

How Common Standards Might Support Improved State Assessments

Laurens L. Wise, HumRRO
Draft: January 14, 2010

This paper was prepared to stimulate and support discussion at a workshop on best practice in state assessment, convened by the National Research Council's Board on Testing and Assessment, December 10-11, 2009, in Washington DC. A wide range of ideas for improving state assessments was presented in earlier papers. The focus of the present paper is on ways in which the adoption of common content standards would support needed improvements. This discussion is timely in that the NGA and CCSSO collaboration is about to release common core standards for K-12 student achievement and 46 states have agreed to consider their adoption (McNeil, 2009).

The paper is organized into three sections. The first section discusses why support for these improvements is needed, including ensuring the feasibility and sustainability of potential improvements. The second section describes new research to estimate potential cost savings if several states pooled resources to build common assessments. These savings could then be invested in improvements to the assessments that are developed. In addition to savings through development of common assessments, a common set of well-articulated content standards will support deeper cognitive analyses of how the standards are learned as well as how mastery of them might be assessed. The final section of this paper describes specific improvements that might be supported through increased funding and cognitive content analyses. These include: (a) more meaningful reporting scales, (b) more timely results, (c) richer diagnostic feedback, (d) greater teacher engagement, and (e) validity studies to support wider interpretations and uses.

Limitations of Current State Assessments

Limited Coverage of Instruction and Student Achievement

Current state assessments do not support the very wide range of uses that policy-makers would like to make of them. Under NCLB requirements, most state assessments are reasonably valid for measuring the extent to which some students are being "left behind." The assessments

focus on core subjects (reading, mathematics, and science) and report whether students have met levels of achievement expected of *all* students. Current state assessments are quite limited, however, for other potentially important uses including broad accountability measures for schools, performance indicators for teachers and principals, criteria for evaluating specific educational programs, and provision of diagnostic information on individual students.

By focusing on a limited number of subjects, current assessments measure only a portion of what we want schools to teach, particularly at the high school level. They are, therefore, quite limited as school accountability measures. Also, by focusing on a single proficiency level, current accountability systems are not good indicators school's impact on all students, including those well above or below minimal proficiency.

Current assessments are also quite limited as measures of instructional or teacher effectiveness. Again, they measure only a small portion of what is taught. Rarely are topics such as history, civics, or foreign language acquisition included. How can we evaluate instruction in these topics? Here too, the focus on minimal proficiency often limits our ability to gauge teacher's contributions to students well above or below this level.

Finally, current assessments provide little or no diagnostic information on instructional needs for individual students. End-of-year or end-of-course assessments provide information that is not available until after the student has either moved on to the next grade or class. In addition, current assessments cover a very broad range of content and are unable to provide reliable information on specific student strengths and weaknesses. Where sub-score information is provided, the measures are highly correlated and differences in mastery of different objectives or clusters of objectives are not estimated reliably.

To fill the gap left by the lack of diagnostic and instructional effectiveness measures, states and districts are increasingly turning to benchmark or interim assessments. These assessments are largely unconnected with the summative state assessments and may or may not provide much valid additional information.

Need for Improved Reliability and Validity

Improvements are needed to increase the reliability and fairness of score results and the validity of a wider range of uses of these results. Currently, content alignment studies provide validity support for the interpretation of test scores as mastery of specific objectives. But there is little evidence of relationships between the reported test scores and instructional approaches and

also little evidence that score levels at one grade predict readiness for what comes next. The National Assessment Governing Board is just beginning a range of validity studies to support “preparedness” interpretations of 12th grade results from the National Assessment of Educational Progress (NAEP). To date, there is little corresponding research for state assessments.

Current content standards are quite broad and the pressure to test a wide range of content in a limited amount of time, once toward the end of the school year, severely limits the precision or reliability of the resulting scores. Measures of mastery of specific clusters of objectives or sub-domains of contents are based on a few items at most and not reliable at all.

