Thoughts on Linking and Comparing Assessments of Common Core Standards

Stephen Lazer, Vice President, Assessment Development, ETS

John Mazzeo, Vice President, Statistical Analysis & Psychometrics Research, ETS

Walter D. Way, Senior Vice President, Psychometric & Research Services, Pearson

Jon S. Twing, Executive Vice President, Assessment & Information, Pearson

Wayne Camara, Vice President, Research & Development, The College Board

Kevin Sweeney, Executive Director, Psychometrics, The College Board

May 2010

ETS® PEARSON CollegeBoard

Introduction

It is likely that even given the federal interest in developing assessments of common standards, a single national test will not emerge. The purpose of this paper is to discuss the types of comparisons that can and cannot be made among students who take different assessments supposedly developed to measure a single set of standards.

There are many reasons for developing new assessments of common core standards in mathematics and English language arts (ELA). The assessments provide a way to promote the standards themselves; people value what is tested. Having a common assessment shared by a consortium of states could allow for efficiencies and economies of scale in the assessment development and analysis processes, which should, in turn, free up some funds to enable the use of more innovative assessment techniques. Moreover, common assessments provide a way for states within a consortium to measure students against a common yardstick. Today, there is no easy and rigorous way to compare the performance of individual students or schools in different states. While the National Assessment of Educational Progress (NAEP) provides data to compare states and, in some cases, large school districts, it does not support school or individual reporting. Also, NAEP exists only at three grades (4, 8, and 12). At grade 12, NAEP remains a national, not a state-level, survey.

If students take the same assessment under the same conditions, a given score in one place has the same meaning as it does in all others. As all students have been presented the same tasks, it is possible to discuss the sorts of things they can do at different points on the scale. This includes the sorts of performances that it takes to reach levels such as “basic” or “proficient.” These behavioral characteristics are much harder to compare if students are taking different tests, in that the specific tasks used to provide evidence may vary substantially.

While a common assessment will maximize comparability of results, from a policy standpoint, it is unrealistic to expect that all states will agree to use the same assessment. A common assessment would not permit the level of local control and customization states will likely want to employ. In addition, different state consortia may use different common core assessments. Even if tests are built to measure the same standards, the scores will not necessarily mean the same things, particularly if the number and type of exercises on the assessments vary, or content emphasis is different. This leads to several questions: In the absence of common assessments, how can one make performance comparisons that purport to have been built to a common set of standards? What kinds of comparisons are and are not possible? And what types of techniques can be used to allow for comparisons?

The answer to these questions will depend on the nature and rigor of the comparisons that are desired. If stakeholders want to be able to make statistical comparisons across jurisdictions as to the growth shown by specific students, or the skills profiles of specific schools, then a common assessment or very similar assessments are required. If, on the other hand, states want a general way to compare the rigor of their performance levels, or a method to help set performance levels of comparable rigor, then greater variability in assessments may be tolerable and various linking approaches can prove helpful.

These questions are more than academic. The United States Department of Education (USED) is planning to give consortia of states grants to develop assessments of common core standards. Awarding multiple grants would allow for experimentation and differentiation but will also affect the comparability of results between consortia. The degree to which the results will be comparable will depend on a number of factors, including the extent to which consortia are willing to work together to facilitate comparisons and the overall similarity of the assessment systems. The purpose of this paper is to provide some initial thought on what may be possible in the way of comparing results across consortia if multiple assessment systems emerge.

Discussions in the professional assessment literature about linking different assessments as a way of comparing their results have often been highly technical. To better communicate with policymakers, educators, and members of the public, this paper takes a different approach. Specifically, we discuss linking and comparing assessments under four scenarios that represent different degrees of cooperation among consortia and different degrees of similarity among assessment systems.

- The first scenario imagines a situation in which consortia work together to develop a substantial common core of their summative assessment system, while allowing some within-consortium customization.
- The second scenario presumes less cooperation and a smaller amount of common material, but still assumes reasonable similarity in the nature of tasks included in the assessments of the different consortia.
- The third scenario assumes some common item types and some small amount of common material between the two assessments, but substantial differences in the non-common portions of the assessment.
- The fourth scenario assumes two very different assessment systems, and little cross-consortium cooperation in assessment design (though it does allow for some, in terms of special studies).

