additional evidence about the commonality of expectations.

We offer these alternatives so that policy makers may see that considerable comparability could be realized without having to require that all students take the same test at the same time under the same conditions. We argue that innovation and rigor should be a high priority in assessment design and are concerned that an overly intense focus on comparability could shut off other important design priorities. Further, we are pessimistic that psychometric comparability can be achieved because it will be almost impossible to have all states agree to strict administration conditions. In other words, we are worried that we could end up with a "lose-lose" situation, where the focus on comparability shuts off potential innovation and we still do not end up with the comparability desired.

References

Braun, H. I. & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.) *Linking and aligning scores and scales*. New York, NY: Springer.

Burton, E. & Linn, R., (1993). Report on Linking Study--Comparability across Assessments: Lessons from the Use of Moderation Procedures in England. Project 2.4: Quantitative Models To Monitor Status and Progress of Learning and Performance. Washington, DC: U.S. Department of Education.

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

Hambleton, R. H. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

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*. New York, NY: Springer.

Impara, J. C., Giraud, G. & Plake, B. (2000, April). The influence of providing target group descriptors when setting a passing score. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Koretz, D.M., Bertenthal, M.W., and Green, B.F. (1999). *Embedding questions: The pursuit of a common measure in uncommon tests*. Washington, DC: National Academy Press.

Loomis, S. C. & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G.J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.

Newton, P., Baird, JA, Goldstein, H., Patrick, H. & Tymms, P., Eds. (2007). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority (QCA).

Perie, Marianne. (2008). A guide to understanding and developing performance level descriptors. *Educational Measurement: Issues and Practice, 27*(4) pp. 15-29.