We would also like to be able to interpret assessment results as providing valid comparisons for all groups of students. Principles of universal design for making assessment questions maximally accessible for all students have been espoused, but are not well understood, particularly by those who set testing policies. Research on the validity of different testing accommodations has not been widely generalized for state assessments. In addition, the current focus on what is known, more than what is taught, gives advantages to students from higher socio-economic families that distorts the use of assessment results as measures of value-added by schools or specific instructional programs.

Support for Feasibility and Sustainability of Targeted Improvements

The considerable funding that may be available for initial development of common assessments should be invested in ways that ensure the feasibility of targeted improvements. This could mean funding infrastructure development to support computer-based testing that would both improve the quality of the assessments and reduce ongoing costs by eliminating printing, shipping and scanning costs. Developing automated scoring methods would also reduce ongoing costs and improve the timeliness of results. Finally, deeper cognitive analyses of the content standards are needed to support better diagnostic information on individual students.

While considerable funding may be available for initial development of common assessments, there are no guarantees with respect to ongoing support. Many states may be hesitant to implement improved assessment systems if they cannot also see reduced costs down the road. In addition to administration, scoring, and reporting costs, ongoing development costs must also be considered. Investments in authoring and item-tracking systems that reduce ongoing development costs may also be needed.

Another issue of sustainability is ongoing teacher involvement and benefit from participation in assessment development and use of assessment results to improve instructional effectiveness. Plans for improved assessments should also include consideration of professional development opportunities that allow teachers to take full advantage of the improvements. Ways of sustaining teacher interest in and use of assessment results may also be needed.

Support for Improved Assessments

Potential Savings through Common Assessments

A brief survey of current state assessment costs was conducted to provide information on potential cost savings from combining state assessments. The survey focused only on contract costs and did not include internal staff costs, which is often hard to track. CCSSO is conducting a similar, albeit more comprehensive survey of assessment costs. However, results from that survey are not yet available. A copy of the survey instrument is included as Attachment A.

In analyzing current assessment costs, development costs which might be shared by more than one state were separated from costs of administration, scoring, and reporting which would have to be borne by each participating state. The idea is to see how much funding for development might be increased if several states adopted common standards and pooled their resources to develop a common assessment.

In fact, the New England Common Assessment Program is an example of just such an approach. Vermont, New Hampshire, Rhode Island, and now Maine have pooled resources to develop a common assessment that is significantly better than what each state could afford to develop on their own. Working with the Center for Assessment to coordinate technical oversight, this consortium of states has been remarkably successful (DePascale, 2009a, 2009b).

Recently, the U.S. Department of Education announced a grants program to support consortia of states in developing common assessments. A total of \$350M will be available for states willing to adopt common standards and work together on the development of common assessments.

The assessment cost survey was completed by 15 state testing directors and 2 development contractors. Most respondents provided data on more than one assessment. Commonly, many states have one contract for grade 3-8 assessments, another one for high school assessments, and yet another for an alternate assessment program.

Key data elements included annual contract costs along with information on the nature of the assessment supported by the contract. Most states and both testing contractors provided estimates of the proportion of the cost that went for initial development (fixed costs) and the proportion that went for administration, scoring, and reporting (variable with the number of tests administered).

It was quickly evident that costs for regular and alternate assessments were quite different. Another key cost driver for ongoing (variable) costs was whether extended constructed response questions, requiring human scoring, were included. Table 1 shows the mean and range of annual development and administration (per student) costs.

Table 1. Average Development and Administration Costs by Assessment Type

Assessment Type	N	Mean	S.D.	Min	Max
Annual Development Costs (in thousands of dollars)					
Alternate	9	363	215	100	686
Regular – ECR	13	1329	968	127	3600
Regular - MC Only	5	551	387	220	1130
Admin Cost/Student (in dollars)					
Alternate	9	376	304	40	851
Regular – ECR	16	26	18	4	65
Regular - MC Only	6	3	3	1	9

Note: Extended constructed response (ECR) tests include writing assessments and other tests requiring human scoring using a multi-level scoring rubric. Multiple choice (MC) tests are normally machine scored.