In each case, the paper discusses the sorts of comparisons that might be possible, and what sorts of procedures might be implemented to allow for those comparisons. All scenarios are based on the idea that different tests are built to a single set of standards.

Scenario 1

This scenario is similar to the assessment design we suggested for a single consortium in a previous paper, entitled *Thoughts on an Assessment of Common Core Standards* (Lazer et al., 2010). Two different consortia agree to work together on the common core of their assessment systems. This core represents roughly 80 – 85 percent of the assessment materials any student will receive and could provide a score on its own. In the remaining 15 – 20 percent, each consortium decides to do testing specific to that consortium. Note that this scenario could also apply to different states within a single consortium, as some states may wish to supplement the common core assessment with state-specific content.

There are two ways in which cross-consortium comparisons are possible, and that involve different degrees of rigor and different limitations on inference. The cleanest and most general comparisons will be those that are made by simply treating the core assessment common to both consortia as a stand-alone test. In this case, students will have completed a common test, so normative and criterion data will have the same meaning. The only technical requirement is ensuring that the core assessment is administered in the same way in both consortia, so there are no context or fatigue differences that influence results and comparisons. The consortium-specific materials would be added to the core elements separately for each consortium. These elements would provide data that could be part of within-consortium comparisons (e.g., between states, schools, or students within a consortium) but not cross-consortium comparisons.

While this approach provides for direct cross-consortium comparisons, it does introduce one set of challenges. Since comparisons would be made based only on the common core, each consortium would, in effect, have two sets of results — one for cross-consortium comparative purposes and one for within-consortium use. The existence of two sets of results, each appropriate for somewhat different uses, could be cumbersome for score users and confusing to the policymakers and the public. An alternative approach that could mitigate such challenges might be to use the common test material as a “linking block” in an attempt to express all results (common, Consortium-A-specific, and Consortium-B-specific) on a common scale. There are various ways one might accomplish this linking, including a concurrent calibration of all the items onto a single Item Response Theory (IRT) scale.

Such an approach seems immediately attractive — it would allow comparisons on a single cross-consortium scale. However, there are a number of considerations that must be addressed concerning the validity of the results of such a linking. The results will be most valid in situations in which the consortium-specific components are similar to the common core elements of the assessment and to each other. In other words, this strong linking approach will make the most sense if the consortia are using their custom section in similar ways.

It is instructive to consider this situation from both logical and technical perspectives. In placing items from different assessments on a common scale, one is essentially arguing that based on the tasks students received, one can estimate with a high degree of certainty what they would have scored on the exercises they did not receive. When are such assumptions warranted? To use the mathematics test as an example, consider a situation in which both consortia used the non-common elements for extended problem-solving exercises. These items shared areas of focus but were somewhat more advanced and cognitively complex in one consortium. Or consider an ELA situation in which the complexity of texts used in the non-common sections differed, but the item types and skills measured were the same. In these cases, one might justify a single scale approach. If the two consortia wish to achieve a single scale, they should limit the differences between their custom pieces.

However, the type of inferences it is appropriate to make on the basis of this common scaling must be carefully considered. For example, consider the case in which Consortium A adds extended answer questions and essay writing in its non-common section, while Consortium B uses short constructed-response and simulation-based questions. Perhaps one could create a single IRT scale; this is not certain. But statistical success is not the whole story. Total scores would not have the same meaning in Consortium A and Consortium B when the nature of the skills required for the material specific to each of the tests differs. In situations where the non-common elements are radically different, one might need to be comfortable with less ambitious inferences, such as examining the relative alignment of performance levels. We discuss this situation further in Scenario 3.