As shown in Table 1, one key finding was the very significant variability in assessment costs from one state to another. This variability may, in part, reflect differing degrees of involvement of state personnel and local teachers in ways that could reduce contractor costs. Results from this survey were judged adequate for estimating development and administration costs for different types of test to an order of magnitude. A more detailed survey would be required to achieve accuracy to several significant digits.

Given current interest in more innovative assessments, the primary focus is on current costs for tests requiring human scoring. The majority of states responding to this survey provided data for this form of assessment. On average, each state is spending well over \$1 million dollars (average \$1,329,000) annually to develop new forms for this type of assessment and an average of \$26 per student to score them. Not surprisingly, alternate assessments which are usually administered one-on-one and usually require human scoring are more expensive to administer, but less is being spent on development. It is possible that fewer new forms of the alternate

assessments are being developed or that less extensive field testing is involved in their development.

Support for the development of new state assessments could come from a combination of common assessment grants and pooling current annual development costs. The \$350 million targeted for the development of common assessments will be divided across an unknown number of grants. Funding for a particular consortium would depend on the number of different grants and also on the number of states pooling their existing development costs. Figure 1 shows how available funding would vary as a function of these two factors.

Development Funds by Number of Grants and Number of States

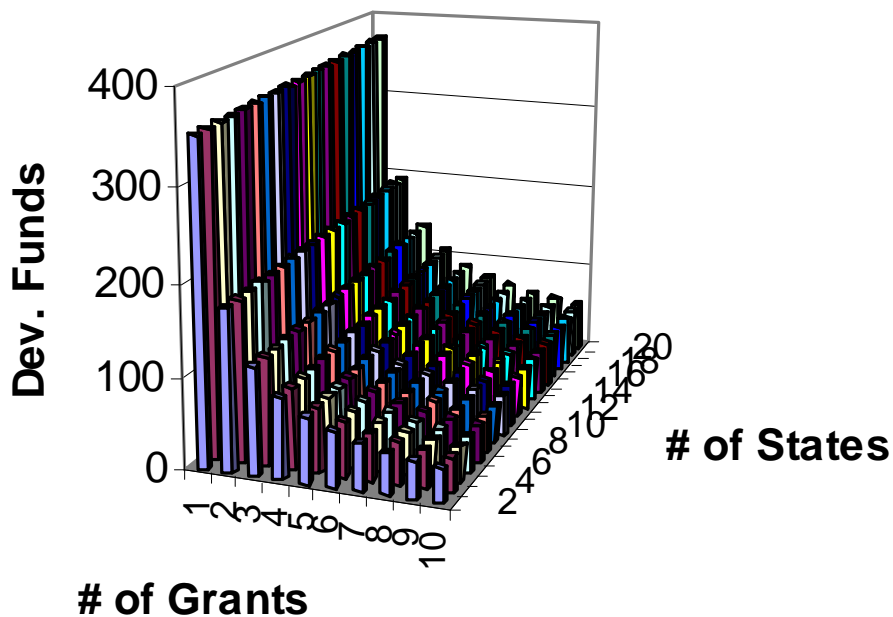


Figure 1. Development funds available to consortia of states through common assessment grants and pooling of development costs.

The most significant support for the development of common assessments is likely to come from the Race-to-the-Top assessment grants. At present requirements for these grants are

being reviewed and it is not known how many grants will be awarded. The U.S. Department of Education seems unlikely to fund more than a few consortia. At present, five potential consortia have been identified (Sawchuk, 2010). If we assumed that there are 5 grants averaging \$70M million each and , in addition, if 10 states pool the money they are currently spending on assessment development, they would have an additional \$13M to spend on initial development for a total of \$83 million. After initial development is complete and the RTT grant money goes away, the consortium of 10 states might still have about \$13M to spend on ongoing development without any increase above current costs.

As stated before, there will be little or no savings for administration, scoring, and reporting costs from pooling resources across states. There might, however, be savings in these areas from design and implementation of a more efficient assessment system. Computer-based testing might eliminate printing, shipping, and scanning costs. Additional savings might result from advances in automated scoring of constructed responses.

Support for Deeper Cognitive Analyses of Common Standards

Common standards would make it possible for states to pool their current assessment resources and also to compete for RTT assessment grants. Common standards would also facilitate deeper cognitive analyses of standards and objectives for student performance than is currently possible with separate state standards. A key goal is that common standards will not only be common, but also better. ACHIEVE touts the new standards as being fewer, clearer, and higher than most state standards. In addition, the common core standards are expected to show thoughtful progressions from one grade to the next leading up to 12th grade standards for readiness for college and work.

Some of the funding for initial development might usefully be spent on a careful cognitive analysis of how the skills to be mastered are taught as well as how they should be assessed. The cognitive analysis should identify barriers to master of specific content. Common standards might also support more extensive analyses of how to help students overcome specific barriers that are identified for individual students through improved diagnostic assessments.

If several states adopt common standards, it should be possible to investigate the effectiveness of strategies used by different states and districts for sequencing instruction. Establishing an expected path for learning (learning trajectory) will allow assessments to

pinpoint student progress along the path as well as identify ways in which students might get off of the expected path. (See, for example, Clements & Sarama, 2009.)

Building learning trajectories requires a model that shows how the content a student must master at one grade is clearly related to expectations for achievement at the next grade. Macro-level learning trajectories will chart progress across grades toward 12th grade readiness. It is hoped that such grade-to-grade relationships will be evident in the common core standards when they are completed.

It would also be useful to develop a good model for within-year sequencing to support the development of micro-level trajectory targets. Of course, within year growth will likely be somewhat multivariate, with different trajectories for different content strands or different types of skills for each subject. Within-year sequencing may very well not be evident in the initial version of the common core standards. Ongoing analyses and evaluation may be required to establish the most effective trajectories for mastering grade-level expectations in each subject.

Figure 2 shows an illustration of how learning objectives for elementary and middle school mathematics might be sequenced for two facets of mathematics. . The example is taken from existing grade K-7 content standards for a single state. The trajectory for numbers and operations begins very simply with the ability to count objects (such as fingers) up to 10 and progresses to understanding and using both positive and negative exponents. Figure 2 also shows a parallel trajectory for measurement skills that ranges from comparing lengths or knowing units of time to solving much more complex measurement problems. The example illustrates that possibility that learning trajectories may be only partially ordered as some points. The measurement trajectory begins with separate paths for measurement of lengths and areas and for measurement of time. At some point these paths converge, leading to more advanced topics. The numbers and operations trajectory includes a separate path after mastery of decimal arithmetic for understanding and computing percents and for numerical operations with fractions.