Scenario 2

Scenario 1 assumes substantial overlap between two assessments and a great deal of similarity. It is important, as well, to consider scenarios in which greater disparity exists across consortia with respect to assessment content, item types, and test delivery modes. Scenario 2 reflects a moderate degree of disparity. Both consortia build assessments that mostly make use of the same item types, although the mix of exercises and the delivery mechanisms differ. Consortium A builds summative exams for each grade level based on the common standards that make extensive use of computer delivery. Its summative assessment consists primarily of multiple-choice items and a limited number of short constructed-response or performance tasks, which are not necessarily machine scored. Consortium A's summative exam is adaptive — the sets of items administered to each student are selected sequentially, in real time, to be at an appropriate level of difficulty based on responses to prior sets of machine-scored items. Because it is adaptive, Consortium A's exam is relatively efficient, requiring a total of 60 minutes of student testing time — 45 minutes for the multiple-choice section and another 15 minutes for the constructed-response section.

Consortium B builds a somewhat different assessment, though again, based on the common standards. Its summative assessment is given on paper and consists of a multiple-choice section and a substantial constructed-response section. The constructed-response section consists, for the most part, of exercise types similar to those used by Consortium A, but here, they make up a greater percentage of the assessment, include some more extended tasks, and are not machine scored. Consortium B's test is not adaptive — within an assessment year all students are administered the same collection of items. Because the exam is not adaptive and includes a greater amount of constructed-response items, it requires considerably more *total* testing time than is required for Consortium A's summative exam. Total testing time is 90 minutes — 60 minutes for the multiple-choice section and 30 minutes for the constructed-response section.

In contrast to Scenario 1, the consortia in Scenario 2 have not agreed to regularly administer a common test. However, the summative exams for the two consortia in this scenario do have some important similarities — they are based on the common standards and, in large part, make use of the same item formats (multiple-choice and short constructed-response items). There are, however, some notable differences with respect to delivery mode (computer adaptive versus conventional paper-and-pencil), emphasis on performance tasks (Consortium B's exam includes some extended performance tasks), and student testing time required. The similarities between the exams, and the fact that both are based on the common standards, provide an opportunity for establishing empirical relationships and thereby, establishing comparability between the scores from the two exams, though, to be sure, the scores will not be as rigorously comparable as those available through the common test under Scenario 1. However, the differences between the exams pose data collection challenges that will need to be addressed to allow for the estimation of empirical relationships, and will require effective cooperation between the two consortia to support valid interpretations of the results.

The most efficient empirical approach to establishing comparability would involve Consortia A and B agreeing to share items — either in a given year or on an ongoing basis. As both exams are developed from the same framework and share some common item types (in our example, multiple-choice and short constructed-response items), the two consortia could agree to include a reasonable number of common items in their exams. These items would be delivered adaptively, by computer, in Consortium A and conventionally, in paper-and-pencil format, in Consortium B. Like the alternative approach described under Scenario 1, the resulting data from the two consortia could be jointly analyzed with commonly used psychometric models based on IRT, allowing the results from the two different exams to be expressed on a single common scale. However, as was the case for Scenario 1, the assumptions that support the validity of assuming a common scale need to be carefully evaluated. The difference from Scenario 1 is that the common items make up a much smaller share of the combined assessments, and, as a collection, do not provide comprehensive coverage of the common standards. Therefore, results on the common items stand on their own as a useful source of information for direct comparisons of results across consortia.

Even more than was the case in Scenario 1, the validity of results arising from this kind of "common-item" approach depends on some important assumptions — the most important being that the psychometric functioning of the items (e.g., the difficulty of the items) is the same, regardless of whether they appear in Consortium A's or in Consortium B's exam. In the current context, that would amount to assuming that item functioning is not affected by whether: (1) the item is administered by computer or in paper-and-pencil formats, (2) it is administered as part of an adaptive test or as part of a conventional test, and (3) the amount of time available for each item and the other kinds of items being presented along with the items differ. It would be prudent to check these sorts of assumptions empirically. Moreover, it would be important that the items represent aspects of the standards that receive similar emphasis in both consortia (and therefore, could be reasonably assumed to function comparably) and are sufficient in number to produce a dependable, generalizable result.