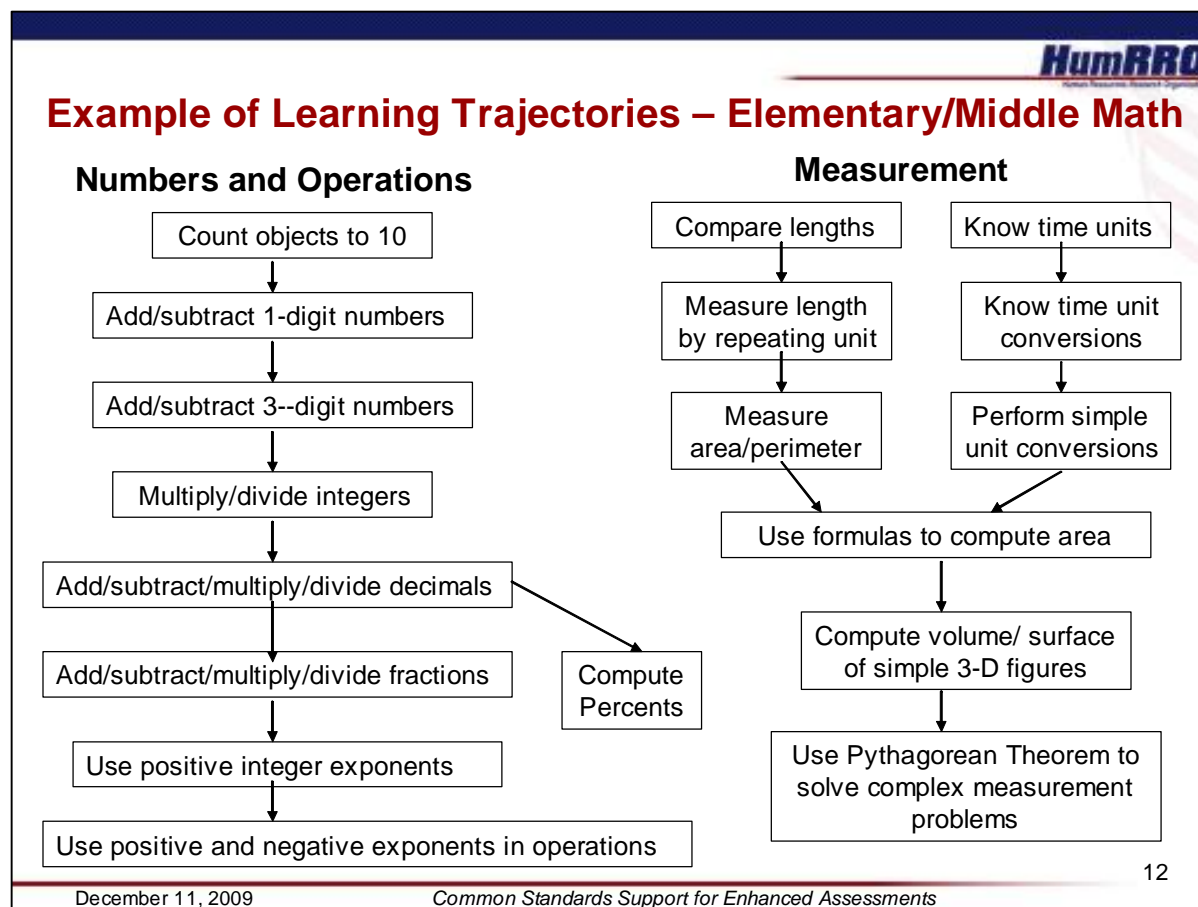


Figure 2. Example of a Multivariate Learning Trajectory Model for Elementary and Middle School Mathematics.

Improvements in Need of Support

Common standards could support increased funding for the development of common assessments as well as encouraging deeper cognitive analyses of the knowledge and skills to be mastered. So what specific improvements in state assessments might result from this support?

First, if we are to do more than simply count the number of children being left behind, we need **more meaningful scales for reporting** individual and group results. Even NAEP is limited to saying things like average achievement went up 5 points without providing any basis for understanding what these 5 points mean. Alternatively, the percentage of students scoring above some arbitrary achievement level is related to very broad and general descriptions of each of the performance levels and no specific information about what an individual student knows and can do.

If we are to go beyond simple accountability uses of state assessments, we also need much **more timely information**. Diagnostic information that is not available until well after the student has completed a course is of limited use. End-of-year assessments do not provide detailed information on the effectiveness of individual instructional units.

Third on this wish list is **better diagnostic feedback**. Specific information on what the student does not yet understand and perhaps why would be allow assessment results to be used to improved individual student achievement.

Fourth on our wish list is **greater teacher engagement** in the development and use of the improved assessments. Real improvements will come only through greater teacher effectiveness.

Finally, we also need studies to support he **validity of additional uses** of assessment information.

Reporting Scales

Current grade-by-grade achievement level reporting leaves much to be desired. Performance standards are typically not related across grades, other than perhaps by normative information. Many states are experimenting with vertical scales that purport to support better information on student progress. However, restrictions requiring “grade-level” testing limit the validity of such scales. If a 4th grader is performing well above his peers, the vertical scale might tell us he or she is performing at the 8th grade level, but we have never checked to see if he or she has met ANY of the 8th grade expectations. Also note that so-called “equal interval” properties of vertical scale are not at all related to equal levels of effort to move from one point to the next or to equal value in doing so.

A useful alternative to current reporting scales could involve development of learning trajectory scales, based on a careful analyses and sequencing of cross-grade content. Different trajectory scales might be developed for separate content strands, allowing us to chart student progress from one grade to the next along multiple expected paths. Placing students along the trajectory would also provide clearer information on what the student should work on next in a way that current scale scores cannot.

If grade-level testing restrictions were removed, an adaptive approach to testing could provide much better information about students who are well below or well above expectations for their current grades. In this approach, all students start answer questions related to expectations for their grade. If it is clear that the student meets these expectations, the assessment

moves on to higher grade content. If it is clear that the student does not meet current grade expectations, the assessment would move back to expectations for earlier grades. Adaptive testing would provide accurate estimates of each student's placement across the entire macro-level learning trajectory.

More Timely Results

Another idea for improving state assessments is to integrate mid-term or quarterly assessments with end-of-year testing. With prompt feedback, such an approach could provide timely information to teachers to improve student achievement during the year.

One concept for meeting this need is to combine an adaptive test using machine (immediately) scored questions with a small set of extended constructed response questions designed to identify particular student difficulties at the level estimated by the machine-score portion of the test. As soon as students finish the test, their teachers would be provided with both initial scores and the responses to the targeted open-ended questions. Teachers could be trained to score and interpret specific student needs based on the open-ended responses. Teachers could also use the open-ended responses to confirm the initial learning trajectory estimates. If they believed the initial estimates were in error, teachers might be able to submit open-ended responses for independent adjudication. Key features of this approach are; (1) immediate feedback to teachers and students and (2) greater teacher engagement in diagnosing specific learning problems.

Richer Diagnostic Information

The bookmark approach was a major breakthrough in setting performance standards. Items are ordered by difficulty and panelists place bookmarks at the points where they believe a student answering correctly up to that point has demonstrated a specified level of performance. It is possible to look at items just above the bookmark as indicators of what a student at one level needs to focus on to reach the next. Since the items usually reflect diverse content, this is rarely that useful of an approach. By contrast, the learning trajectories approach orders achievement by content and can give a better picture of what the student needs to master next. Thus learning trajectories can provide better diagnostic information on individual student needs.

Better diagnostic information is also needed in evaluating program or even teacher effectiveness. Multiple assessments per year would provide clearer data on student growth in

mastering grade-specific content. This concept of growth could support better measures of program effectiveness in comparison to growth based on end-of-year tests covering different content.

Better Summative Information

In addition to better diagnostic information and within-year growth, we also need better summative information for school and student accountability. We need to combine results from multiple within-year assessments to identify student progress in reaching ultimate readiness standards. The use of multiple assessments throughout year would support improved precision in estimating student achievement levels in comparison to the currently limited time available for end-of-year testing. This design would also support the use of within-year as well as across-year growth in school accountability models.

Greater Teacher Engagement

There are several ways of involving teachers in the development and use of improved assessments that could significantly increase their engagement in diagnosing individual student problems and evaluating and improving their instructional practices. First, teachers could be intimately involved in assessment development, creating or reviewing test questions or exercises. Teachers would benefit from the training routinely provided to item-writers and reviewers and could also contribute a classroom perspective that is not always part of the test development process. Second, teacher involvement in the scoring and interpreting responses to extended constructed response questions might deepen their understanding of the type of evidence that supports or fails to support mastery of specific areas of the curriculum. Interpretation of learning trajectory results would also help teachers see the relevance of skills they are teaching for future student achievement. Finally, support for training in the use of assessment results could significantly increase teacher participation in continuous process improvement with respect to their teaching.