If the consortia are able to cooperate in the early stages of establishing their respective testing programs, opportunities for obtaining comparability data may be improved. For example, Consortium A's development effort would likely include a field test that would not involve adaptive administration, but rather, a large number of field-test forms that were spiraled among students in order to obtain the IRT item parameter estimates needed for subsequent computer adaptive administration. If comparability data were obtained at this time, some of the challenges inherent in the adaptive versus linear administration approaches might be overcome.

Alternative designs for establishing comparability, beyond the common item approach, might also be considered, particularly if the sorts of actions just discussed are not practical or are tried and demonstrated to be ineffective in establishing comparability relationships. For example, one might involve creating a common linking test from material deemed appropriate for either exam and administering that test in the same administration mode and under the same timing conditions to students from both consortia. So, to be concrete, a common 60-minute test consisting of multiple-choice and short constructed-response items might be created and administered as a conventional paper-and-pencil test to students from both Consortium A and Consortium B. Most likely, such a test would be given to a *sample* of students from the two consortia as part of a separate special study. Alternatively, this same test could be delivered by computer in Consortium A and as a paper-and-pencil test in Consortium B, to keep the mode of presentation as familiar as possible for the students in each of the consortia. Under either alternative, for students in the special study sample, two sets of exam scores would be obtained — those that arise from their normal required summative exam and those that arise from the special study. Given such data, a number of statistical approaches are available to establish either predictive or concordance relationships between the scores from the two different summative exams through the common linking test.

The common linking test design has potential advantages over the use of embedded common items. Administering the exact same test under the exact same conditions (or at least as a conventional test in both cases with common timing) may provide better control of the kinds of potential context effects that could invalidate the IRT assumptions required for establishing a valid link in the embedded item design. However, the common linking design presents its own unique challenges, as well. It requires a separate additional exam administration, and such a requirement may be inconvenient or impractical to arrange. Consortia would need to work cooperatively and creatively to establish commonly standardized administration conditions, to minimize the intrusiveness of the extra testing, to obtain adequate and representative sample sizes for the special study, and to sufficiently motivate students taking the common linking test to give their best effort.

Scenario 3

Scenario 3 has one common element with Scenario 2: the two consortia have agreed to share a modest number of common items. However, beyond this commonality, the two assessments are quite different. For example, assume the common items are all multiple choice. Consortium A uses short-answer questions amenable to automated scoring in addition to these questions. The ELA assessment includes measurement of listening, as well as items that assess reading and writing skills. Consortium B uses extended-answer questions and dynamic problem-solving tasks involving computer simulations, but does not test listening.

In this case, even though the assessments share common items, these items really do not help much from a linking perspective. Since the items do not represent enough coverage of the standards to stand alone, comparisons on the common materials will not be sufficient. In addition, the common items cannot be used as an equating block or as the basis for a single scaling because they do not adequately cover the constructs measured in either of the two assessments. Even if one could use statistical machinery to create an apparent scale, it is not reasonable to assume one can create scale scores that have a common meaning across consortia. Therefore, any attempt to compare results from these two assessments would most likely not make use of the common items but would need to rely on other methods. In one approach, students from the two consortia could be asked to take the tests from the other. However, the results from this sort of a study might well be suspect, in that student motivation on the two tests could be variable, and the different instructional contexts might influence the results in complicated ways. Perhaps a more promising approach is to do what has been done in studies relating state assessments to NAEP. In this case, since all states participate in NAEP, analysts have been able to use statistical approaches to map state achievement-level cut scores onto the NAEP scale. However, these sorts of comparisons — whether, in a normative sense, the achievement levels on one test appear higher than those on another — likely represents the limit of what is possible. Any interpretations that suggest the sorts of specific skills students who take one assessment would display on the other are unlikely to be supportable.

Interestingly, the comparison situation would in fact be better for two assessments that shared no common items but had highly similar structures, content, and exercise types. In that case, a study that asked students to take both assessments might yield sensible results and support either concordance tables or projections. In the scenario described here, because the common items are not representative of either assessment, they serve little benefit from a linking perspective.

Scenario 4

Compared to the other scenarios, this represents the fewest possibilities for establishing valid relationships between the results of the assessments. In this scenario, the assessments contain a different mix of exercise types and have no common materials. Any rigorous type of linking is not possible because there are no common items in the two assessments, and no common students who would take both assessments. More importantly, each assessment arguably measures different types of cognitive skills in responding to the test questions. Consequently, there is no straightforward mechanism for making comparisons. Any comparative statements about performance across the two assessments should reflect these constraints and be general and descriptive in nature, so as not to imply more precision in the relationship between the two assessments than actually exists.

What sorts of comparative statements might be possible, and what steps would need to be taken to defend those statements? As mentioned above, the assessments do share a common set of content standards that they are measuring and a common set of achievement-level definitions. This commonality provides the basis for making comparative statements. In other words, one could try to compare the rigor of the achievement levels, or make comparative statements about what students on each assessment who are ranked in the same achievement level know and can do.

Given the degree of difference between the two consortia's assessments, perhaps the most tractable way to approach the making of comparative statements about the results from the two consortia would involve the adoption of one of the following approaches. First, as described under Scenario 3, one could conduct a special research study (or series of studies) to determine an empirical relationship between the two assessments by relating each to a third, independent assessment (such as NAEP). This might allow one to place achievement-level cut scores on a common metric, to determine, again, solely from a normative perspective, the relative rigor of the achievement levels on the two tests. In other words, one might be able to determine whether, when placed on the NAEP scale, the achievement levels on the

Consortium A tests were higher or lower than those of Consortium B. Of course, this approach only works where there is a suitable independent test to use. For reasons described above, having students take the test from the other consortium is likely to be problematic.

The other possible comparative approach is descriptive in nature. As mentioned above, assume that both consortia used the same a priori achievement-level descriptors to inform development and in their level-setting processes. If one then takes the resulting levels as "givens," one could look empirically at the sorts of tasks students rated at a single achievement level ("proficient," for example) tended to achieve on both assessments. This could make use of an approach called "scale anchoring."

To be a bit more specific, let us assume that both assessments have populations of students who are called proficient based on their test scores. For each test, one can look at the sorts of items or tasks on which these students score well. This, in turn, will allow us to create a description of what proficient students can do on each of the two tests. These descriptions can be compared so people can make judgments about the meaning of the scores. Of course, these comparisons are not quantitative, but they may still be of use to policymakers.

Summary

These four scenarios are discussed to provide some sense of what sort of comparisons should be possible if there are multiple common core assessments. In the first scenario, in which common elements can stand alone, the full range of comparisons at individual and group levels are possible if those comparisons are limited to common items. No special analyses and data collections are needed, and no major assumptions must be made. It may also be possible to have almost the same freedom of valid comparisons if non-common elements are included in the analysis. However, this possibility exists to the extent to which the non-common elements are, in fact, fairly similar.

As the amount of common materials decline, and as they become less representative of the construct measured, underlying assumptions become a larger part of making comparisons between the results of different assessments. This markedly increases the degree of caution required in interpreting the results. It also makes it more likely that comparison will require special data collections and analyses.

Finally, where there are very different assessments, one cannot expect to support comparisons at the individual test-taker level. Scores will mean different things in terms of what students know and can do. The best one can hope for are linkages that let us compare general attributes, such as the rigor of achievement levels.

In closing, testing experts are often asked, "How good is good enough?" when it comes to comparing the results of different assessments. The answer is that it depends on the purposes of the comparison and uses of the data. If one wishes to make tight comparisons on a common yardstick that have high stakes, then a common assessment is called for. If one wishes to make more general comparative statements, other approaches may be sufficient. The price of getting this wrong can be major. One would be uncomfortable if some schools were rewarded because their students appeared to show higher annual gains than others, but this was a function not of true growth, but rather, of a difference in assessments. This is not an argument for doing nothing, but rather for matching the rigor of approach and inference to the realities of the data.

References

Lazer, S., Mazzeo, J., Twing, J.S., Way, W.D., Camara, W., & Sweeney, K. (2010). *Thoughts on an assessment of common core assessments* (ETS, Pearson, & the College Board white paper). Princeton, NJ: Educational Testing Service.