Validity Evidence

Along with richer assessments and wider uses of assessment results come requirements for broader and better validity evidence. For example, a dynamic assessment would require somewhat more complex alignment studies to show clear relationships between test results and underlying content standards. Also, assumptions underlying learning trajectories would have to

be tested to be sure that placing students a one point on the trajectory implies general mastery of content associated with previous points and at most limited mastery of content associated with succeeding points.

Extensive work has been done to support alignment studies for state assessments, but relatively little has been done to establish the validity of test results as measures of learning effectiveness for specific curricular content or as predictors of readiness for future learning objectives. The increased availability of longitudinal student data bases make it possible to conduct a much wider range of validity studies.

Summary

Three main points were described in this paper. The first point is that state assessments are sorely in need of improvement. Current assessments are, at best, limited measures of school effectiveness, let alone program or teacher effectiveness. While a reasonable gauge of the extent to which students in different groups are being “left behind,” current assessments provide information of limited diagnostic use and is also not available in time to help individual students.

The second point of the paper is that adoption of common content standards would, in fact, support many of the needed improvements in state assessments. If states work together on common assessments, there will be greater resources to support improvements. Of course, federal grants for common assessments will also provide one-time resources for even more significant developments. Adopting common standards will also provide a basis for deeper analysis of how students learn, particularly if the common standards are well articulated across grades and are clearer than the current content standards of most states.

Finally, several highly promising avenues of improvement were described. Combining interim and end-of-year assessments would provide more timely and, overall, more reliable information than is currently available. More frequent assessments would also support the study of within-year learning trajectories and within-year progress along these trajectories. Adaptive testing would allow accurate placement of students along broad learning trajectories and could be used to select highly informative open-ended questions. Responses to these questions might be provided to teachers along with learning trajectory “scores.” Finally, greater teacher involvement in test development, scoring, interpretation, and use could yield very positive improvements to instruction and learning.

References

- Clements, D. and Sarama, J. (2009). *Learning and Teaching Early Math: The Learning Trajectories Approach*. New York: Routledge Press.
- DePascale, C.A. (2009a). The New England Common Assessment Program: Notes on the collaboration among four New England states.. Dover, NH: National Center for Educational Assessment. (Available at www.nciea.org.)
- DePascale, C.A. (2009b). Establishing a state consortium for assessment. Dover, NH: National Center for Educational Assessment. (Available at www.nciea.org.)
- McNeil, M. (2009). “46 States Agree to Common Academic Standards Effort”. *Education Week*, June 10, 2009.
- Sawchuk, S. (2010). “States Rush to Join Assessment Consortia””. *Education Week*, February 3, 2010.

Attachment A: State Assessment Cost Survey Form

Survey of Cost Data for State Assessments

State:			
Questions:	1 st Contract	2 nd Contract	3 rd Contract
1. Approximate cost of contract (in thousands of dollars)			
2. Number of years covered			
3. Grades covered			
4. Subjects (circle all that apply): R-Reading, M- Math, S- Science SS- Social Studies, W-Writing, O-Other	R M S SS W O	R M S SS W O	R M S SS W O
5. Assessment Type (circle all that apply): R-Regular, A-Alternate (1%) M – Modified (2%), E- English Proficiency (include only if part of a larger contract)	R A M E	R A M E	R A M E
6. Approx. total number of tests administered per year (thousands)			
7. Includes short-answer questions: N-No, M-Yes machine scored, Y-Yes with human scoring	N M Y	N M Y	N M Y
8. Includes extended response questions? N – No, W-Writing Essay Only, Y-Yes	N W Y	N W Y	N W Y
<i>Please skip the remaining questions if information is not easily available.</i>			
9. Approximately what percentage of the total costs are for test development?			
10. Approximately what percentage of the costs are for administration, scoring, and reporting?			
11. Contractor (short name is fine)			

Use multiple pages if there are more than 3 relevant assessment contracts.

Completed by: _____ Date: _____

Please e-mail electronic versions to LWISE@humrro.org
Or fax paper copy to 831-375-4021

If you are interested in a copy of the paper based on this information, please provide an e-mail address: _____